A data-driven prediction approach for sports team performance and its application to National Basketball Association

Article *in* Omega · September 2019

DOI: 10.1016/j.omega.2019.102123

Feng li
Southwestern University of Finance and Economics
15 PUBLICATIONS **122** CITATIONS

Contents lists available at ScienceDirect

Omega

journal homepage: www.elsevier.com/locate/omega

A data-driven prediction approach for sports team performance and its application to National Basketball Association[☆]

Yongjun Li[a], Lizheng Wang[a], Feng Li[b,*]

[a] School of Management, University of Science and Technology of China, Hefei, Anhui Province 230026, China
[b] School of Business Administration, Southwestern University of Finance and Economics, Chengdu, Sichuan Province 611130, China

ARTICLE INFO

Article history:
Received 17 February 2019
Accepted 24 September 2019
Available online xxx

Keywords:
Data envelopment analysis (DEA)
Performance prediction
Data-driven
Sports team
National basketball association

ABSTRACT

Performance prediction is an issue of vital importance in many real managerial applications. This paper will propose a prediction approach for sports team performance based on data envelopment analysis (DEA) methodology and data-driven technique. The proposed approach includes two steps: The first one conducts a multivariate logistic regression analysis to examine the relationship between the winning probability and game outcomes at the team-level. The other one addresses a DEA-based player portfolio efficiency analysis to optimally choose players and plan the playing time among players in the court. The second step aims to use players' and team's historical data to train the future and obtain the most promising outcomes in terms of their average inefficiency status. Finally, we apply the proposed performance prediction approach to National Basketball Association and take Golden State Warriors as an example to illustrate its usefulness and efficacy. We obtain the prediction results for the 2015–16 regular season based on a four-season dataset from the 2011–12 season to the 2014–15 season. Further, we carry out multiple experiments to provide deeper discussion and analysis on according prediction results. It shows that the DEA-based data-driven approach can predict the sports team performance very well and can also provide interesting insights into the performance prediction problem.

1. Introduction

Nowadays, performance prediction of different production units is a very important issue in the competitive environment [15]. A good prediction on the future performance can benefit various actions and objectives such as resource allocation, production adjustment, revenue management, etc. In addition, the performance prediction is also of vital significance to set a development target for these production units. For instance, a manufacturer would address its production planning based on the production efficiency prediction from various aspects. A country can draft and manage the national economic plan very well through a well-conducted production analysis and prediction. More specially, the predictive performance of collective sports has attracted more and more attention in recent years and can be useful for designing practice tasks, training and competition [25,33,44]. Among all these performance prediction applications in the real managerial applications, one of the most important sectors is the sport industry. The performance prediction and analysis are of major interests of all related coaches, players, sport scientists, investors and performance analysts [20].

Sports all over the world have attracted an ever-increasing attention in the past several decades. The sports industry has earned great value and revenue all over the world and take the global sports apparel market for instance, it had a total revenue of about 145.34 billion dollars in 2015, which was estimated to grow up to 181 billion in 2019.[1] Because of the big market of sport industries and its popularity among all human beings, much attention has been paid to forecast the match outcomes of sporting events [22,79]. Among all sports activities, the National Basketball Association (NBA) from the United States of America is one of the most valuable leagues within which only the average franchise value of NBA teams recorded 1.923 billion dollars in 2019.[2] The NBA has become one of the biggest sports business because of its broadcasting rights, advertising and merchandising sales [62]. For collective sports like NBA where a group of players competes against another, an issue of most importance is to gain advantages against competitors for each season and achieve a better performance as much

[☆] This manuscript was processed by Associate Editor Aparicio.

[*] Corresponding author.

E-mail addresses: lifeng1990@swufe.edu.cn, lfeng90@mail.ustc.edu.cn (F. Li).

[1] https://www.statista.com/statistics/254489/total-revenue-of-the-global-sports-apparel-market/, "Total revenue of the global sports apparel market 2012-2025", first accessed on Oct. 3, 2016 and re-accessed Aug. 14, 2019.

[2] https://www.statista.com/statistics/193696/franchise-value-of-national-basketball-association-teams-in-2010/, "Value of National Basketball Association franchises 2019", accessed on Aug. 14, 2019.

https://doi.org/10.1016/j.omega.2019.102123
0305-0483/© 2019 Elsevier Ltd. All rights reserved.

as possible. For this purpose, the manager (or coach) would focus much on the analysis of future possible performance and accordingly adjust the team setting. Here in this paper, we will address the issue of predicting the future performance using NBA teams as an example. For simplification of consideration and without loss of generality, we address the problem of how to optimize a certain team's future performance in the competition with other teams by choosing players and determining their playing time in the court, and eventually maximize its possible winning probability for the next regular season.

When address the winning probability prediction problem, some certain assumptions should be made. To this end, we first make a basic assumption that the history will occur again, implying that the past performance is a good reference/predictor of future performance [88]. In other words, we believe that the data in the given sample can depict all possibilities in future and an average inefficiency concept is adopted to define the relative efficiency status in the prediction approach. Although this basic assumption may not always hold in any case, it is an important and natural assumption for predicting future performance based on past performance, as these extreme events cannot be analyzed for data-driven approaches. Then the production technology derived from history data can be used to approximate the future production activities. Further, to address the performance prediction problem based on future production activities, a frontier approach is required to construct the production possible set and production frontier, upon which we can estimate the possible outcomes level in future. Here we will use the non-parametric Data Envelopment Analysis (DEA) method as an ideal production frontier approach. DEA, first introduced by Charnes et al. [17] and further extended by Banker et al. [12], is a mathematical programming method originally proposed to evaluate the relative performance of peer decision making units (DMUs) based on certain inputs and outputs in the sample. Now the DEA methodology has been applied to many disciplines since its seminal work [4,5,45,49,99]. In this paper, DEA was selected from several aspects: first, the DEA methodology is good at dealing with multiple inputs and multiple outputs; second, it considers the trade-off among various inputs and outputs; third, this non-parametric approach does not use any subjective weights, which shows great flexibilities in developing the performance prediction approach. Further, player transactions are very common in NBA and the DEA methodology provides us opportunities to predict game outputs considering different player portfolios, which is very suitable for NBA operations in which the coach will address such a work. Besides, it is possible to conduct DEA-based experiments and simulations to find best player portfolios that can determine a maximal winning probability, which has also been frequently considered in NBA games, thus we can predict the NBA team performance well. Note that there are various assumptions on the returns to scale properties in the DEA literature, here we will use DEA methods with variable returns to scale assumption, for it captures the nonlinear relations between the playing time and outputs for players.

In this paper, we will develop a data-driven sports team performance prediction approach based on data envelopment analysis methodology. And the proposed approach will be applied to the National Basketball Association to verify its usefulness and efficacy. The prediction process can be divided into two steps: The first step is to conduct a multivariate statistics regression analysis to estimate the quantity relationship between the winning probability and team's various game outputs obtained in games. Here an S-shape logistic regression method was adopted to ensure that the winning probability is located in the interval of [0, 1]. Also, through this step we divide all game outcomes into desirable outputs and undesirable outputs according to the sign of estimated parameters. By taking the regression equation obtained in the first step as the objective function, the other step conducts a DEA-based player portfolio efficiency analysis to optimally choose players and allocate the playing time among players in the court. This step aims to use the historical data in the sample to train the "*future*" and obtain the most promising performances. Within this step, we use an average inefficiency concept to limit the relative efficiency status for both players and the team considered. Afterwards, we apply the proposed DEA-based data-driven prediction approach to a dataset of Golden State Warriors in NBA from the 2011–12 season to the 2014–15 season. Then we obtain the prediction results for the next 2015–16 NBA regular season and compare it with the real result to gain some valuable insights. Also, some deeper experiments and discussions on the results are conducted. The results show that the proposed data-driven approach can predict the future performance very well. In addition, sensitivity analysis can be used to investigate which player is the most valuable one in that team. These multiple experiments are interesting and show the usefulness of the proposed approach and how it can be used to gain insight into the problem.

In this paper, we provide an insight to develop DEA-based data-driven approaches for predicting future performance of collective sports teams. Here we contribute to the literature from the following several aspects: first, we apply DEA-based methodology for the performance prediction problem. DEA is traditionally proposed for ex-post evaluations, while this paper uses it for ex ante analysis. To the best of our knowledge, this paper is the first one that builds DEA approaches for predicting sports performance at the team-level. Second, we propose a data-driven approach based on a frontier production method. It would be of especially vital significance in the big data context. Third, a real dataset derived from Golden State Warriors of NBA is used to demonstrate the usefulness and efficacy of the proposed approach. Through the empirical application, we not only built an applicable performance prediction approach, but also provided valuable and interesting insights on NBA.

The reminder of this paper is organized as follows: Section 2 reviews some relevant literature. Section 3 first briefly describes the problem setting and then develops a data-driven prediction approach based on data envelopment analysis models. In Section 4, the proposed approach is applied to Golden State Warriors by using a real dataset from the 2011–12 season to the 2014–15 season to predict the winning probability in the next 2015–16 regular season. And some experiments are also provided to discuss and analyze further on according prediction results. Finally, Section 5 summarizes this paper.

2. Literature review

Note that this paper proposes a DEA-based data-driven prediction approach for sports team performances and applies it to a real case of NBA, hence the relevant literature mainly unfolds from three aspects, performance prediction, DEA methodology and performance in NBA.

2.1. Performance prediction

The first research stream involves the prediction of future performance. Prediction, which means to estimate a certain level or a possibility, has been studied in many applications. Through a well-processed performance prediction procedure, a production unit can be able to predict the future performance, analyze the abnormal situation, and consequently, take measures to prevent performance from deteriorating [36]. Performance prediction is usually investigated in business activities. For example, Nyhuis et al. [67] applied queuing, simulation and mathematical approximation approach to

Please cite this article as: Y. Li, L. Wang and F. Li, A data-driven prediction approach for sports team performance and its application to National Basketball Association, Omega, https://doi.org/10.1016/j.omega.2019.102123

predict logistics firms' performance. Liu and Frangopol [57] designed an optimal bridge maintenance plan according to the probabilistic performance prediction. Variyam et al. [84] proposed a fast-transient testing methodology to predict the performance indicators of analog circuits. Huang [36] developed a production performance prediction system for the semiconductor manufacturing. Facing with high volatility in the market, the performance prediction or output prediction in the financial sector has also attracted lots of attention. For example, Lam [40] used neural networks to integrate fundamental and technical analysis for financial performance prediction, in which the rate of returns on common shareholders' equity was predicted based on 16 financial statement variables and 11 macroeconomic variables. Ravi et al. [74] proposed a soft computing system for bank performance prediction, in which the performance of a bank in the next year was predicted based on its previous financial data within two years. Acknowledging the advantages of combination of multiple prediction methods, Xiao et al. [90] applied the Dempster-Shafer evidence theory to addressing the prediction problem, and further used rough set to determine the weights attached to individual prediction methods. And consequently, the authors approached the prediction of financial distress for listed companies. Also, there are some studies focusing on enterprises bankruptcy prediction such as Premachandra et al. [73], Ouenniche and Tone [71] and du Jardin [27].

There are also some sport team performance prediction studies, although the task of predicting the outcomes of a sport event would be very difficult [33,85]. The most common sport team performance prediction is that of predicting outcomes of specific football matches [23]. For example, Min et al. [61] integrated Bayesian inference, rule-based reasoning, and an in-game time-series approach to predict the results of football matches. Leitner et al. [44] analyzed several different approaches in assessing sport participants' abilities and winning probabilities, and further integrated these methods in a common framework to forecast the results of the European football championship in 2008. Constantinou et al. [21] developed a Bayesian network approach for forecasting the results of Football matches in English Premier League. Using a dataset of 203 national football teams, Omondi-Ochieng [69] predicted football match performances based on per capita gross national income and football workers.

It is rather remarkable that regression approaches are usually used in sports performance studies [2,75]. For example, Atkinson and Nevill [2] suggested ordinary linear or multiple linear regression methods when continuous and unbounded sport performance indicators were considered, while logistic regression or discriminant function analysis for categorical indicators. Based on a sample of five seasons in NBA, Melnick [60] used regression analysis to find a significant relationship between team assists and game win-to-loss records. Andrade et al. [6] used regression to examine the relation between sleep quality, mood, and game results in the elite athletes participating in Brazilian volleyball competitions. Robertson et al. [66] used logistic regression method to investigate the relation between game outcomes (win/loss) and discrete team performance indicators in Australian Football League regular seasons. Saavedra-García et al. [75] applied a logistic regression additive model to examining the relative age effect on sport performance based on a sample of 21,639 players from 1908 to 2012. Omondi-Ochieng [69] used binary logistic regression analysis to investigate the relation between national economic prosperity, the acquisition of football workers and predict football performances based on per capita gross national income and football workers. Amatria et al. [8] developed a multiple logistic regression model as well as two simple logistic regression models to investigate the association between game formats, skill learned and game performances. Yang et al. [92] used multinomial logistic regression to identify the key physical and technical performance variables related to team quality in the Chinese Super League. Recent studies on regression approaches in sports performance can also be seen in Silva et al. [77], Gómez et al. [32], Chalitsios et al. [14], Gamble et al. [31], etc.

It can be seen that all these previous articles addressed the performance prediction issue using traditional methods, and most papers focused on normal tendency while the sport results might be affected by rarely seen events or extreme events [23]. Therefore, in this paper we will develop the prediction approach based on a combination of frontier production method and logistic regression and apply it to the sport sector. We will use the frontier production method to measure the inefficiencies in the previous period and also use it to generate possible outputs in future.

2.2. Data envelopment analysis

The second category of relevant researches focuses on data envelopment analysis, which is known as a non-parametric production frontier approach used for performance evaluation. This method can date back to Farrell [28], in which the author suggested to compare the existing production setting with the ideal status to evaluate the relative efficiency. Also, Shephard [76] defined an input distance function to measure the cost efficiency. Further, Charnes et al. [17] creatively combined the Farrell measure and Shephard distance function to propose a new mathematical programming method called DEA, aiming to measure the relative efficiency of peer decision making units (DMUs). Afterwards, Banker et al. [12] developed the Charnes et al. [17] model to consider the variable returns to scale assumption, and consequently obtained a pure technical efficiency being free from the scale effect. Since their seminal work, the DEA-based approaches and its applications have been extensively studied in the literature [50,54,98], ranging from hospitals [70], universities [3,42], sports [55,91] to banks [19,48,80], mergers and acquisitions [46,89], resource allocation [47,51] and other production processes [52,56].

More specifically, many scholars developed DEA-based approaches to evaluate the performance for NBA teams and players. For example, Cooper et al. [24] used non-zero weights DEA to evaluate the NBA players, and the weight results can also be used to identify individual players' relative strengths and weaknesses. Chen et al. [18] developed a bounded integer DEA approach to evaluate the player performance in NBA by noting that most indicators in basketball games are integers and bounded. Similar research of DEA approaches on NBA player performance can also be seen in Lee and Worthington [43], Mansoor and Sinah [58] and Asghar et al. [10]. Meanwhile, Moreno and Lozano [62] used a network DEA approach to evaluate the team efficiency in NBA, and the authors also calculated the possible team budget reduction and games won by the team. Yang et al. [91] also evaluated the team efficiency in NBA under a network environment, and the authors used an additive two-stage decomposition framework to estimate the wage efficiency and on-court efficiency. Villa and Lozano [86] used a dynamic network DEA model to address the basketball games efficiency, and the authors considered the difference of home and visitor teams.

Note that almost all DEA models are designed for ex-post efficiency analysis based on pre-specified inputs and outputs data, very few studies focus on the performance prediction for future. For example, Yang et al. [93] proposed a hybrid minimax reference point-DEA model to incorporate the judgments from central decision maker and individual DMU, and that model was supposed to address the future performance forecasting and target setting. By considering the uncertainty in future performance, Chang et al. [16] took past, present and future performances measures to evaluate efficiencies in highly volatile operating environments. Zhang and Wang [96] integrated support vector machine and information granulation with DEA models to evaluate the future efficiency of

DMUs. Radovanović et al. [65] integrated DEA and machine learning algorithms to predict a new virtual player' efficiency in NBA based on existing players' historical data. Here in this paper, we will try to develop DEA-based method to predict the future sport performance at the team-level. The historical data is used to *train* the decision-making units, and the possible game outputs in future are generated by learning from previous practices. Therefore, both past performance evaluation and future performance prediction are considered based on the DEA methodology.

2.3. NBA and its performance

NBA is one of the most popular sport in the world [62], and thus has attracted attention from both academic researchers and practitioners. Many papers can be found in the published literature working on NBA. For example, Hofler and Payne [34] applied the stochastic production frontier approach to evaluate the efficiency of 27 NBA teams during the 1992–1993 season, and teams' efficiencies were determined by examining how closely they performed relative to their potentials. Berman et al. [13] studied the relationship between team performance and shared team experience in NBA, and results implied a predicted positive relationship which will decline along with the increase of shared experience until changing into a negative relationship. Katayama and Nuch [37] studied the causal effect of within-team salary dispersion on team performance using game-level panel data in NBA, while results shown that there was not a statistically significant relationship between salary dispersion and team performance. Podlog et al. [72] examined the influence of injury/illness to team performance in NBA based on a 25-year series data from 1986 to 2010. Recent studies can be seen in Kester et al. [38], Feddersen et al. [29], Koster and Aven [39], Spiteri et al. [78], etc.

As discussed previously, NBA is also an important application area of DEA methodology, and many scholars developed DEA-based approaches to evaluate the performance for NBA teams and players. Except for above studies in Section 2.2, Aizemberg et al. [1] used a game cross-efficiency method to assessing NBA teams' performance from 2006 to 2010, in which the payroll and the average attendance were inputs while the wins and the average points per game are outputs. Masoumzadeh et al. [59] developed a new DEA approach without explicit inputs and applied the proposed approach to evaluating 35 players. Meanwhile, Lee and Berri [41] estimated the team efficiency through constructing different efficient frontier for different player positions. Considering interval numbers and absolute dominance relations, Ang et al. [7] evaluated and ranked the performance of 30 NBA teams using an interval cross-efficiency approach. Moreton and Lozano [63] evaluated the productivity change of NBA teams from 2006–07 to 2012–13 seasons, in which a Malmquist network DEA approach was used.

Within the sport industry like NBA, the performance or outcome prediction is of vital importance, since the popularity and number of fans of any NBA team will depend heavily on its results [62]. For example, Baghal [11] used structural equation modeling methodology to investigate the prediction indicators of winning probability. Radovanović et al. [65] first used a DEA model to evaluate existing players' efficiencies in NBA, then considered DEA scores as inputs and integrated linear regression, neural network and support vector machines to determine and predict the efficiency for a new player. Moxley and Towne [64] used a growth mixture model and past performances to predict the NBA career. Vračar et al. [87] proposed a novel method to generate a simulation of a basketball match between two teams according to team-level play-by-play events. That method was applied to NBA and such a simulation can be taken as a well representative of prediction results. Many researchers have built sport performance prediction approaches in NBA, but a common disadvantage is low prediction accuracy [33]. Thabtah et al. [82] built an intelligent machine learning framework to predict game results in NBA. Note in addition that there exists also a prediction report supposed as the official prediction to some extent, which is provided by Entertainment Sports Programming Network (EPSN, www.espn.com). EPSN defines an indicator called player efficiency rating (PER), which uses data obtained by a certain player in the court to measure his performance [35]. This indicator accounts for positive and negative playing statistics but pays more attention to the offense aspect than the defense aspect. Also, the famous data blog called FiveThirtyEight proposes a prediction method to calculate each NBA team's rating and its chances of advancing to the playoffs (projects.fivethirtyeight.com/2016-nba-picks). Although there exist methods approaching the team performance prediction in NBA, here in this paper we will propose a new data-driven prediction approach based on data envelopment analysis and logistic regression. This method will provide some new valuable insights for both performance prediction and NBA.

Note in addition that there exist also some data-driven studies, which speak highly of the value of data, from the sheer volume of data under consideration to the knowledge and information that lies behind data [81]. The data-driven analytics has been mainly investigated in manufacturing systems, as can be seen in Li et al. [53], Yu and Matta [94], Omar et al. [68], Zhang et al. [95], etc. For example, Li et al. [53] proposed a data-driven technique for detecting bottleneck with a production line in both short and long term. That method used data derived from the production line blockage and starvation probabilities and buffer content records. Yu and Matta [94] worked also on data-driven bottleneck identification approaches, but they used a statistical framework to reduce the data-driven detection inaccuracy. Zhang et al. [95] used a data-driven approach for the purpose of optimizing product specifications. The authors used operating data to construct customer satisfaction function and customer choice model, which will be further used to maximize profits and market shares and minimize cost. In particular, Daraio and Simar [26] presented a novel work by integrating data-driven approach with directional distance function in DEA and applied the proposed approach to banks. As can be seen that few researches have been proposed to integrate data-driven technique with DEA methodology and apply to sports, therefore it makes sense for this paper to propose a new data-driven prediction approach for sports team performance based on DEA models and explores its application in National Basketball Association.

3. Problem and methodology

We first describe the problem setting in Section 3.1 and then develop mathematical models in Section 3.2. For simplification of studying the performance prediction problem, here we narrow the topic to predict the NBA teams' winning probability in the next regular season based on data derived from several previous seasons.

3.1. Problem setting

For the NBA team under consideration, there are n players listed in the next season. For the simplification of research, we assume that there will be no transactions of players and only these given players are taken into account. Further, we consider no player injuries and all players can be allocated with playing time and play in the next season. For a pre-specified data sample, there are q games recorded for this NBA team, with the p-th $(p = 1,\ldots,q)$ game being characterized with playing time $t^p > 0$ and game outcomes $x_r^p \geq 0 (r = 1, \ldots, s)$. Besides, for each player $j = 1, \ldots, n$, he obtains the output data $x_{rj}^p \geq 0$ within the total playing time $t_j^p \geq 0$ in the p-th$(p \in P_j)$ game.

Please cite this article as: Y. Li, L. Wang and F. Li, A data-driven prediction approach for sports team performance and its application to National Basketball Association, Omega, https://doi.org/10.1016/j.omega.2019.102123

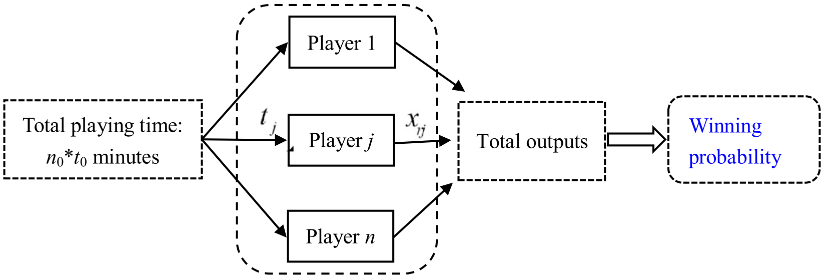

Fig. 1. The problem structure of winning probability prediction.

In the next competitive season, the coach will allocate the playing time across all players and expect outcomes for each player. And then the total outcomes summed from its players can be converted into its possible winning probability. In an appropriate way, the expected winning probability in the next competitive season can be maximized through all players' great efforts. Therefore, the problem arises about of how to allocate the playing time among players and plan their outcomes to maximize the winning probability and expected number of wins for all 82 games in the next regular season.

Here we depict the problem in Fig. 1. For a regular basketball game, there will be four quarters with each lasting twelve minutes (for NBA $t_0 = 12*4$). At the same time, five players (for NBA $n_0 = 5$) are allowed in the court simultaneously. Therefore, the total playing time $n_0 * t_0$ will be allocated across these n players,[3] and each player will obtain some outcomes through playing in the court. As a result, by summing the outcomes of individual player a quantified total outcome is predicted for the team as a whole. Further, the total outcome will be converted into a possible winning probability according to the quantity relationship between game outcomes and winning probability, and such a winning probability can be taken as the future performance prediction result for the considered NBA team. For simplicity, we introduction some key notations as follows:

Decision variables:

$t_j \geq 0$, the playing time allocated to player j, $t_0 \geq t_j \geq 0$;
$x_{rj} \geq 0$, the expected outcomes obtained by player j, $r = 1, \ldots, s$;
$\hat{x}_r \geq 0$, the total expected outcomes for the considered team such that $\hat{x}_r = \sum_{j=1}^{n} x_{rj}$.

Parameters:

t_0, the full time for a game, which is 48-minute in the NBA;
n_0, the number of players in the court;
j, index of players, $j = 1, \ldots, n$;
r, index of outputs, $r = 1, \ldots, s$;
P_j, set of games played by player j ($j = 1, \ldots, n$; $|P_j| = q_j$);
p, superscript for games the team played in the sample (p=1,...,q), or superscript for games player j played in the sample ($p \in P_j$);
$t^p > 0$, the playing time the team played in the p-th game;
$x_r^p \geq 0$, the r-th outcome achieved by the considered team in the p-th game;
$x_{rj}^p \geq 0$, the r-th outcome achieved by player j in the p-th game ($p \in P_j$);
$t_j^p \geq 0$, playing time of player j in the p-th game ($p \in P_j$).

To sum up, we would maximize a certain objective function f by allocating the playing time in an efficient way. Note that the objective function f is determined according to the targeted goals of analysis, therefore it can be all kinds of formulas. This basic prediction model can be formulated as model (1):

$$\begin{aligned} &\max \quad f \quad = f(t_1, \ldots, t_n) \\ &s.t. \sum_{j=1}^{n} t_j = n_0 \cdot t_0 \\ &\quad 0 \leq t_j \leq t_0, \forall j = 1, \ldots, n. \end{aligned} \tag{1}$$

Model (1) presents a simple idea that the objective function f is the function of players' playing time, and by appropriately distributing the overall playing time $n_0 \cdot t_0$ the objective function f can be maximized. Note that if f is a single objective function, then the optimal prediction result would be related to only one player's playing time. And as a result, we would focus on that player and his outcomes and accordingly neglect all other players. However, in all team sports like NBA the results would be comprehensive, so we need to use regression methods to obtain the formula of multivariable objective function, which shows the quantity relationship between the winning probability and the playing time allocated to individual player. Note in addition that we will use a production frontier method called data envelopment analysis (DEA) to convert the playing times $t_j (j = 1, \ldots, n)$ into various outputs $x_{rj} (r = 1, \ldots, s; j = 1, \ldots, n)$ in the court, so we only need to estimate the quantity relationship between the winning probability and various outcomes.

3.2. Mathematical modeling

3.2.1. Regression

In this subsection, the prediction procedure will be implemented through two steps. In the first step, we conduct a multivariate statistics regression analysis to estimate the quantity relationship between the winning probability and various game outcomes. Also, the outcomes are divided into desirable outputs and undesirable outputs based on the estimated parameters. Specifically, the winning probability is taken as the dependent variable, while the data of game outcomes obtained by the team in each game is the independent variable. Note that the results for each game would be binary (i.e., win or loss) and the traditional linear regression cannot ensure the winning probability be located in the interval of [0, 1], thus here we use the S-shape logistic regression. The definition equation is given by the following formula (2):

$$\frac{P_p}{1 - P_p} = e^{\beta_0 + \sum_{r=1}^{s} \beta_r x_r^p} \tag{2}$$

where $P_p = \Pr(Y_p = 1)$ denotes the winning probability for the p-th game, so the formula $P_p/(1 - P_p)$ is the ratio of win-to-loss; x_r^p is the value of outcome $r = 1, \ldots, s$ for the p-th game, and β_r is the sensitivity parameter to be estimated for outcome $r = 1, \ldots, s$, while β_0 is the intercept term implying the original value. To estimate the parameters in above Eq. (2), each game of the considered

[3] Normally, only some players (for example, 12 players in NBA) are allowed to play in each game. Here in this paper, we predict the winning probability and expected number of wins for all games (each team will play 82 games for each regular season in NBA) in the next regular season from a statistics perspective, thus we assume that all players can be allocated with playing time and play for a virtual game.

Please cite this article as: Y. Li, L. Wang and F. Li, A data-driven prediction approach for sports team performance and its application to National Basketball Association, Omega, https://doi.org/10.1016/j.omega.2019.102123

team would be taken as an observation. It is clear that to obtain the objective function for a particular team there are q observations (i.e., games). Formula (2) can be changed into formula (3) to estimate the parameters, where $Y = 1$ if the team wins, and $Y = 0$ if the team loses.

$$Y = \beta_0 + \sum_{r=1}^{s} \beta_r x_r + \varepsilon \tag{3}$$

There exist many estimation methods such as Least Square Method (LSM) and Maximum Likelihood Estimation (MLE) that can be used to estimate these parameters. Different parameters may be obtained by using different estimation methods. Suppose the estimated parameters are $(\beta_0^*, \beta_1^*, \dots, \beta_s^*)$, then all game outcomes can be divided into two categories according to the sign, namely, negative for undesirable outputs and positive for desirable outputs. Suppose in addition that there are m desirable outputs and (s-m) undesirable outputs. As a result, maximizing the following linear Eq. (4) can be equivalent to maximize the winning probability.

$$\hat{f} = \beta_0^* + \beta_1^* x_1 + \dots + \beta_m^* x_m + \dots + \beta_s^* x_s \tag{4}$$

3.2.2. Efficiency analysis for players and the whole team

To address the performance prediction problem, we need to specify a possible efficiency or inefficiency status dusing the predicted period. For this purpose, we should first approach the efficiency evaluation for both players and the team as a whole. There are many methods can be used for the efficiency analysis purpose, here we develop our approach based on a production frontier method called data envelopment analysis. This method uses historical data to construct an efficiency frontier, on which all decision making units are projected, and consequently the real units are compared with these projections to evaluate their relative efficiencies.

Here in this paper, we use the same model for both players and the team. Consider player j for example and denote its game set as P_j and the number of games played as $q_j = |P_j|$. Then, the relative efficiency of player $j (j = 1, \dots, n)$ in the o-th game $(o \in P_j)$ can be evaluated as follows:

$$\begin{aligned} & Max\left(\sum_{r=1}^{m} s_{rj}^{+o} + \sum_{r=m+1}^{s} s_{rj}^{-o}\right) \\ & s.t. \sum_{k \in P_j} \lambda_k t_j^k \le t_j^o \\ & \quad \sum_{k \in P_j} \lambda_k x_{rj}^k \ge x_{rj}^o + s_{rj}^{+o}, \forall r = 1, \dots, m \\ & \quad \sum_{k \in P_j} \lambda_k x_{rj}^k \le x_{rj}^o - s_{rj}^{-o}, \forall r = m+1, \dots, s \\ & \quad \sum_{k \in P_j} \lambda_k = 1 \\ & \quad \lambda_k, s_{rj}^{+o}, s_{rj}^{-o} \ge 0, \forall k \in P_j; r = 1, \dots, s. \end{aligned} \tag{5}$$

Model (5) is a slack-based directional distance function (DDF) model with endogenous direction vector, which varies from DMU (i.e., game) to DMU to maximize desirable outputs and minimize undesirable outputs simultaneously. Here we just consider the output-oriented slacks-based direction, for (1) there is only one input (i.e., playing time); (2) it can be easily demonstrated that projections of inefficient DMUs based on the optimal output slacks are strongly efficient if only one input is taken into account; (3) the optimal direction vector is non-radial, so model (5) would not overestimate the efficiency in situations where there exist non-zero slacks [30,97]. For the optimal direction vector, readers can refer to Arabi et al. [9]. As a result, model (5) estimates the largest performance inefficiency in terms of feasible decreases in undesirable outputs and feasible increases in desirable outputs.

Suppose the optimal solution of model (5) is $(\lambda_k^{o*}, s_{rj}^{+o*}, s_{rj}^{-o*})$, then the inefficiency ratio for each measure of player j (j=1,...,n) can be calculated according to Tone [83]:

$$\rho_{rj}^{+o*} = \frac{s_{rj}^{+o*}}{x_{rj}^o} (r = 1, \dots, m), \rho_{rj}^{-o*} = \frac{s_{rj}^{-o*}}{x_{rj}^o} (r = m+1, \dots, s) \tag{6}$$

Although the possible efficiencies for the team and players in the prediction period are unknown and full of uncertainty, a possible approximation of the possible efficiency is the average efficiency score among the past seasons. And this is a natural and most feasible way used in many applications. As a result, we can get a weighted average inefficiency for each measure, where the playing time in the o-th game $(o \in P_j)$ is taken as the weight attached to each individual inefficiency ratio.

$$\rho_{rj}^{+*} = \frac{\sum_{o \in P_j} t_j^o \rho_{rj}^{+0*}}{\sum_{o \in P_j} t_j^o} (r = 1, \dots, m), \rho_{rj}^{-*} = \frac{\sum_{o \in P_j} t_j^o \rho_{rj}^{-0*}}{\sum_{o \in P_j} t_j^o} (r = m+1, \dots, s) \tag{7}$$

3.2.3. Prediction model

To conduct the DEA-based performance prediction, at any rate we believe that the average inefficiency ratio $(\rho_{rj}^{+*}, \rho_{rj}^{-*})$ and $(\rho_r^{+*}, \rho_r^{-*})$ are good proxies of the inefficiency status of players and team within the prediction period, respectively. This assumption can be linked to the observation that well performed players and teams based on past performances may have better performances in subsequent competitions as compared with opponents [88]. As a result, we develop the following prediction model:

$$\begin{aligned} & Max \quad \hat{f} = \beta_0^* + \beta_1^* \hat{x}_1 + \dots + \beta_s^* \hat{x}_s \\ & s.t. \sum_{j=1}^{n} t_j = n_0 \cdot t_0 \\ & \quad \sum_{k \in P_j} \lambda_{kj} t_j^k \le t_j, \forall j = 1, \dots, n \\ & \quad \sum_{k \in P_j} \lambda_{kj} x_{rj}^k \ge x_{rj} + \rho_{rj}^{+*} x_{rj}, \forall r = 1, \dots, m; j = 1, \dots, n \\ & \quad \sum_{k \in P_j} \lambda_{kj} x_{rj}^k \le x_{rj} - \rho_{rj}^{-*} x_{rj}, \forall r = m+1, \dots, s; j = 1, \dots, n \\ & \quad \sum_{k \in P_j} \lambda_{kj} = 1, \forall j = 1, \dots, n \\ & \quad \sum_{j=1}^{n} x_{rj} = \hat{x}_r, \forall r = 1, \dots, s \\ & \quad \sum_{l=1}^{q} \lambda_l x_r^l \ge \hat{x}_r + \rho_r^{+*} \hat{x}_r, \forall r = 1, \dots, m \\ & \quad \sum_{l=1}^{q} \lambda_l x_r^l \le \hat{x}_r - \rho_r^{-*} \hat{x}_r, \forall r = m+1, \dots, s \\ & \quad \sum \sum_{l=1}^{q} \lambda_l = 1 \\ & \quad 0 \le t_j \le t_0, \forall j = 1, \dots, n \\ & \quad \lambda_{kj}, \lambda_l \ge 0, \forall j = 1, \dots, n; \forall k \in P_j; l = 1, \dots, q. \end{aligned} \tag{8}$$

In above model (8), the decision variable t_j is the playing time of player j and x_{rj} are his corresponding optimal outcomes, λ_{kj} is the intensity variable used to construct the efficiency frontier for each player j (j=1,...,n). The first constraint implies that the allocated playing time of all players sums precisely to the total playing time of a single game, while the subsequent four constraints ensure that the planned input-output for each player in the prediction period is within the production possible set (PPS), which is constructed using historical data under the variable returns to scale (VRS) assumption. Here the product of the planned input/output and a weighted inefficiency ratio gives an inefficiency slack, and this can be characterized as an inefficiency status adopted in the prediction period. The constraint $\hat{x}_r = \sum_{j=1}^{n} x_{rj}$ implies that the team's output is only derived from the outputs for all players. Besides, the rest three constraints are used to ensure that the expected outputs are within the team's production possible set. Again, the average inefficiency ratio is imposed on the team. Here we remove the constraint on the team playing time, as we consider a regular game environment and the team playing time will be identical for all games. In addition, acknowledging the regression equation in Section 3.2.1, the objective function of model (8) aims to maximize the winning probability by determining the optimal decision variable t_j and corresponding x_{rj}. It is no-

Table 1
Data of Golden State Warriors from 2011–2012 to 2014–2015 ($N = 297$).

	Two point	Three point	Free throw	Defensive rebound	Offensive rebound	Assist	Steal	Block	Turnover	Personal foul
Max	43	19	30	22	48	39	19	14	26	37
Min	18	1	5	1	18	10	1	1	5	8
Mean	30.1818	9.0640	15.8788	10.4579	33.1347	24.0067	7.9966	5.1515	14.2997	20.9562
Std.ev	4.6142	3.2538	5.0737	3.6250	5.6576	5.0290	3.1724	2.4330	3.7849	4.4889

Table 2
Data of players in regular seasons from 2011–2012 to 2014–2015.

NO.	Player name	Game count	Playing time	Two point	Three point	Free throw	Defensive rebound	Offensive rebound	Assist	Steal	Block	Turnover	Personal foul
1	Brandon Rush	138	17.62	164	118	75	32	295	126	45	79	102	127
2	Marreese Speights	294	16.44	978	16	479	489	886	199	75	150	302	659
3	Andre Iguodala	282	32.26	831	303	430	235	1119	1354	1263	430	125	509
4	Klay Thompson	306	32.24	1203	784	583	121	842	963	719	296	162	544
5	Andrew Bogut	178	35.29	578	1	77	413	1146	390	118	313	262	519
6	Anderson Varejao	141	35.29	590	0	245	491	956	1447	302	160	86	196
7	Leandro Barbosa	191	29.23	460	133	221	69	233	302	285	121	25	178
8	James Michael McAdoo	15	9.14	24	0	14	15	22	37	2	5	9	6
9	Ian Clark	53	6.86	24	21	14	7	28	35	27	17	5	23
10	Draymond Green	240	22.28	430	180	259	258	1058	1316	496	267	196	275
11	Stephen Curry	262	35.01	1202	874	916	176	898	1074	1961	456	50	847
12	Festus Ezeli	124	13.09	150	0	92	201	267	468	31	30	116	95
13	Shaun Livingston	278	21.54	723	3	370	180	530	710	841	217	102	349
14	Harrison Barnes	241	37.30	658	205	382	236	861	1097	330	178	53	256

Remark: Kevon Looney jointed the NBA since 2016, thus he is neglected here. Also, Jason Thompson was moved out since he momently played for Golden State Warriors and then departed. The data in the above table is the total value except for the average playing time in the unit of minute.

table that in model (8), the higher the inefficiency ratio $(\rho_{rj}^{+*}, \rho_{rj}^{-*})$ and $(\rho_r^{+*}, \rho_r^{-*})$ are, the lower the optimal objective function $\hat{f}^*$ is.

Model (8) develops a basic framework to address the prediction, but additional constraints can be still added. For example, these five starters are expected to play longer than those substitutes. Accordingly, we may insert an additional constraint $t_{st} > t_{su}$, where *st* and *su* imply starters and substitutes, respectively. Besides, some players may have an upper limit of playing time because of their energy and strength, thus a constraint such as $t_j \leq U_j$ would be available. More specially, if only $\bar{n}$ players instead of all n players will play in the next season, then we can introduce some binary variables $y_j \in \{0, 1\}$, which takes the value of one if player j is given minutes to play and otherwise takes zero. And further we can take the player count requirement into account by substituting the convexity constraint by $\sum_{k \in P_j} \lambda_{kj} = y_j$ inserting additional constraints $\sum_{j=1}^{n} y_j = \bar{n}$ and $t_j \leq t_0 \cdot y_j$.

Suppose the optimal solution of model (8) is $(t_j^*, \forall j; x_{rj}^*, \forall r, j;\ \hat{x}_r^*, \forall r)$, then the optimal objective function is $\hat{f}^* = \beta_0^* + \beta_1^* \hat{x}_1^* + \cdots + \beta_s^* \hat{x}_s^*$. As a result, by remaining the weighted average inefficiency ratio for each measure the maximal possible winning probability can be calculated via following formula (9).

$$P^* = \frac{1}{1 + e^{-\hat{f}^*}} = \frac{1}{1 + e^{-(\beta_0^* + \beta_1^* \hat{x}_1^* + \cdots + \beta_s^* \hat{x}_s^*)}} \tag{9}$$

Suppose there are N games for each team in each regular season, thus the expected number of wins would be statistically calculated as $N \cdot P^*$. In particular, the expected number of wins of games would be denoted as $82P^*$ in the NBA regular season.

4. Application to Golden State Warriors in NBA

In this section we apply the proposed DEA-based data-driven approach to the Golden State Warriors. We obtain the prediction results and also carry out multiple experiments to provide deeper discussion and analysis on according prediction results.

4.1. Data description

The Golden State Warriors was established in 1946 and jointed NBA in the same year. The Golden State Warriors is one of the early eleven team members of NBA, and so far the Golden State Warriors has gotten six grand champions. Here in this section, we use the data derived from the 2011–12 season to the 2014–15 season to predict the ideal performance in the next 2015–16 regular season. In addition, we will remove those games with overtimes for two reasons: on one hand, we will allocate the total playing time of 240 min for a virtual game, whereas the total playing time of these games with overtimes will exceed that value. On the other hand, within the standard playing time (i.e., 240 min), the win-to-loss results for these games with overtime are difficult to be quantified to one or zero, since these games ended in a draw. Also, the games players have played more than 48 min are neglected in this paper. As a result, we got the empirical data for Golden State Warriors and its fourteen players from *Basketball Reference* (http://www.basketball-reference.com/), as given in Tables 1 and 2.

Note that there exist some zero values for these players and it will cause some troubles in conducting the slacks-based efficiency evaluation, here we transfer these zeroes into positive values according to Tone [83]. Further, we process these original positive values in a similar way. As a result, we will use $(x_{rj} + 1)$ to replace x_{rj} for all players.

4.2. Preliminary prediction results

Firstly, we use the team outcome data to conduct the logistic regression analysis, and the results are given in Table 3 and Fig. 2. We find that the logistic regression equation can fit the winning probability very well. In fact, if we set the threshold to

Table 3
Regression results.

Variable	Coefficients	Std. Error	z-Statistic
β_0	−16.94654***	2.593090	−6.535268
β_1-two point	0.190505***	0.052596	3.622038
β_2-three point	0.325316***	0.076971	4.226503
β_3-free throw	0.137032***	0.036770	3.726745
β_4-defensive rebound	0.208003***	0.035743	5.819339
β_5-assist	0.106414**	0.047322	2.248719
β_6-steal	0.206344***	0.059389	3.474461
β_7-turnover	−0.130368***	0.045205	−2.883917
β_8-perfonal foul	−0.112095***	0.038009	−2.949159
	McFadden R^2	0.3791	
	LR statistic	150.6680	
	p-value (LR)	0.0000***	

Remark: *, ** and *** imply a significant level of 0.1, 0.05 and 0.01, respectively.

be 50%, then only 53 out of 297 games are found to be inconsistent with the real results. So, the accuracy can be nearly 82.15% ((297-53)/297). The McFadden R^2 0.3791, and LR statistic 150.6680, also verify the good fitness. All the above findings show that we can use the estimated regression equation to illustrate the quantity relationship between various game outcomes and the winning probability. As a result, the objective function to be used in the performance prediction model is given as below:

$$\begin{aligned} &-16.94654 + 0.190505 * \text{two point} + 0.325316 * \text{ three point} \\ &\quad +0.137032 * \text{free throw} + 0.208003 * \text{defensive rebound} \\ &\quad +0.106414 * \text{assist } + 0.206344 * \text{ steal } - 0.130368 * \text{turnover} \\ &\quad -0.112095 * \text{personal foul} \end{aligned}$$

Note that the offensive rebound and block are no longer in the regression equation, since both their estimated parameters are not statistically significant. Therefore, these two measures will be removed in the following part. Further, there are eight outputs, but according to the results in Table 3, two outputs (turnover and personal foul) are undesirable, which is also an intuitive finding for the real situations. Further, the three-point is the most important one among these eight measures. Besides, the three-point is subsequently followed by defensive rebound, steal and two point, implying that not only scoring is important, but also preventing the opponents from scoring is very important. This is just the basic idea behind the NBA games.

Further, we solve model (5) to calculate inefficiencies in each game for players and team. Afterwards, the inefficiency ratio is calculated based on formula (6), and then based on formula (7) the average inefficiency ratio can be obtained by weighting with the playing time, as given in Table 4.

Then we take these results into model (8) to optimally allocate the playing time and obtain the most promising performance results predicted for the next 2015–16 regular season, as shown in Table 5.

According to Table 5, we can conclude that by optimally choosing players and allocating the playing time, the most optimal winning probability for the Golden State Warriors is 73.95% and the expected victory in the 2015–16 regular season would be 60.64. An underlying assumption behind our results is that the average inefficiency ratio is adopted to characterize players' and team's inefficiency status in the prediction approach, so statistically the temporary shocks (or termed *noise* in individual game in the prediction season) from the average inefficiency ratios are not supposed to affect the overall prediction results, whereas those permanent shocks (or termed *change* on the inefficiency status of the team and players in the prediction season) will lead to different winning results. Note that the Golden State Warriors broke the record in NBA to obtain 73 victories in the 2015–16 season, so we can conclude that our prediction approach can give a good prediction result. Also, we can verify that the Golden State Warriors as a whole played very well in this season, since the real winning probability is even higher than the predicted one. This result can be also due to the fact that the Golden State Warriors shows an increase tendency on the winning probability from 2011–12 season to 2014–15 season, and as a result the average inefficiency ratio based on a four-season dataset may underestimate the potentials of Golden State Warriors. In addition, according to the results in Table 5 we are able to claim that Anderson Varejao, Stephen Curry, Harrison Barnes, Andrew Bogut and Andre Iguodala are the top five most valuable players in the Golden State Warriors, as they are supposed to obtain the most scores, and besides, there are also many other outputs achieved by these players.

Note that the proposed prediction approach given in model (8) may obtain a unique optimal objective function and accordingly the predicted winning probability is unique, but the solutions can

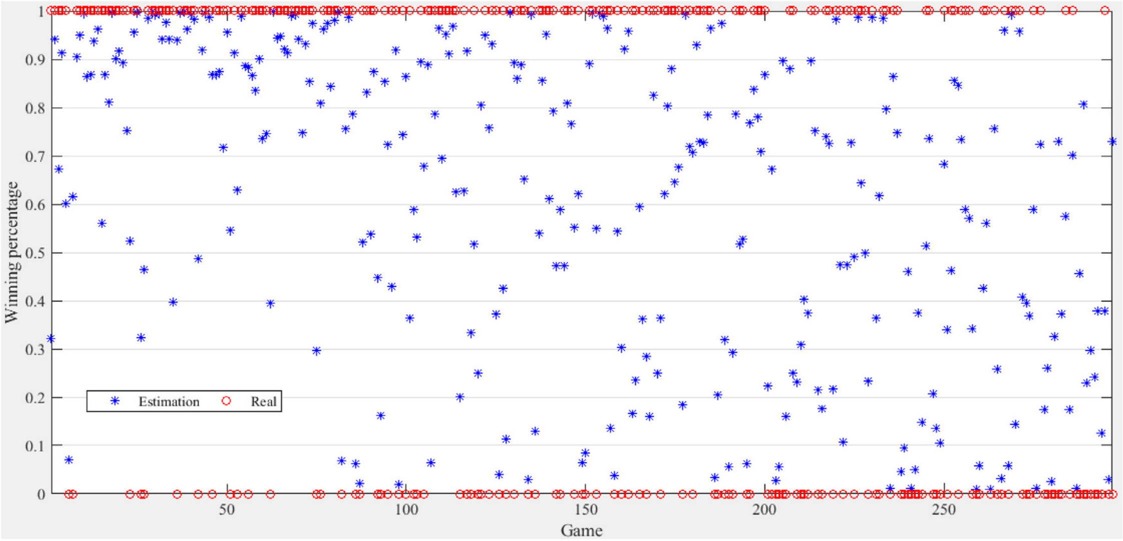

Fig. 2. Scatter diagram of winning probability.

Table 4
Average inefficiency ratio for the team and its players.

Player NO	Two point	Three point	Free throw	Defensive rebound	Assist	Steal	Turnover	Personal foul
1	0.4800	0.4458	0.1274	0.6081	0.1339	0.1946	0.0699	0.0984
2	0.4334	0.0092	1.9250	0.4619	0.8576	0.2748	0.2070	0.2135
3	0.4801	0.4462	0.5769	0.9181	0.7061	0.8514	0.1032	0.1144
4	0.8563	0.7657	1.9924	0.9597	0.7910	0.4015	0.1829	0.1630
5	0.5993	0.0015	0.3193	0.5165	0.7504	0.4333	0.1840	0.1054
6	0.5384	0.0000	0.7551	0.7177	0.5285	0.4993	0.1659	0.1289
7	0.1356	0.1584	1.1910	0.8810	0.9107	0.2841	0.0469	0.1044
8	0.0157	0.0000	0.0000	0.0000	0.0313	0.0000	0.0000	0.0000
9	0.6668	0.5148	0.0000	0.6876	0.6772	0.1845	0.0047	0.1198
10	0.7042	0.1462	0.3506	0.5406	0.7116	0.8367	0.1840	0.1091
11	0.4578	0.5303	1.0177	0.8348	0.8200	0.3622	0.0431	0.0520
12	0.4066	0.0000	1.4643	0.8518	0.7320	0.0711	0.0896	0.0739
13	0.9949	0.0000	0.9130	1.0616	1.1566	0.2995	0.1065	0.2719
14	0.4753	0.2274	0.5153	0.8016	0.9415	0.3332	0.0454	0.0810
Team	0.1084	0.4744	0.3325	0.1658	0.4168	0.9120	0.0938	0.1564

Table 5
Prediction results.

NO	Playing time	Two point	Three point	Free throw	Defensive rebound	Assist	Steal	Turnover	Personal foul
1	8.85	0.5718	0.4654	1.2037	2.1236	0.5111	0.4338	1.0751	1.1294
2	6.17	1.1871	0.5519	0.3877	1.7850	0.3689	0.2953	1.2611	1.3228
3	30.03	1.9914	0.7569	1.5599	2.8723	1.9669	0.4406	1.1151	1.2073
4	20.91	1.7947	0.7817	0.9012	1.6558	1.1409	0.4057	1.2239	1.2725
5	22.22	2.9435	0.5309	0.7143	6.4963	0.3646	0.2623	1.2255	1.1418
6	34.89	6.3526	0.5380	0.6506	8.5573	6.5100	0.2767	1.1989	1.1874
7	13.72	2.2499	0.7348	0.8924	1.2020	0.7332	0.3963	1.0493	1.1681
8	6.52	1.5126	0.5380	0.9476	2.2579	1.3194	0.3274	1.0000	1.0015
9	4.88	0.6998	0.4940	1.6741	1.0619	0.5944	0.3886	1.0047	1.1448
10	18.10	1.3936	0.6427	1.6880	3.5451	2.0619	0.3242	1.2254	1.1634
11	18.08	2.7308	0.9104	1.5089	1.5352	0.9894	0.5261	1.0450	1.1713
12	12.19	0.9638	0.5380	0.9537	2.1658	1.9237	0.3326	1.0984	1.0906
13	17.25	1.4206	0.5282	0.9823	1.3625	0.8511	0.4298	1.1191	1.4382
14	26.21	2.1551	0.8065	1.6957	2.8370	1.8388	0.3907	1.0476	1.1570
Team	240	27.9672	8.8173	15.7602	39.4579	21.1742	5.2301	15.6892	16.5961
	Winning probability			73.95%		Expected number of wins			60.64

be multiple. The solutions may be changed once we add some additional constraints. For example, in many real applications and for the sake of popularity, the coach prefers to arrange some players to others. Specially, those top-class players have many fans all over the world and their show time will favor the ticket revenue, and accordingly the coach will allocate more playing time to those top-class players than others. Here we simply consider the case discussed previously in which these five starters are expected to play longer than those substitutes. As declared by the Golden State Warriors, Klay Thompson, Andrew Bogut, Draymond Green, Stephen Curry and Harrison Barnes are starters for the 2015–16 regular season. Through solving model (8) again, we obtain new prediction results in Table 6. As shown in Table 6, the predicted winning probability will remain unchanged, but the allocation of playing time and according game outcomes achieved by different players would be largely different. Again, we focus on the valuable players with most scores, the rank would be Stephen Curry, Anderson Varejao, Draymond Green, Harrison Barnes, and Andrew Bogut. This result is very similar with the one obtained previously.

In addition, we will compare our prediction results with some other predictive methods. Although many academics have built sport performance prediction approaches in NBA, a common disadvantage is low prediction accuracy [33] and none of these approaches is widely accepted. Here we consider two prediction results that are widely used by sports industry practitioners and sport fans. One is provided by Entertainment Sports Programming Network (EPSN, www.espn.com) and the other is provided by FiveThirtyEight (fivethirtyeight.com/sports). ESPN uses a forward-looking measure called Basketball Power Index (BPI) to measure team quality, which uses advanced statistical analysis to measure each team's offensive and defensive levels relative to an average team. The BPI can be used to predict a given team's average scores and winning probability. It is announced that ESPN's BPI is one of the most successful prediction approaches and it has won over 72% of NBA games.[4] On the contrary, FiveThirtyEight combines an Elo-based model with the so-called CARMELO player projections (a system that compares current NBA players with similar players throughout the league history) to predict NBA teams' "CARM-Elo" rating, expected number of win-to-loss and possibility of advancing to the playoffs (and beyond). Results of the two methods can be found online, but limited information on its prediction models and technical parameters is known, so we only compare prediction results on mathematical values.

We got prediction results of ESPN[5] with 60 wins and 22 losses. Interestingly, identical pre-season results were predicted for Golden State Warriors by FiveThirtyEight,[6] with also 60 wins and 22 losses being predicted. The comparison results are shown in Table 7. First of all, it is no wonder that all three predictions including this paper underestimated the winning probability and

[4] www.espn.com/nba/story/_/id/13984129/what-espn-nba-basketball-power-index.

[5] https://www.espn.com/nba/story/_/id/13980821/projected-standings-2015-16.

[6] https://projects.fivethirtyeight.com/2016-nba-picks/. We take the pre-season prediction by setting the prediction date as October 27th 2015, which is at the beginning of 2015–2016 regular reason.

Please cite this article as: Y. Li, L. Wang and F. Li, A data-driven prediction approach for sports team performance and its application to National Basketball Association, Omega, https://doi.org/10.1016/j.omega.2019.102123

Table 6
Prediction results with starters playing longer.

NO	Playing time	Two point	Three point	Free throw	Defensive rebound	Assist	Steal	Turnover	Personal foul
1	6.70	0.6039	0.3981	0.9399	2.0658	0.5404	0.3210	1.0751	1.1299
2	6.35	1.4567	0.5221	0.3572	2.1773	0.3689	0.2724	1.2611	1.3439
3	23.03	1.9285	0.5470	0.9604	2.5707	1.9999	0.4499	1.1151	1.1995
4	25.86	1.9601	0.9063	1.0122	2.2950	1.6173	0.4565	1.2239	1.2712
5	24.60	2.8777	0.4956	0.7241	6.4879	0.3847	0.2700	1.2255	1.1423
6	24.27	4.1598	0.4989	0.7670	5.3876	4.4861	0.3535	1.1989	1.1776
7	12.14	2.0857	0.6628	0.8220	1.2115	0.7976	0.4015	1.0493	1.1676
8	8.32	1.6941	0.4989	0.9659	2.7016	0.7767	0.2865	1.0000	1.0015
9	4.46	1.1188	0.6963	1.0995	1.3072	0.7701	0.3159	1.0047	1.1451
10	24.68	1.8526	0.7775	2.6181	4.2872	2.8979	0.3725	1.2254	1.1637
11	25.82	3.8824	1.0620	1.9393	2.0970	1.4498	0.5424	1.0450	1.1717
12	12.00	0.9390	0.4989	0.9999	2.2389	1.9224	0.3069	1.0984	1.0909
13	15.09	1.2545	0.4975	0.8598	1.3869	0.9126	0.4820	1.1191	1.4367
14	26.70	2.1533	0.7552	1.6948	3.2434	2.2497	0.3993	1.0476	1.1544
Team	240	27.9672	8.8173	15.7602	39.4579	21.1742	5.2301	15.6892	16.5961
	Winning probability			73.95%		Expected number of wins			60.64

Table 7
Different predictions for Golden State Warriors in the 2015–2016 season.

Indicators	This paper	ESPN	FiveThirtyEight	Real results
Expected number of wins	60.64	60	60	73
Expected number of losses	21.36	22	22	9
Winning probability (%)	73.95	73.17	73.17	89.02

expected number of wins, as Golden State Warriors broke the record in NBA to obtain 73 wins. Further, it can be seen that our prediction is extremely closed to the prediction results provided by ESPN and FiveThirtyEight, with our approach only having almost one more win. This finding demonstrates that our predictive approach has a similar efficacy as that of ESPN and FiveThirtyEight, and even better in the 2015–2016 reason considering the fact that Golden State Warriors made a new record. Since both ESPN and FiveThirtyEight are good examples of successful predictive approaches, through the comparison analysis we can conclude that the proposed DEA-based data-driven approach has relatively good efficacy and effectiveness in sports team performance prediction.

4.3. Discussion and analysis

In the previous section, we have taken Golden State Warriors as an example to illustrate the usefulness and efficacy of our proposed data-driven prediction approach. Here in the following part we will provide a further discussion on the proposed prediction approach and its application results, with the purpose of showing how it can be used to gain insight into the problem. In particular, we address the prediction of inefficiency sensitivity, different data sample, least promising performance, critical player analysis and prediction with stochastic inefficiency slacks.

Note that in model (8) we conduct a deterministic best performance prediction by assuming the inefficiency ratio to be constant for all outputs. However, there will be some uncertainty. Here we consider the uncertainty in the inefficiency ratio for players, team as a whole, and both the team and players simultaneously. The sensitivity analysis results in terms of the change percentage of inefficiency ratio are shown in Fig. 3. We find that the optimal winning probability is more sensitivity to the change of team's inefficiency ratio as compared with players. This is due to the fact that for collective sports like NBA, not all players play well or bad in each game at the same time. In fact, in almost all games some players will play well whereas the other players will play bad. As a result, the team as a whole will obtain game outcomes derived from well played players and badly played players, simultaneously. And consequently, when we impose the team's production possible set and production frontier constraint, the team's inefficiency ratio is a sensitive factor whereas the players' inefficiency ratio would be less sensitive. From this perspective, the manager and coach of Golden State Warriors are suggested to pay more attention to the team cooperation and communication and try to eliminate much more team inefficiency.

It is clear that the proposed prediction approach is data-driven and the results would be data-based, so different data sample may lead to different results. The results obtained previously are based on a data sample of four-season from the 2011–12 season to the 2014–15 season. By using the most recent 2014–15 season, here we would consider another three samples with only one season, two seasons and three seasons, respectively. We solve the proposed prediction approach in the same way, and the prediction results are given in Table 8. It shows that the predicted winning probability will decrease in the sample size. This is due to the fact that the Golden State Warriors won more and more in the regular season from 2011 to 2015, and its average inefficiency ratio will be smaller and smaller. The winning probability under different data samples is shown in Fig. 4, in which it involves the real winning probability for each individual season, the accumulative winning probability and predicted winning probability based on different samples. And consequently, the predicted results would also be better and better. In particular, the result based on the two seasons of 2013–14 and 2014–15 is very near the real results in the 2015–16 regular season in which Golden State Warriors broke the record in NBA to obtain 73 wins.

Fig. 5 shows the playing time for each player under different data samples. It can be seen that although there exist some variations, players show a main trend, implying that the playing time with the optimal team winning probability is relatively stable.

Previously we use the production frontier method to predict the future performance, and such a prediction is the most promising one. Here we would address the least performance given also the average inefficiency ratio calculated by model (5) and formula (6) and (7) and shown in Table 4. For this purpose, we would first

Please cite this article as: Y. Li, L. Wang and F. Li, A data-driven prediction approach for sports team performance and its application to National Basketball Association, Omega, https://doi.org/10.1016/j.omega.2019.102123

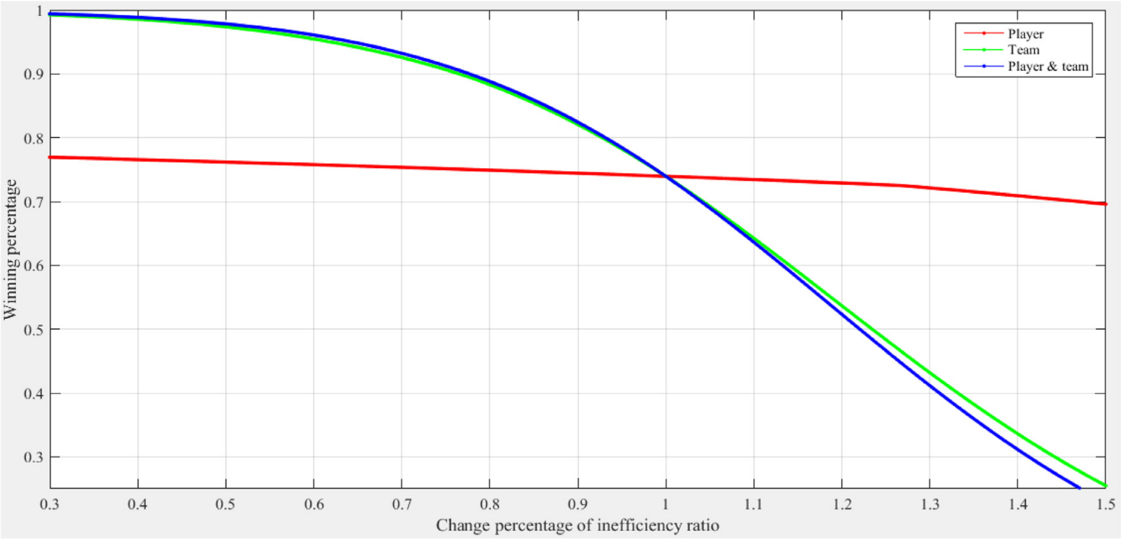

Fig. 3. Sensitivity analysis of inefficiency.

Table 8
Prediction results under different data samples.

	2014–15	2013–14 to 2014–15	2012–13 to 2014–15	2011–12 to 2014–15
Two point	28.7977	28.2326	28.1232	27.9672
Three point	10.9775	9.9907	9.1520	8.8173
Free throw	16.4336	16.4567	16.2527	15.7602
Defensive rebound	41.7886	40.6424	40.5134	39.4579
Assist	24.6235	22.3997	21.5429	21.1742
Steal	7.0134	6.0619	5.3653	5.2301
Turnover	15.3585	15.7692	15.9202	15.6892
Personal foul	15.8805	16.3251	16.4339	16.5961
Winning probability	96.58%	89.47%	82.12%	73.95%
Expected number of wins	79.19	73.35	67.34	60.64

divide all decision making units (here refer to the games played by players or the team) for each player j (or the team) into two sets based on model (5), with set E_j for efficient DMUs and set I_j for inefficient DMUs. And consequently, we use the following model (10) to approach the issue of least winning probability prediction.

$$\begin{array}{ll} Min \;\; \hat{f} = \beta_0^* + \beta_1^* \hat{x}_1 + \cdots + \beta_s^* \hat{x}_s & \\ s.t. \sum_{j=1}^n t_j = n_0 \cdot t_0 & \\ \quad \sum\limits_{k \in E_j} \lambda_{kj} t_j^k = t_j, \forall j = 1, \ldots, n & \\ \quad \sum\limits_{k \in E_j} \lambda_{kj} x_{rj}^k = x_{rj} + \rho_{rj}^{+*} x_{rj}, \forall r = 1, \ldots, m;\; j = 1, \ldots, n & \\ \quad \sum\limits_{k \in E_j} \lambda_{kj} x_{rj}^k = x_{rj} - \rho_{rj}^{-*} x_{rj}, \forall r = m+1, \ldots, s;\; j = 1, \ldots, n & \\ \quad \sum\limits_{k \in E_j} \lambda_{kj} = 1, \forall j = 1, \ldots, n & (10) \\ \quad \sum_{j=1}^n x_{rj} = \hat{x}_r, \forall r = 1, \ldots, s & \\ \quad \sum\limits_{l \in E} \lambda_l x_r^l = \hat{x}_r + \rho_r^{+*} \hat{x}_r, \forall r = 1, \ldots, m & \\ \quad \sum\limits_{l \in E} \lambda_l x_r^l = \hat{x}_r - \rho_r^{-*} \hat{x}_r, \forall r = m+1, \ldots, s & \\ \quad \sum\limits_{l \in E} \lambda_l = 1 & \\ \quad 0 \le t_j \le t_0, \forall j = 1, \ldots, n & \\ \quad \lambda_{kj}, \lambda_l \ge 0, \forall j = 1, \ldots, n;\; \forall k \in P_j;\; l = 1, \ldots, q. & \end{array}$$

Table 9 list the least prediction results based on model (10). We find that the predicted least performance is very different from the most promising result given in Table 5. Although all players are characterized with the average inefficiency ratios, different planning will lead to different results, which even differ a lot. Such a finding demonstrates the phenomenon that the appropriate selection of players and the allocation of playing time in the court are of most significant importance for NBA teams. Also, the strategic arrangement to obtain different combinations of these desirable and undesirable outputs is closely related with the winning probability. Further, note that Stephen Curry almost will not participate in games under the worst situation, which may due to the fact that he is a well-played basketball player and freezing Stephen Curry out will cause very bad results.

In many sport events there exists a “hot player” phenomenon which indicating that some players will perform superior to other players. Also, almost in each team in all sports there will exist one, two or more critical players that contribute the most. From this perspective, we are interested in discussing the impact of absence of some players, after all, the injury problem is inevitable for almost all players. Without loss of generality, we think that there exist some critical players who will contribute much more if other players are out of function. An issue of considerable significance is that how many changes would occur in the future performance if a player is injured and will not play any game in the prediction period, and accordingly, how to reallocation the playing time. Towards this end, we consider situations where a certain player is not here across all the next season, and the results are given in Table 10. According to the results, we find that by remaining the average inefficiency ratios for the team and players, the optimal predicted winning probability will be unchanged whoever is absent. One possible reason for this result would be that, within

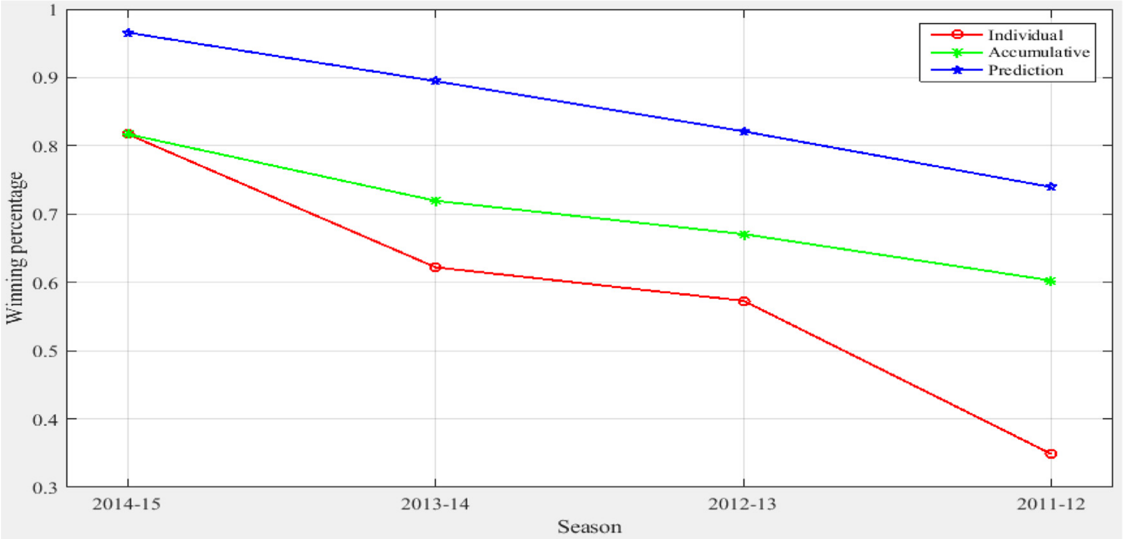

Fig. 4. Winning probability under data different samples.

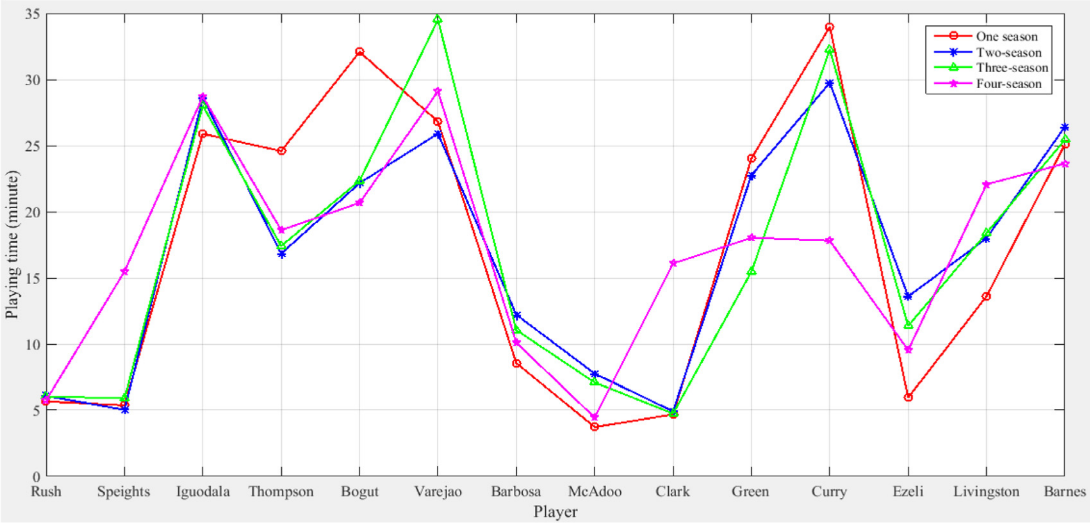

Fig. 5. The playing time for each player under different data samples.

Table 9
Least prediction results by remaining average inefficiency ratios.

NO	Playing time	Two point	Three point	Free throw	Defensive rebound	Assist	Steal	Turnover	Personal foul
1	30.68	2.7028	0.6917	1.7740	6.2186	2.6458	0.8371	2.1502	1.1091
2	36.79	6.5000	0.9909	0.7758	4.6284	2.0242	1.0705	1.2611	4.3604
3	28.43	2.7025	0.6914	1.2683	3.6495	4.6891	1.0802	1.1151	1.1292
4	10.83	1.6161	0.5663	0.3342	1.5309	2.7917	0.7135	1.2239	1.1948
5	31.92	3.7517	0.9985	0.7580	3.9564	2.8565	0.6977	1.2255	1.1178
6	9.33	1.9501	1.0000	0.5698	1.7465	2.6170	0.6670	1.1989	1.1480
7	11.07	3.6875	0.8633	1.0020	0.7435	0.9404	0.7788	1.8854	1.5614
8	9.22	1.9691	1.0000	1.0000	3.0000	2.9088	1.0000	1.0000	1.0000
9	13.87	0.6000	0.6602	3.5699	1.0163	1.6188	0.8443	1.3639	3.0020
10	3.38	0.5868	0.8724	0.9202	1.4372	1.2936	0.5445	1.5230	1.1225
11	0.05	0.6860	0.6535	0.4956	0.5450	0.5495	0.7341	1.0450	1.0549
12	16.37	2.1329	1.0000	1.6232	2.1600	2.3094	0.9336	1.0984	1.0797
13	24.05	2.5064	1.0000	3.6592	0.4851	0.4637	0.7695	1.1191	4.1200
14	14.01	0.8142	0.8147	1.0582	0.8900	0.9295	1.3493	1.8845	1.9574
Team	240	29.3631	8.9602	15.9655	29.1645	25.7953	9.1774	16.2513	24.2827
	Winning probability			40.48%		Expected number of wins			33.19

Please cite this article as: Y. Li, L. Wang and F. Li, A data-driven prediction approach for sports team performance and its application to National Basketball Association, Omega, https://doi.org/10.1016/j.omega.2019.102123

Table 10
Critical player analysis.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Original	8.85	6.17	30.03	20.91	22.22	34.89	13.72	6.52	4.88	18.10	18.08	12.19	17.25	26.21
1		6.30	30.24	19.68	22.02	34.49	13.21	8.75	4.54	20.00	25.07	12.72	16.85	26.12
2	8.67		30.02	20.03	22.19	34.77	13.26	8.63	4.75	20.80	20.30	13.06	16.98	26.55
3	9.24	6.75		24.40	22.49	35.15	14.33	8.69	5.12	20.07	36.35	12.76	18.03	26.63
4	10.32	6.65	31.38		22.49	34.93	14.15	8.27	5.06	19.43	29.97	12.47	18.04	26.86
5	7.83	6.35	30.77	21.28		35.30	13.60	8.71	4.59	18.86	34.76	13.99	17.04	26.92
6	7.15	6.77	33.19	29.98	22.00		13.45	8.68	4.54	17.32	33.11	13.78	21.06	28.98
7	9.25	6.41	30.35	21.82	22.09	34.92		7.77	4.91	17.92	28.66	12.11	17.47	26.32
8	8.53	6.02	30.03	19.41	22.13	34.56	12.69		4.42	17.42	29.23	12.72	16.44	26.42
9	7.78	6.17	29.67	18.89	22.15	34.61	13.12	8.53		26.04	17.51	12.54	16.75	26.22
10	9.97	6.22	30.86	22.18	22.14	34.92	13.03	8.57	4.56		30.26	12.63	17.04	27.60
11	11.17	6.40	31.24	24.24	22.09	34.76	14.08	7.81	4.78	26.13		12.11	18.23	26.96
12	9.17	6.37	30.89	22.24	22.08	33.67	13.65	8.64	4.72	18.22	26.17		17.37	26.79
13	9.29	6.39	30.32	22.12	22.22	34.94	13.81	8.40	5.01	17.61	31.69	11.86		26.34
14	12.30	6.51	31.00	23.54	22.18	34.98	13.79	8.72	4.84	20.36	31.41	12.93	17.44	
Average change (%)	0.44	0.24	0.74	1.39	−0.04	−0.12	−0.16	1.95	−0.12	1.91	10.72	0.55	0.35	0.62

Table 11
The player increasing the most playing time under two players' absences.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	11	11	11	11	11	11	11	11	11	11	10	11	11	11
2		10	11	11	11	11	11	11	11	11	1	11	11	11
3			11	11	11	11	11	11	11	11	4	11	11	11
4				11	11	11	11	11	11	11	1	11	11	11
5					11	11	11	11	11	11	4	11	11	11
6						11	11	11	11	11	4	11	11	11
7							11	11	11	11	10	11	11	11
8								11	11	11	4	11	11	11
9									10	11	1	11	11	11
10										11	1	11	11	11
11											10	10	4	1
12												11	11	11
13													11	11
14														11

Table 12
Prediction results with stochastic inefficiency ratios.

Times	100	200	500	1000	2000	5000
Two-point	30.3457	29.4949	28.3304	29.4165	28.6462	27.7977
Three-point	11.3040	7.2992	9.5504	8.9233	9.3790	9.1923
Free throw	15.7380	17.8335	16.8418	16.0301	14.8009	15.4075
Defensive rebound	29.8773	37.8587	36.4901	38.3319	39.8781	38.6982
Assist	21.1111	23.8053	20.0428	23.3757	19.9083	21.4906
Steal	9.1059	6.4165	6.6939	5.5641	6.2919	5.9038
Turnover	17.8169	17.6964	16.6332	18.1679	18.2946	16.4729
Personal foul	21.6431	20.1347	19.7301	17.8731	17.7174	17.6573
Winning probability	56.45%	65.89%	64.32%	72.98%	71.72%	70.65%
Expected number of wins	46.29	54.03	52.74	59.85	58.81	57.93

DEA-based approaches different DMUs are homogenous and can be substituted with a convex combination of peer DMUs. And as a result, anyone's absence can be offset by others, and the winning probability keep unchanged. Another reason comes from the fact that except for the PPS of each players, we have used also the team's PPS to restrict the predicted possible input-output. Since NBA is a team sport and the players of Golden State Warriors in each year change a lot, the gap between the team's PPS and the sum of the players' PPS would be large. When we use model (8) to address the winning probability prediction, there may exist some more outputs that can be achieved by the sum of players but not by the team. As a result, if such kind of redundant outputs exists, the winning probability would be unchanged no matter which player cannot participate in the games.

However, if we investigate the change of playing time, we would find that Stephen Curry's playing time would increase the most. Table 11 shows the player who will increase the most playing time if two players are removed simultaneously, and the results show that in almost all cases Stephen Curry is the target. The one who increases the most can be supposed as the critical player, since by playing longer he would try his best to offset the negative effect of one's absence and achieve the optimal game outcomes for the team. Note in addition that previously we find Stephen Curry is one of the top five valuable player based on possible scores and game outcomes, therefore, here we may conclude that according to our data-driven prediction approach and the data sample used in this paper, Stephen Curry is the most valuable player for the Golden State Warriors. In fact, Stephen Curry got the honor of most valuable player (MVP) in the 2015–16 regular season of NBA.

Further, we would consider such a situation in which both the individual inefficiency ratio on each outcome for all players and

Please cite this article as: Y. Li, L. Wang and F. Li, A data-driven prediction approach for sports team performance and its application to National Basketball Association, Omega, https://doi.org/10.1016/j.omega.2019.102123

Table 13
Percentage of outputs outperform the average value.

NO	Two point	Three point	Free throw	Defensive rebound	Assist	Steal	Turnover	Personal foul
1	25.36%	47.83%	25.36%	34.06%	57.25%	30.43%	49.28%	44.93%
2	40.82%	5.10%	46.94%	36.73%	44.22%	23.13%	71.77%	57.14%
3	59.22%	29.43%	45.39%	51.42%	48.94%	44.33%	65.60%	47.16%
4	53.92%	49.67%	53.59%	49.67%	37.58%	40.52%	60.13%	45.75%
5	42.13%	0.56%	25.84%	46.07%	39.89%	47.19%	56.74%	41.01%
6	39.72%	0.00%	51.77%	48.94%	46.10%	37.59%	54.61%	62.41%
7	40.84%	47.64%	33.51%	36.65%	42.93%	40.84%	87.96%	40.84%
8	33.33%	0.00%	40.00%	46.67%	40.00%	13.33%	53.33%	66.67%
9	35.85%	33.96%	13.21%	35.85%	41.51%	39.62%	100%	66.04%
10	49.17%	44.58%	31.67%	47.50%	47.92%	33.75%	52.50%	69.17%
11	48.47%	40.84%	45.04%	41.98%	40.08%	47.33%	85.11%	58.40%
12	33.06%	0.00%	40.32%	36.29%	50.00%	21.77%	38.71%	52.42%
13	45.68%	1.08%	38.85%	53.96%	45.32%	37.77%	70.50%	62.95%
14	48.13%	57.26%	51.04%	43.57%	47.30%	37.76%	80.50%	69.29%
Team	42.76%	42.42%	50.17%	45.45%	45.12%	52.53%	45.22%	46.46%

that ratio of the team as a whole are determined randomly. In other words, note that the results in Table 4 are average values based on model (5) and formula (6) and (7), here we assume that the inefficiency ratios are randomly selected from those of previous games in the sample. As a result, the prediction result would be also changed randomly, as given in Table 12. We find that the stochastic prediction results are more likely to be lower than that of the deterministic case, this may due to the fact that for those desirable (undesirable) outputs only a very low percentage of all games for these players is higher (lower) than its average value. For details, we can see the statistics given in Table 13, which shows that almost all percentage values are less than 0.5 for desirable outputs and more than 0.5 for undesirable outputs.

5. Conclusions

Performance evaluation has become one of the most important tasks in the competitive environment. A well-defined prediction method has significant impacts on production planning, resource allocation, revenue management, and so on. In this paper, we propose a two-step data-driven approach based on data envelopment analysis to predict the winning probability in team sports like National Basketball Association. It first uses a multiple statistics regression analysis to estimate the quantity relationship between the winning probability and various game outcomes at the team-level, and then applies DEA-based production frontier models to obtain the optimal game outcomes. Given the average inefficiencies, the predictive result is generated by optimally allocating playing time across players and planning possible game outcomes. Through the application to Golden State Warriors in NBA, we find that the proposed approach has a good prediction power.

This paper has provided an insight to develop DEA-based data-driven prediction approaches for sports team performance prediction, and the proposed approach can be considered as a reference and benchmark for future research on the same objective, more specifically DEA-based approaches in the sport performance prediction. Meanwhile, it can be extended from some directions. First, we only applied the proposed approach to one NBA team, so in future we will try to illustrate it with more comprehensive data and situations. This point would be of particular attraction in the big data context. Besides, note that we don't take the play-by-play agenda and competitive strategies of opponents into account, it would be of significance to integrate more behavior theories into the data-driven prediction approach. Also, it is an interesting and promising problem of predicting the individual game, for which we should take more factors such as player limitation, player portfolio and playing time planning into account. Future research can develop approaches for sports team performance prediction at the game level. In addition, the proposed prediction approach adopts the average inefficiency ratios to quantify the efficiency status in the prediction period. This is a natural and feasible way, but other ways are also possible to approach the data-driven prediction problems. And we should note that an accurate estimation of the possible efficiency status in the prediction period is very important for the prediction approaches and results. Further, since the DEA method is extremely optimistic, the prediction results may be exaggerated and too sanguine. A possible reason depends on the production frontier, which might be closely associated with games with weak opponents. A possible research would be designed to overcome this drawback and obtain better results. Last, we will also try to develop a faster and easier tool that can be used in real-time in the staff of the teams, which might be implemented through creating a latent variable or some multivariate techniques in a comprehensive model.

Acknowledgments

The authors would like to thank the editor and three anonymous reviewers for their kind work and insightful comments and suggestions. This research was financially supported by the National Natural Science Foundation of China (Nos. 71901178, 71910107002, 71901225, 71671172, 71631006) and the Fundamental Research Funds for the Central Universities at Southwestern University of Finance and Economics (Nos. JBK1901013 and JBK190504).

References

[1] Aizemberg L, Roboredo MC, Ramos TG, de Mello JCCS, Meza LA, Alves AM. Measuring the NBA teams' cross-efficiency by DEA game. Am J Oper Res 2014;4(3):101–12.
[2] Atkinson G, Nevill AM. Selected issues in the design and analysis of sport performance research. J Sports Sci 2001;19(10):811–27.
[3] An Q, Wen Y, Chu J, Chen X. Profit inefficiency decomposition in a serial-structure system with resource sharing. J Oper Res Soc 2019. doi:10.1080/01605682.2018.1510810.
[4] An Q, Wen Y, Ding T, Li Y. Resource sharing and payoff allocation in a three-stage system: integrating network DEA with the shapley value method. Omega 2019;85:16–25.
[5] An Q, Wang Z, Emrouznejad A, Zhu Q, Chen X. Efficiency evaluation of parallel interdependent processes systems: an application to Chinese 985 project universities. Int J Prod Res 2019;57(17):5387–99. doi:10.1080/00207543.2018.1521531.
[6] Andrade A, Bevilacqua GG, Coimbra DR, Pereira FS, Brandt R. Sleep quality, mood and performance: a study of elite Brazilian volleyball athletes. J Sports Sci Med 2016;15(4):601–5.
[7] Ang S, Yang C, Zhao F, Yang F. Ranking of DMUs with interval cross-efficiencies based on absolute dominance. Int J Inform Decis Sci 2016;8(4):325–40.

Please cite this article as: Y. Li, L. Wang and F. Li, A data-driven prediction approach for sports team performance and its application to National Basketball Association, Omega, https://doi.org/10.1016/j.omega.2019.102123

[8] Amatria M, Lapresa D, Arana J, Anguera MT, Garzón B. Optimization of game formats in U-10 soccer using logistic regression analysis. J Hum Kinet 2016;54(1):163–71.
[9] Arabi B, Munisamy S, Emrouznejad A. A new slacks-based measure of Malmquist–Luenberger index in the presence of undesirable outputs. Omega 2015;51:29–37.
[10] Asghar F, Asif M, Nadeem MA, Nawaz MA, Idrees M. A novel approach to ranking National Basketball Association players. J Glob Econ, Manag Bus Res 2018;10(4):176–83.
[11] Baghal T. Are the "Four factors" indicators of one factor? An application of structural equation modeling methodology to NBA data in prediction of winning percentage. J Quant Anal Sports 2012;8(1):1–17.
[12] Banker RD, Charnes A, Cooper WW. Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manag Sci 1984;30(9):1078–92.
[13] Berman SL, Down J, Hill CW. Tacit knowledge as a source of competitive advantage in the National Basketball Association. Acad Manag J 2002;45(1):13–31.
[14] Chalitsios C, Nikodelis T, Panoutsakopoulos V, Chassanidis C, Kollias I. Classification of soccer and Basketball players' jumping performance characteristics: a logistic regression approach. Sports 2019;7(7):163.
[15] Chang TM, Hsu MF, Lin SJ. Integrated news mining technique and AI-based mechanism for corporate performance forecasting. Inf Sci 2018;424:273–86.
[16] Chang TS, Tone K, Wu CH. DEA models incorporating uncertain future performance. Eur J Oper Res 2016;254(2):532–49.
[17] Charnes A, Cooper WW, Rhodes E. Measuring the efficiency of decision making units. Eur J Oper Res 1978;2(6):429–44.
[18] Chen Y, Gong Y, Li X. Evaluating NBA player performance using bounded integer data envelopment analysis. Inf Syst Oper Res 2017;55(1):38–51.
[19] Chu J, Wu J, Chu C, Zhang T. DEA-based fixed cost allocation in two-stage systems: leader-follower and satisfaction degree bargaining game approaches. Omega 2019. doi:10.1016/j.omega.2019.03.012.
[20] Clemente FM, Couceiro MS, Martins FM, Mendes R. An online tactical metrics applied to football game. Res J Appl Sci, Eng Technol 2013;5(5):1700–19.
[21] Constantinou AC, Fenton NE, Neil M. Pi-football: a bayesian network model for forecasting association Football match outcomes. Knowl Based Syst 2012;36:322–39.
[22] Constantinou AC, Fenton NE, Neil M. Profiting from an inefficient association football gambling market: prediction, risk and uncertainty using bayesian networks. Knowl Based Syst 2013;50(5):60–86.
[23] Constantinou A, Fenton N. Towards smart-data: improving predictive accuracy in long-term football team performance. Knowl Based Syst 2017;124:93–104.
[24] Cooper WW, Ruiz JL, Sirvent I. Selecting non-zero weights to evaluate effectiveness of basketball players with DEA. Eur J Oper Res 2009;195(2):563–74.
[25] Couceiro MS, Dias G, Araújo D, Davids K. The ARCANE project: how an ecological dynamics framework can enhance performance assessment and prediction in football. Sports Med 2016;46(12):1781–6.
[26] Daraio C, Simar L. Efficiency and benchmarking with directional distances: a data-driven approach. J Oper Res Soc 2016;67(7):928–44.
[27] du Jardin P. Failure pattern-based ensembles applied to bankruptcy forecasting. Decis Support Syst 2018;107:64–77.
[28] Farrell MJ. The measurement of productive efficiency. J R Stat Soc Ser A 1957;120(3):253–90.
[29] Feddersen A, Humphreys BR, Soebbing BP. Sentiment bias in national basketball association betting. J Sports Econ 2018;19(4):455–72.
[30] Fukuyama H, Weber WL. A directional slacks-based measure of technical inefficiency. Socioecon Plann Sci 2009;43(4):274–87.
[31] Gamble D, Bradley J, McCarren A, Moyna NM. Team performance indicators which differentiate between winning and losing in elite Gaelic football. Int J Perform Anal Sport 2019;19(4):478–90.
[32] Gómez M, Ibáñez S, Parejo I, Furley P. The use of classification and regression tree when classifying winning and losing basketball teams. Kinesiol: Int J Fundam Appl Kinesiology 2017;49(1):47–56.
[33] Haghighat M, Rastegari H, Nourafza N. A review of data mining techniques for result prediction in sports. Adv Comput Sci: An Int J 2013;2(5):7–12.
[34] Hofler RA, Payne JE. Measuring efficiency in the national basketball association. Econ Lett 1997;55(2):293–9.
[35] Hollinger J. Pro basketball forecast, 2005-06. Potomac Books; 2005.
[36] Huang CL. The construction of production performance prediction system for semiconductor manufacturing with artificial neural networks. Int J Prod Res 1999;37(6):1387–402.
[37] Katayama H, Nuch H. A game-level analysis of salary dispersion and team performance in the national basketball association. Appl Econ 2011;43(10):1193–207.
[38] Kester BS, Behery OA, Minhas SV, Hsu WK. Athletic performance and career longevity following anterior cruciate ligament reconstruction in the National Basketball Association. Knee Surg, Sports Trauma, Arthrosc 2017;25(10):3031–7.
[39] Koster J, Aven B. The effects of individual status and group performance on network ties among teammates in the National Basketball Association. PLoS ONE 2018;13(4):e0196013.
[40] Lam M. Neural network techniques for financial performance prediction: integrating fundamental and technical analysis. Decis Support Syst 2004;37(4):567–81.
[41] Lee YH, Berri D. A re-examination of production functions and efficiency estimates for the national basketball association. Scott J Polit Econ 2008;55(1):51–66.
[42] Lee BL, Worthington AC. A network DEA quantity and quality-orientated production model: an application to Australian university research services. Omega 2016;60:26–33.
[43] Lee BL, Worthington AC. A note on the 'Linsanity' of measuring the relative efficiency of National Basketball Association guards. Appl Econ 2013;45(29):4193–202.
[44] Leitner C, Zeileis A, Hornik K. Forecasting sports tournaments by ratings of (prob) abilities: a comparison for the EURO 2008. Int J Forecast 2010;26(3):471–81.
[45] Li F, Emrouznejad A, Yang GL, Li Y. Carbon emission abatement quota allocation in Chinese manufacturing industries: an integrated cooperative game data envelopment analysis approach. J Oper Res Soc 2019. doi:10.1080/01605682.2019.1609892.
[46] Li F, Liang L, Li Y, Emrouznejad A. An alternative approach to decompose the potential gains from mergers. J Oper Res Soc 2018;69(11):1793–802.
[47] Li F, Song J, Dolgui A, Liang L. Using common weights and efficiency invariance principles for resource allocation and target setting. Int J Prod Res 2017;55(17):4982–97.
[48] Li F, Zhu Q, Chen Z. Allocating a fixed cost across the decision making units with two-stage network structures. Omega 2019;83:139–54.
[49] Li F, Zhu Q, Chen Z, Xue H. A balanced data envelopment analysis cross-efficiency evaluation approach. Expert Syst Appl 2018;106:154–68.
[50] Li F, Zhu Q, Liang L. Allocating a fixed cost based on a DEA-game cross efficiency approach. Expert Syst Appl 2018;96:196–207.
[51] Li F, Zhu Q, Liang L. A new data envelopment analysis based approach for fixed cost allocation. Ann Oper Res 2019;274(1–2):347–72.
[52] Li F, Zhu Q, Zhuang J. Analysis of fire protection efficiency in the United States: a two-stage DEA-based approach. OR Spectr 2018;40(1):23–68.
[53] Li L, Chang Q, Ni J. Data driven bottleneck detection of manufacturing systems. Int J Prod Res 2009;47(18):5019–36.
[54] Li X, Li F, Zhao N, Zhu Q. Measuring environmental sustainability performance of freight transportation seaports in China: a data envelopment analysis approach based on the closest targets. Expert Syst 2018. doi:10.1111/exsy.12334.
[55] Lins MPE, Gomes EG, de Mello JCCS, de Mello AJRS. Olympic ranking based on a zero sum gains DEA model. Eur J Oper Res 2003;148(2):312–22.
[56] Liu JS, Lu LY, Lu WM, Lin BJ. A survey of DEA applications. Omega 2013;41(5):893–902.
[57] Liu M, Frangopol DM. Optimal bridge maintenance planning based on probabilistic performance prediction. Eng Struct 2004;26(7):991–1002.
[58] Mansoor MS, Sinah S. Technical and scale efficiency of all-time NBA leaders. J Glob Econ, Manag Bus Res 2018;10(3):158–65.
[59] Masoumzadeh A, Toloo M, Amirteimoori A. Performance assessment in production systems without explicit inputs: an application to basketball players. IMA J Manag Mathe 2016;27(2):143–56.
[60] Melnick MJ. Relationship between team assists and win-loss record in the National Basketball Association. Percept Mot Skills 2001;92(2):595–602.
[61] Min B, Kim J, Choe C, Eom H, McKay RB. A compound framework for sports results prediction: a football case study. Knowl Based Syst 2008;21(7):551–62.
[62] Moreno P, Lozano S. A network DEA assessment of team efficiency in the NBA. Ann Oper Res 2014;214(1):99–124.
[63] Moreno P, Lozano S. Estimation of productivity change of NBA teams from 2006-07 to 2012-13 seasons. Int J Sport Financ 2015;10(3):217–41.
[64] Moxley JH, Towne TJ. Predicting success in the National Basketball Association: stability & potential. Psychol Sport Exerc 2015;16:128–36.
[65] Radovanović S, Radojičić M, Sacić G. Two-phased DEA-MLA approach for predicting efficiency of NBA players. Yugosl J Oper Res 2014;24(3):347–58.
[66] Robertson S, Back N, Bartlett JD. Explaining match outcome in elite Australian rules football using team performance indicators. J Sports Sci 2016;34(7):637–44.
[67] Nyhuis P, Von Cieminski G, Fischer A, Feldmann K. Applying simulation and analytical models for logistic performance prediction. CIRP Ann-Manuf Technol 2005;54(1):417–22.
[68] Omar RSM, Venkatadri U, Diallo C, Mrishih S. A data-driven approach to multi-product production network planning. Int J Prod Res 2017;55(23):7110–34.
[69] Omondi-Ochieng P. Gross national income, football workers and national football team performances: a logistic regression analysis. Team Perform Manag 2015;21(7/8):405–20.
[70] Ouellette P, Vierstraete V. Technological change and efficiency in the presence of quasi-fixed inputs: a DEA application to the hospital sector. Eur J Oper Res 2004;154(3):755–63.
[71] Ouenniche J, Tone K. An out-of-sample evaluation framework for DEA with application in bankruptcy prediction. Ann Oper Res 2017;254(1–2):235–50.
[72] Podlog L, Buhler CF, Pollack H, Hopkins PN, Burgess PR. Time trends for injuries and illness, and their relation to performance in the National Basketball Association. J Sci Med Sport 2015;18(3):278–82.
[73] Premachandra IM, Chen Y, Watson J. DEA as a tool for predicting corporate failure and success: a case of bankruptcy assessment. Omega 2011;39(6):620–6.
[74] Ravi V, Kurniawan H, Thai PNK, Kumar PR. Soft computing system for bank performance prediction. Appl Soft Comput 2008;8(1):305–15.

[75] Saavedra-García M, Matabuena M, Montero-Seoane A, Fernández-Romero JJ. A new approach to study the relative age effect with the use of additive logistic regression models: a case of study of FIFA football tournaments (1908–2012). PLoS ONE 2019;14(7):e0219757.

[76] Sheppard RW. Theory of cost and production function. Princeton, NJ: Princeton University; 1970.

[77] Silva M, Marcelino R, Lacerda D, João PV. Match analysis in Volleyball: a systematic review. Monten J Sports Sci Med 2016;5(1):35–46.

[78] Spiteri T, Binetti M, Scanlan AT, Dalbo VJ, et al. Physical determinants of division 1 collegiate basketball, women's national basketball league, and women's National Basketball Association Athletes: with reference to lower-body sidedness. J Strength Cond Res 2019;33(1):159–66.

[79] Stekler HO, Sendor D, Verlander R. Issues in sports forecasting. Int J Forecast 2010;26(3):606–21.

[80] Sueyoshi T. Mixed integer programming approach of extended DEA discriminant analysis. Eur J Oper Res 2004;152(1):45–55.

[81] Tao F, Qi Q, Liu A, Kusiak A. Data-driven smart manufacturing. J Manuf Syst 2018;48(C):157–69.

[82] Thabtah F, Zhang L, Abdelhamid N. NBA game result prediction using feature analysis and machine learning. Ann Data Sci 2019;6(1):103–16.

[83] Tone K. A slacks-based measure of efficiency in data envelopment analysis. Eur J Oper Res 2001;130(3):498–509.

[84] Variyam PN, Cherubal S, Chatterjee A. Prediction of analog performance parameters using fast transient testing. IEEE Trans Comput-Aided Des Integr Circuits Syst 2002;21(3):349–61.

[85] Vaz de Melo PO, Almeida VA, Loureiro AA, Faloutsos C. Forecasting in the NBA and other team sports: network effects in action. ACM Transn Knowl Discov Data (TKDD 2012;6(3):1–27 no13.

[86] Villa G, Lozano S. Dynamic network DEA approach to basketball games efficiency. J Oper Res Soc 2018;69(11):1738–50.

[87] Vračar P, Štrumbelj E, Kononenko I. Modeling basketball play-by-play data. Expert Syst Appl 2016;44:58–66.

[88] Waguespack DM, Salomon R. Quality, subjectivity, and sustained superior performance at the olympic games. Manage Sci 2015;62(1):286–300.

[89] Wu DD, Luo C, Wang H, Birge JR. Bi-level programing merger evaluation and application to banking operations. Prod Oper Manage 2016;25(3):498–515.

[90] Xiao Z, Yang X, Pang Y, Dang X. The prediction for listed companies' financial distress by using multiple prediction methods with rough set and Dempster–Shafer evidence theory. Knowl Based Syst 2012;26:196–206.

[91] Yang CH, Lin HY, Chen CP. Measuring the efficiency of NBA teams: additive efficiency decomposition in two-stage DEA. Ann Oper Res 2014;217(1):565–89.

[92] Yang G, Leicht AS, Lago C, Gómez MÁ. Key team physical and technical performance indicators indicative of team quality in the soccer Chinese super league. Res Sports Med 2018;26(2):158–67.

[93] Yang JB, Wong BYH, Xu DL, Liu XB, Steuer RE. Integrated bank performance assessment and management planning using hybrid minimax reference point-DEA approach. Eur J Oper Res 2010;207(3):1506–18.

[94] Yu C, Matta A. A statistical framework of data-driven bottleneck identification in manufacturing systems. Int J Prod Res 2016;54(21):6317–32.

[95] Zhang L, Chu X, Chen H, Yan B. A data-driven approach for the optimisation of product specifications. Int J Prod Res 2019;57(3):703–21.

[96] Zhang Q, Wang C. DEA efficiency prediction based on IG–SVM. Neural Comput Appl 2018. doi:10.1007/s00521-018-3904-4.

[97] Zhou P, Ang BW, Wang H. Energy and CO_2 emission performance in electricity generation: a non-radial directional distance function approach. Eur J Oper Res 2012;221(3):625–35.

[98] Zhu Q, Wu J, Ji X, Li F. A simple MILP to determine closest targets in non-oriented DEA model satisfying strong monotonicity. Omega 2018;79:1–8.

[99] Yin P, Chu J, Wu J, Ding J, Yang M, Wang Y. A DEA-based two-stage network approach for hotel performance analysis: An internal cooperation perspective. Omega 2019. doi:10.1016/j.omega.2019.02.004.

Please cite this article as: Y. Li, L. Wang and F. Li, A data-driven prediction approach for sports team performance and its application to National Basketball Association, Omega, https://doi.org/10.1016/j.omega.2019.102123