

Table of Contents

List of tables

List of figures

1. aims and objectives.
2. Background and Context

Sports analytics has experienced rapid growth both for teams and individual players in recent years. Statistical tools can be particularly useful for assessing performances and defining efficient gaming strategies. In this paper, we focus on basketball, a sport that has been at the forefront of using analytic tools, utilizing data from the National Basketball Association (NBA).

In the context of basketball statistics literature, many of the proposed approaches can be traced back to Bill James's Sabermetrics approach in baseball (James, B. (1984). *The bill James baseball abstract 1987*. Ballantine Books). However, Oliver's seminal work in 2004 (Oliver, D. (2004). *Basketball on paper: Rules and tools for performance analysis*. Potomac Books, Inc) is considered a significant contribution. Oliver identified the "Four Factors" that influence game outcomes, assigning different weights to them. These factors include Effective Field Goal Percentage, Turnovers per Possession, Offensive Rebounding Percentage, and Free Throw Attempt Rate. He presents a valuable perspective that suggests player evaluation should not solely revolve around scoring statistics but should also consider their offensive contributions in relation to the negative impacts on defense. The author factors in elements such as shooting efficiency, assists, and free throws as contributions to scoring. Additionally, an attempt is made to incorporate statistical data related to shot-blocking, opponent field goal percentages, and more to assess defensive value. While this represents a step in the right direction, his paper also has some issues. For instance, it tends to oversimplify problems when catering to the general public and frequently relies on assumptions that can undermine the depth of the arguments. John Maynard Keynes has reminded us that the significant challenge in statistics lies not in the methods themselves but in the absence of a means to validate the data we input.

Kubatko (Kubatko, J., Oliver, D., Pelton, K., & Rosenbaum, D. T. (2007). *A starting point for analyzing basketball statistics*. *Journal of Quantitative Analysis in Sports*, 3(3), 1–24.) presents the quantitative analysis of sports as a rapidly growing branch within the scientific domain. Particularly, the concept of equal possessions for opponents in a game is deemed crucial in basketball analysis. While estimates of possessions have been in existence for approximately two decades, the various formulas employed have occasionally led to confusion. Their objective is to clarify the relationship between possessions and various statistical data by demonstrating how most previous formulas are special cases of a more general formulation. Furthermore, they aspire to establish their new estimates as a common foundation for future possession estimations. In addition to outlining data sources for statistical research in basketball, the paper also delves into various concepts and methodologies. These encompass offensive and defensive efficiency, tactical strategies, per-minute statistics, pace adjustments, true shooting percentage, effective field goal percentage, rebound rates, the Four Factors, plus/minus statistics, counterpart statistics, linear

weights metrics, individual possession usage, individual efficiency, the Pythagorean method, and the Bell Curve method. It is worth noting that while this list is not exhaustive in terms of the methodologies employed in the field, the author contends that these tools are well-suited within the possession framework and form the fundamental basis for discussions in the realm of statistical research in basketball.

Yang et al. (Yang, C. H., Lin, H. Y., & Chen, C. P. (2014). *Measuring the efficiency of nba teams: Additive efficiency decomposition in two-stage dea*. *Annals of Operations Research*, 217(1), 565–589.) conducts an evaluation of the efficiency of NBA teams using a two-stage DEA framework. Employing an additive efficiency approach, the overall team efficiency is decomposed into first-stage wage efficiency and second stage on-court efficiency, with individual endogenous weights determined for each stage. Empirical results indicate that NBA teams perform better in terms of wage efficiency compared to on-court efficiency, primarily due to the latter being influenced by numerous uncontrollable factors. Furthermore, NBA teams tend to have higher average weights in the first stage, suggesting that team managers can enhance organizational efficiency more effectively by carefully recruiting players.

Piette (Piette, J., Pham, L., & Anand, S. (2011). *Evaluating basketball player performance via statistical network modeling*. In *The 5th MIT Sloan sports analytics conference*.) and his colleagues identified the primary challenge in assessing individual player performance in basketball as the need to account for interaction effects among teammates. To address this issue, they adapted a network-based algorithm to estimate centrality scores and their corresponding statistical significance. This approach quantifies aspects that were previously challenging to measure and adjusts for potential confounders. They determined a player's statistical contribution within this network by examining the frequency with which the player is visited during a random walk on the network. Additionally, they employed bootstrap techniques on the original weights to create reference distributions for significance testing.

Metulini et al. (Metulini, R., Manisera, M., & Zuccolotto, P. (2018). *Modelling the dynamic pattern of surface area in basketball and its effects on team performance*. *Journal of Quantitative Analysis in Sports*, 14(3), 117–130.) Using a time series of basketball player coordinates and focusing on the dynamics of the surface area occupied by the five players on the court, this study has two primary objectives: (i) to provide tools for a detailed description and analysis of game dynamics concerning surface area and (ii) to investigate its impact on both team and opponent scoring. We propose a three-step integrated procedure employing various statistical modeling approaches. Firstly, we utilize a Markov Switching Model (MSM) to detect structural changes in the surface area. Subsequently, we conduct descriptive analyses to highlight associations between these structural phases and relevant game variables. Finally, we assess the relationship between phase probabilities and points scored by the team and the opponent using Vector Auto Regressive (VAR) models. This procedure is applied to real-world data, and in the analyzed case studies, structural changes are found to be strongly linked to offensive and defensive game phases, with observable associations between surface area dynamics and the points scored by both the team and the opponent.

Sandholtz et al. (Sandholtz, N., Mortensen, J., & Bornn, L. (2020). *Measuring spatial allocative efficiency in basketball*. *Journal of Quantitative Analysis in Sports*, 16(4), 271–289.) argue that in

basketball, every shooting opportunity carries an opportunity cost; a player's shot eliminates all potential opportunities for their teammates in that particular offensive play. Therefore, a player's shooting efficiency should ultimately be considered relative to the lineup. This aspect of efficiency – the optimal way to distribute shots within a lineup – is a focal point of their research.

Cervone et al. (Cervone, D., D'Amour, A., Bornn, L., & Goldsberry, K. (2016). A multiresolution stochastic process model for predicting basketball possession outcomes. *Journal of the American Statistical Association*, 111(514), 585–599.) present Basketball games constantly evolve in space and time as players engage in continuous interactions with their teammates, the opposing team, and the basketball itself. However, current analyses of basketball game outcomes rely on discrete summaries, simplifying these intricate interactions into aggregated statistics such as points and assists. Therefore, a framework is proposed in this study that utilizes optical player tracking data to provide real-time estimates of the expected number of points that will be scored by the end of a possession. This metric, known as Expected Possession Value (EPV), is derived from a stochastic process model that captures the dynamic evolution of basketball possessions. The proposed framework models this process at multiple levels of granularity, distinguishing between the players' continuous, infinitesimal movements and discrete events such as shot attempts and turnovers. Transition kernels are estimated using hierarchical spatiotemporal models, enabling the sharing of information across players while maintaining computational feasibility even with extensive datasets. In addition to estimating EPV, these models offer fresh insights into players' decision-making tendencies in relation to their spatial strategies. In summary, this framework represents a significant advancement in the analysis of basketball games, allowing for a more nuanced understanding of the sport's dynamics by considering both space and time, as well as the complex interactions among players, opposing teams, and the basketball itself.

Regression-based approaches, such as the Bayesian linear regression model by Deshpande and Jensen (Deshpande, S. K., & Jensen, S. T. (2016). Estimating an nba player's impact on his team's chances of winning. *Journal of Quantitative Analysis in Sports*, 12(2), 51–72.), They introduced a Bayesian linear regression model to estimate the impact of individual players while controlling for the factors of other players on the court. We introduced several posterior summaries to determine the rankings of players within their team and across the league. This enables us to identify players who have high salaries, but a lower impact compared to their teammates, as well as those players whose high impact is not adequately captured by existing metrics.

In the realm of neural networks, notable works include Shen et al. (Shen, J., Zhao, Y., Liu, J. K., & Wang, Y. (2020). Recognizing scoring in basketball game from AER sequence by spiking neural networks. In *2020 international joint conference on neural networks (IJCNN)* (pp. 1–8). IEEE), and Babaee Khobdeh et al. (Babaee Khobdeh, S., Yamaghani, M. R., & Khodaparast Sareshkeh, S. (2021). Clustering of basketball players using self-organizing map neural networks. *Journal of Applied Research on Industrial Engineering*, 8(4), 412–428.). For an extensive review of statistics in basketball, Turner and Franks (Turner, Z., & Franks, A. (2021). Modeling player and team performance in basketball. *Annual Review of Statistics and Its Application*, 8, 1–23.) have focused on methods for quantifying and describing basketball gameplay. At the team level, various approaches for characterizing team strategy and performance have been discussed. Meanwhile, at the individual player level, an in-depth exploration of numerous tools for player assessment has been conducted. These encompass metrics for

evaluating overall player value, defensive prowess, and shot modeling, as well as methodologies for comprehending player performance over multiple seasons through player production curves.

However, as Baghal points out, there is a lack of attention to applying multivariate analyses to basketball metrics, despite the well-established understanding that winning a game depends on a plurality of interrelated variables, each with a different weight. To address this gap, Baghal (Baghal, T. (2012). Are the “four factors” indicators of one factor? an application of structural equation modeling methodology to nba data in prediction of winning percentage. *Journal of Quantitative Analysis in Sports*, 8(1).) employs In Structural Equation Modeling (SEM), the explained variance in regression is like that of a linear regression model, highlighting the importance of offensive and defensive quality, with a greater influence observed for offensive quality. Team salary is associated with offensive quality but unrelated to defensive quality or winning. Consequently, a second structural equation model is proposed in which the impact of salary on winning is mediated through its relationship with offensive and defensive quality. Given that salary is linked to offensive quality but not defensive quality, and offensive quality holds greater significance for winning percentages, this suggests that investments in player salaries are directed towards enhancing offensive performance, thereby influencing success through the performance they support.

with the aim of expanding analysis using a multivariate approach, we propose the use of Bayesian networks (BNs). BNs are probabilistic expert systems capable of identifying complex dependence structures among performance indicators. They enable what-if analysis and provide a pictorial representation of dependency structures through graph topology. The belief updating process can help identify impactful variables and suggest optimal strategies in various contexts.

To handle non-Gaussian distributions without the need for variable discretization, we consider an extension to the Gaussian copula. Specifically, we apply Non-Parametric Bayesian Networks (NPBNs), as proposed by Hanea et al. (2006) and Kurowicka and Cooke (2006).

In conclusion, this work introduces the use of BNs in sports analytics, emphasizing their versatility and natural suitability for prediction, strategy suggestion, and decision-making in sports. We employ BNs to identify the primary determinants influencing a player's winning percentage, considering key performance indicators from the literature and incorporating data and expert information. // 接下来的内容，需要根据实际情况补充

2.1 Basic concepts of basketball

In this section, we present different terminologies that are related to the game of basketball.

2.1.1 Court

Basketball games are played on a rectangular court with a basket at each end. The court is divided into two halves by a midcourt line. In the center of the court, there is a center circle, and before the game begins, only two players from each team are allowed inside this circle. The court is further divided into two-point and three-point areas, which are separated by the three-point line

2.1.2 Team structure

Each team consists of five players. The five players on each team are assigned to different positions on the court. The positions assigned to players are point guard, shooting guard, small forward, power forward, and center.

2.1.3 Basic rules

Here are the basic rules of basketball:

- 1.The objective of each team is to score by putting the ball into the opponent's basket.
- 2.Each team can have a maximum of five players on the court at a time.
- 3.The game is divided into four quarters, each lasting 12 minutes, resulting in a total game time of 48 minutes. Overtime of 5 minutes is played in case of a tie, and this continues until a winner is determined.
- 4.Scoring two points occurs when the ball is shot into the basket from within the three-point arc.
- 5.Scoring three points occurs when the ball is shot into the basket from beyond the three-point line.
- 6.Scoring one point happens when the ball is successfully thrown into the basket from the free-throw line (also known as a free throw).
- 7.The ball can be passed to another player or dribbled (moved while bouncing the ball) from one point to another. Once a player stops dribbling, they cannot start again. Crossing the midcourt line while in possession of the ball signifies that the team cannot cross back over the midcourt line.
- 8.The team in possession of the ball has a maximum of 24 seconds to attempt a shot.
- 9.Illegally contacting an opponent's player results in a personal foul.

2.1.4 Player statistics

Basic game statistics¹ of a basketball player are given in Table 1

2.2 Related work(改一下题目)

The related work can be divided into ranking and prediction in basketball. In ranking the performance of basketball players is evaluated by using various statistics whereas in prediction the outcome of the basketball game is predicted using machine learning classifiers.

2.2.1 Ranking basketball players

NBA game statistics like points, blocks, rebounds, field goals etc are widely used for rating the basketball players. John Hollinger (per) introduced a formula that uses player box score statistics to measure the efficiency of the player. To know how much a player is efficient, the idea of on and off the court was proposed by [15] (P. Fearnhead, B.M. Taylor, On estimating the ability of nba players, J. Quant.Anal. Sports 7 (3) (2011).). They observed a Novel Model and Approach for Assessing the Abilities of NBA Players. The primary concept revolves around directly measuring a player's proficiency by comparing their performance on and off the court, considering the constantly changing dynamics of players present during different game times. This is achieved through a structured methodology utilizing data from multiple seasons to evaluate a player's abilities within a single season. It separately assesses offensive and defensive strengths and amalgamates these factors to provide an overall rating. Game statistics data will be considered for predicting player capabilities. The impact of an NBA team player is evaluated by [16](S.K. Deshpande, S.T. Jensen, Estimating an NBA player's impact on histeam's chances of winning, J. Quant. Anal. Sports 12 (2) (2016) 51–72.), they used a Bayesian linear regression model for finding an individual player impact on the team winning. The said research ranks the players with respect to their team and across the leagues. Slack based measure method is used by [17] (F. Asghar, M. Asif, M.A. Nadeem, M.A. Nawaz, M. Idrees, A novel approach to ranking national basketball association players, J. Glob. Econ. Manage.Bus. Res. (2018) 176–183) to rank players in NBA. They compared their ranking with player impact measure approach. They conclude that even the players who are ranked top by slack based measure approach are ranked at bottom by player impact measure approach. The reason for getting different ranking on same data is because both methods work in different manner. Authors in [18] (J. Koster, B. Aven, The effects of individual status and group performance on network ties among teammates in the national basketball association, PLoS One 13 (4) (2018) e0196013.) shows how network ties among players are affected through individual status and group performance. Relationship between game statistics and match outcome is explored by [19](S. Zhang, A. Lorenzo, C. Zhou, Y. Cui, B. Gonçalves, M. Angel Gómez, Performance profiles and opposition interaction during game-play in elite basketball: evidences from national basketball association, Int. J. Perform. Anal. Sport 19 (1) (2019) 28–48.). They also consider how player's technical and physical performance is affected by interaction of opposition.(补充一下那个

可视化的东西, 20%, 然后 PER 不可靠这玩意!)

2.2.2 Match outcome prediction

The fuzzy rule-based system (FRBS) is proposed by [20] (K. Trawinski, A fuzzy classification system for prediction of the results of the basketball games, in: *International Conference on Fuzzy Systems*, IEEE, 2010, pp. 1–7.) for the prediction of the basketball match outcome. Feature selection was applied for the selection of best features and various fuzzy models were used for the prediction of match outcome. A model for college basketball was proposed by [21] (F.J. Ruiz, F. Perez-Cruz, A generative model for predicting outcomes in college basketball, *J. Quant. Anal. Sports* 11 (1) (2015) 39–52.) that combined a simple soccer model and Poisson factorization. The simple soccer model identifies each team by its attack and defense coefficients whereas the Poisson factorization considers the elements of the matrix that are independent of the Poisson random variables. For match outcome prediction in basketball an integrated model called HSVMDT (Hybrid Support Vector Machine and Decision Tree) is proposed by [22] (P.-F. Pai, L.-H. Chang Liao, K.-P. Lin, Analyzing basketball games by a support vector machines with decision tree model, *Neural Comput. Appl.* 28 (12) (2017) 4159–4167.) . Feature selection was used to select the best features (7 features were selected out of 17). HSVMDT was tested on both selected features and on all 17 features HSVMDT achieved 82.25% with feature selection and without feature selection, the accuracy was 67%. The decision tree generates many rules that can cause confusion for decision makers. Rules pruning was used to limit the number of rules. For measuring the quality of decision rules, the sum of testing accuracy and coverage index was used. The results showed that decision rules have better quality after pruning. The rules generated by said model aim to help coaches to identify which factors are affecting match outcome. Analysis based on classification and regression tree was performed by [23] (M. A Gómez, S. J Ibáñez, I. Parejo, P. Furley, The use of classification and regression tree when classifying winning and losing basketball teams, *Kinesiology: Int. J. Fundam. Appl. Kinesiol.* 49 (1) (2017) 47–56.) to find best predictor in order to classify teams as winning or losing teams. Their analysis showed that in fast paced games the importance of defensive rebounds is 100%, importance of free throws is 94.7%, assists 86.1% and importance of fouls is 55.9%. On the other hand the importance of variables in slow paced games are: free throws is 100%, defensive rebounds 82.3%, fouls 68.4%, assists 66.9%, 2-points 62.2% and importance of 3-point field goals is 62.1%. Data driven and data envelopment analysis-based techniques were used by [24] (Y. Li, L. Wang, F. Li, A data-driven prediction approach for sports team performance and its application to national basketball association, *Omega* (2019) 102123.) for predicting the

performance of sports team. They used multi-variate logistic regression to find relationship between winning probability and match outcome. Their study suggests that team coaches and managers should focus on communication and co-operation of team. Various machine learning models are used by [25] (F. Thabtah, L. Zhang, N. Abdelhamid, NBA game result prediction using feature analysis and machine learning, *Ann. Data Sci.* 6 (1) (2019) 103–116.) for the prediction of match outcome in basketball. They examined the strength of various features for match outcome prediction. The defensive rebound was observed to be the most suitable feature for match outcome prediction. Discrete-time and finite-state Markov chain has been used by [26] (J. Shi, K. Song, A discrete-time and finite-state Markov chain based in-play prediction model for NBA basketball matches, *Comm. Statist. Simulation Comput.* (2019) 1–9.) to predict the outcome of the match when the game is in progress. The aim of the said model is to model the difference between the home team and the visiting team score at some time point. The predictions for the ongoing match can be made on the current score of the team instead of past data.

2.2.3 Applications of machine learning techniques

Machine learning models have wide range of applications. Here we give an overview of some of the application of machine learning techniques.

SVM [27] (C. Cortes, V. Vapnik, *Support-Vector Networks Machine Learning*, vol. 20, Kluwer Academic Publisher, Boston, MA, 1995.) and Naive Bayes [28] (I. Rish, et al., An empirical study of the naive Bayes classifier, in: *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, vol. 3, no.22, 2001, pp. 41–46.) techniques have been used by [29] (S.V. Wawre, S.N. Deshmukh, Sentiment classification using machine learning techniques, *Int. J. Sci. Res. (IJSR)* 5 (4) (2016) 819–821.) for classification of movie reviews.

Text classification based on document embedding is used by [30] (R.A. Sinoara, J. Camacho-Collados, R.G. Rossi, R. Navigli, S.O. Rezende, Knowledge-enhanced document embeddings for text classification, *Knowl.-Based Syst.* 163 (2019) 955–971.) . One of the application of machine learning in the domain of legal documents is presented by [31] (I. Chalkidis, M. Fergadiotis, P. Malakasiotis, I. Androutsopoulos, Large-scale multi-label text classification on eu legislation, 2019, arXiv preprint arXiv:1906.02192.) , where the authors applied various models for multi-label text classification on legislation

documents. Words in pair neural networks is presented by [32] (W. Yujia, L. Jing, S. Chengfang, J. CHANG, et al., Words in pairs neural networks for text classification, Chin. J. Electron. 29 (3) (2020) 491–500.) for text classification that overcome the limitation of text classification based on single word with multiple meanings. Novel machine learning model SS3 proposed by [33] (S.G. Burdisso, M. Errecalde, M. Montes-y Gómez, A text classification framework for simple and effective early depression detection over social media streams, Expert Syst. Appl. 133 (2019) 182–197.) for text classification that have the ability of early risk detection on social media. Siamese capsule networks that are based on local and global features for text classification has been used by [34] (Y. Wu, J. Li, J. Wu, J. Chang, Siamese capsule networks with global and local features for text classification, Neurocomputing (2020).)

Machine learning has also been actively used for classification of spam messages. A review of soft techniques for classification of sms spam is presented by [35] (O. Abayomi-Alli, S. Misra, A. Abayomi-Alli, M. Odusami, A review of soft techniques for SMS spam classification: Methods, approaches and applications, Eng. Appl. Artif. Intell. 86 (2019) 197–212.) . Discrete Hidden Markov Model is used by [36] (T. Xia, X. Chen, A discrete hidden Markov model for SMS spam detection, Appl. Sci. 10 (14) (2020) 5011.) for spam detection that has the capability to exploit the order of words and can handle the problem of low term frequency. Rule based algorithm with the ability of constant time complexity has been used for detection of spam [37] (T. Xia, A constant time complexity spam detection algorithm for boosting throughput on rule-based filtering systems, IEEE Access 8 (2020) 82653–82661.)

classical machine learning technique are not much efficient in situations where the decisions are time-dependent, for such situations, [38] (Y. Chen, Y. Zhou, Machine learning based decision making for time varying systems: Parameter estimation and performance optimization, Knowl.-Based Syst. 190 (2020) 105479.) presented a machine learning model that have the ability to work in time varying systems. Machine learning techniques based on evolutionary

frame-work has been used in medical domain on clinical data [39]. (J.A. Castellanos-Garzón, E. Costa, J.M. Corchado, et al., An evolutionary framework for machine learning applied to medical data, Knowl.-Based Syst. 185 (2019) 104982.) For prediction of breast cancer, Support Vector Machines and Artificial Neural Networks has been applied by [40] (E.A. Bayrak, P. Kırıcı, T. Ensari, Comparison of machine learning methods for breast cancer diagnosis, in: 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science, EBBT, IEEE,2019, pp. 1–3) on Wisconsin Breast Cancer dataset.

Table 1

S No.	attribute	Definition
1	Minutes Played Per Game (MP)	The average number of minutes a player plays per game.
2	Field Goals Per Game (FG)	The average number of field goals a player makes per game.
3	Field Goal Attempts Per Game (FGA)	The average number of field goal attempts a player makes per game.
4	Field Goal Percentage (FG%)	The ratio of field goals made to field goal attempts, used to measure shooting accuracy.
5	3-Point Field Goals Per Game(3P)	The average number of 3-point field goals a player makes per game.
6	3-Point Field Goal Attempts Per Game(3PA)	The average number of 3-point field goal attempts a player makes per game.
7	3-Point Field Goal Percentage(3P%)	The ratio of 3-point field goals made to 3-point field goal attempts, used to measure 3-point shooting accuracy.
8	2-Point Field Goals Per Game(2P)	The average number of 2-point field goals a player makes per game.
9	2-Point Field Goal Attempts Per Game(2PA)	The average number of 2-point field goal attempts a player makes per game.
10	2-Point Field Goal Percentage(2P%)	The ratio of 2-point field goals made to 2-point field goal attempts, used to measure 2-point shooting accuracy.
11	Effective Field Goal Percentage(eFG%)	A measure of shooting efficiency that considers both 2-point and 3-point field goals.
12	Free Throws Per Game (FT)	The average number of free throws a player makes per game.
13	Free Throw Attempts Per Game (FTA)	The average number of free throw attempts a player makes per game.
14	Free Throw Percentage (FT%)	The ratio of free throws made to free throw attempts, used to measure free throw shooting accuracy.
15	Offensive Rebounds Per Game (ORB)	The average number of offensive rebounds a player grabs per game.
16	Defensive Rebounds Per Game (DRB)	The average number of defensive rebounds a player grabs per game.
17	Total Rebounds Per Game (TRB)	The average number of total rebounds (offensive and defensive) a player grabs per game.
18	Assists Per Game (AST)	The average number of assists a player records per game.
19	Steals Per Game (STL)	The average number of steals a player records per game.
20	Blocks Per Game (BLK)	The average number of blocks a player records per game.

21	Turnovers Per Game (TOV)	The average number of turnovers a player commits per game.
22	Personal Fouls Per Game (PF)	The average number of personal fouls a player commits per game.
23	Points Per Game (PTS)	The average number of points a player scores per game.
24	Player Efficiency Rating (PER)	A comprehensive statistic that evaluates a player's performance in various aspects.
25	Offensive Win Shares (OWS)	Measures a player's offensive contribution to the team's wins.
27	Win Shares (WS)	A comprehensive measure of a player's contribution to the team's wins, combining both offensive and defensive contributions.