An Olympic prediction approach using data envelopment analysis

Yongjun Li[a]

Email: lionli@ustc.edu.cn

Shen Jiang[a]

Email: 1353073931@qq.com

Lizheng Wang*[b1]

Email: lzwang@mail.ustc.edu.cn

[a]School of Management, University of Science and Technology of China, Hefei, P.R. China

[b]International Institute of Finance, School of Management, University of Science and Technology of China, Hefei, P.R. China

[1] *corresponding author

An Olympic prediction approach using data envelopment analysis

Abstract:

Predicting the number of medals a country will win at the Olympics is important for policymakers and commercial activities. In this paper, we use a data-driven prediction method based on data envelopment analysis (DEA) to perform Olympic prediction in two steps. The first step estimates the production function of the number of medals and various Olympic outcomes through negative binomial regression. The second step uses DEA to allocate investment optimally for individual sports and to obtain optimal Olympic predictions. Notably, our approach considers two frontiers, one for each sport, and another for each country's overall outcomes. We trained the proposed approach using Olympic Games datasets from 2008, 2012, and 2016 to predict the performance of China and other countries in the 2020 Olympics. Results show that the proposed approach can yield more precise estimates than previous methods. In addition, this paper presents insights into prediction techniques of large-scale international sports events and provides several future directions for researchers in the Olympic prediction field.

Keywords: Data envelopment analysis; Data-driven; Negative binomial regression; Olympic prediction

1 Introduction

Since the beginning of the first modern Olympic Games, this event has received increasing attention worldwide. Countries from around the globe now participate in the Olympic Games. The Summer Olympics, held every four years, is undoubtedly the world's leading sporting event in terms of the number of athletes and participating countries, the range of sports represented, and the size of the television audience (Forrest et al., 2017). Each Olympic Games publishes a ranking of countries by the number of medals they win. Still, these rankings are not officially recognized by the National Olympic Committee and are published only to indicate each country's results (Fazlollahi et al., 2020). Even so, each country still aims to win more medals. The number of medals reflects not only a participating country's national strength but also a participating country's political prestige and economic status. Research has shown that the self-reported well-being of a nation's citizens increases after an over-performance at the Olympics (Kavetsos and Szymanski, 2010; Hallmann et al., 2013). Therefore, winning medals at the Olympic Games is of great practical importance to enhance international influence and national pride. The Summer Olympics, which includes a variety of sports, is the focus of attention for countries worldwide as athletes from many countries compete to win medals in various sports. Planning the available resources to maximize the number of medals won at the Olympics has become a major concern for each government.

Olympic prediction has received much attention in recent studies. Predicting performance at the Olympics can provide governments with a better allocation of funding and an assessment of the individual programs' performances (Blais-Morisset et al., 2017). In addition, policymakers can adjust their policies and achieve improvements for the next Olympics based on ex-ante predictions. Both policymakers and industries are interested in Olympic predictions. For example, the media and journalists promote the Olympics before the Summer Games begin, and Olympic predictions are often the focus of these reports and media campaigns that are used to garner wide rating attention. In addition, some companies will adjust their sponsorship and support for target sports based on the Olympic predictions, and sports betting companies will adjust their betting odds based on the Olympic predictions. Therefore, Olympic predictions are important not only at the national level but also at the corporate level.

In the literature, most scholars have used political, economic, social, cultural, and environmental variables to predict the number of Olympic medals (Condon et al., 1999; Bernard and Busse, 2004; Vagenas and Vlachokyriakou, 2012; Otamendi and Doncel, 2014; Waguespack and Salomon, 2016; Fazlollahi et al., 2020; Mengjie et al., 2020). However, these studies cannot provide insight into the performance of individual Olympic sports and only explore the relationship between Olympic medal counts and the selected variables. In recent studies, data envelopment analysis (DEA) has been widely used to evaluate the Olympic Games (Wu et al., 2010). However, most studies only focus on the ex-post evaluation; few studies focus on the ex-ante Olympic Games prediction. In contrast, predicting future Olympic performance from an efficiency perspective can help us not only reasonably assess each Olympic sport but also reasonably optimize resources. Li et al. (2021) developed a DEA-based data-driven prediction method by which future team performance can be predicted. Notably, only one team's performance was predicted in Li et al. (2021), which cannot provide sufficient evidence to support that its outstanding predictive power is robust.

Our current study extends the approach developed by Li et al. (2021) to predict a country's performance at the Olympic Games from an individual sport perspective. The method makes predictions in two steps. The first step is to perform a multivariate statistical regression to estimate the production function between the number of medals and relevant Olympic outcomes. Since the number of medals is discrete, we adopt a negative binomial model to estimate the production function. In the second step, we simulate the decision process of policymakers who optimally allocate the investment to each sport to get the maximum number of medals. This step adopts DEA to construct production frontiers for both individual sports and countries. These two production functions aim to yield optimal Olympic outcomes. We also include the average inefficiency ratio to constrain the relative efficiency states across countries and sports.

Then we introduce the generated outcomes in the second step to the production function obtained in the first step to get the final predictions. We applied the method to the 2008-2016 Summer Olympics dataset and used it to predict the performance of China and other countries in the 2020 Summer Olympics. The results show that the proposed approach yields more accurate Olympics predictions than previous ones.

The main contributions of this paper are as follows. First, this paper extends the data-driven prediction method developed by Li et al. (2021) to Olympic prediction scenarios and confirms its effectiveness and applicability in predicting large-scale sports events. Second, this paper accurately predicts Olympics from an individual sport perspective, which is very different from previous studies. Most previous studies on the evaluation of the Olympic Games have focused only on the overall efficiency of a country (Lins et al., 2003; Li et al., 2008; Wu et al., 2009; Lei et al., 2015; Li et al., 2015). Finally, we confirm that DEA is a powerful approach in sports prediction scenarios, and the proposed approach yields more precise Olympic predictions than previous methods for multiple countries.

The remainder of this paper is organized as follows. Section 2 reviews relevant literature. We describe the problem setting and develop a data-driven prediction approach based on data envelopment analysis models in Section 3. Section 4 applies the methodology used in this paper to 2020 Summer Olympics predictions, and experiments are also discussed to analyze further based on the prediction results. Finally, Section 5 summarizes this paper.

2 Literature review

This paper uses a DEA-based data-driven prediction approach to predict China's Olympic performance in 2020 based on previous Olympics. Thus, the literature review in this paper focuses on both Olympic Games predictions and DEA.

2.1 Statistics-based Olympic Games Predictions

Medal prediction in the Olympic Games has long been a hot topic in the research field, and most studies have examined the factors influencing Olympic medal results. Bernard and Busse (2004) considered the production function of Olympic medals, exploring the relationship between medal shares and socioeconomic factors using econometric methods in terms of population, GDP per capita, host or not, national institutions, and medal shares of the previous edition. They used their model to predict the expected medal shares of each country at the Sydney 2000 Olympics. Forrest et al. (2017) extended Bernard and Busse's national-level model and applied it to individual sports, analyzing the impact of various factors on medals won in different sports. Pfau and Donald (2006) developed a model for predicting the Winter Olympics based on the model proposed by Bernard and Busse and applied the model to predict the 2006 Winter Olympics. Vagenas and Vlachokyriakou (2012) further explored the demo-economic factors influencing the Olympic medals and found that the Olympic medals were log-linearly regressed on the factors of land, population, GDP, urban population, inflation, growth rate, unemployment rate, labor force, health expenditures, ex-host, and team size. They found that Olympic team size was the best single predictor of Olympic medals. Similar to the previous study, Johnson and Ali (2004) investigated the economic and political factors that encourage countries to select Olympic athletes and predicted the number of medals won by the major participating countries in the 2002 Winter Olympics. Otamendi and Doncel (2014) developed a new econometric model for predicting the winter Olympics based on previous research and extended the model to apply it to individual sports, thus obtaining Olympic medal prediction results for the whole country. Otamendi and Doncel (2014) also applied their proposed model to the summer Olympics, thus confirming the feasibility of predicting Olympics results by sport. Forrest et al. (2010) considered quasi-host status (the country that will host the next Olympic Games) and public sports subsidies in predicting Olympic medal shares, and they found that both factors had a significant positive effect in predicting Olympic medal shares.

In addition, machine learning algorithms have received extensive attention in the study of Olympic Games forecasting. Condon et al. (1999) predicted the success of countries at the 1996 Olympics through neural networks. Mengjie et al. (2020) used a random forest regression model to predict each country's performance at the 2016 Olympics by regressing four factors: GDP, population, national team size, and host advantage. Fazlollahi et al. (2020) used a theoretical model and neural network model to predict the number of Olympic medals each country would win at the 2020 Tokyo Olympics. As a refinement of overall Olympic prediction, individual Olympic sport prediction has also been a research focus. For example, Huang Chang et al. (2012) used gray system theory to predict track and field results and their trends for the 2012 London Olympics. Van Tuyckom and Jöreskog (2010) used a structural equation modeling approach to explore the impact of different factors on the number of medals won at the Olympics. Yang et al. (2021) predicted the performance of the 2012 London Olympic Men's 100m Track and Field Championship based on an artificial neural network algorithm. Allen et al. (2015) used athletes' historical data and a neural network model to predict the athletes' future performance, thus providing a basis for talent selection. Holub et al. (2021) described the history of Olympic breaststroke and butterfly performance changes, developed a mathematical prediction model, and predicted swimming performance for the 2021 Tokyo Olympics.

Currently, most studies predict Olympic Games future performance through macroscopic predictors; few studies predict future performance from the efficiency perspective. Unlike the above-mentioned studies, this paper uses a data-driven approach based on DEA to forecast the Olympic performance of countries worldwide.

2.2 Olympic Games Prediction using Data Envelopment Analysis

DEA, first proposed by Charnes et al. (1978), is a nonparametric linear programming approach, and the method is used to measure the relative efficiency of multi-input and multi-output homogeneous decision making units (DMUs). DEA has been widely used in performance evaluation, such as analyses of bank performance evaluation (Ouenniche and Carrales, 2018), energy and environment efficiency (Geng et al., 2017), supply chain management (Chen and Yan, 2011), airport efficiency (Tsui et al., 2014), medical efficiency (Khushalani and Ozcan, 2017), and Olympic Games efficiency (Lins et al., 2003).

DEA has been widely studied and applied in the evaluation of Olympic Games. Lins et al. (2003) proposed a zero-sum benefit DEA model based on the classical DEA model for evaluating the relative efficiency among individual countries in the Olympic Games. Li et al. (2008) used the context-dependent AR-DEA (assurance regions DEA) developed by Cook and Zhu (2008) to measure and benchmark countries' performance at six Summer Olympics by incorporating multiple sets of AR into the DEA. In this study, individual participating countries were classified into four different revenue types, and these different groups of countries placed different emphases on gold, silver, and bronze medals. Wu et al. (2009) proposed an improved DEA game cross-efficiency model in which each decision unit is viewed as a competitor via a non-cooperative game, and they applied the model to evaluate the relative efficiency of each participating country in six Summer Olympics.

Among other things, the game cross-efficiency model implicitly includes the relative importance of gold, silver, and bronze medals without specifying the exact assurance regions. Yang et al. (2011) developed a DEA model with a fixed output sum in which the total output of all decision units is fixed, and the decision units compete to maximize their DEA efficiency scores. Unlike previous work that only evaluated the performance of participating countries in the Summer or Winter Olympics, Lei et al. (2015) proposed a parallel DEA approach to evaluate the Summer and Winter Olympics as a parallel system. They proposed an efficiency decomposition procedure to obtain a range of efficiencies for each Olympic subsystem, which was eventually applied to the 2012 Summer Olympics and the 2010 Winter Olympics. While the aforementioned studies treat each participating country in the Olympic Games as a black box, i.e., ignoring the internal processes, Li et al. (2015) used a two-stage DEA approach to open the black box and thus measure the performance of participating countries in the Olympic Games.

Almost all DEA models are based on prespecified input and output data for ex-post efficiency analysis, and few studies focus on future performance prediction. In addition, most studies are based on macroeconomic indicators to predict the performance at future Olympic Games, and few studies have predicted the future Olympic performance from an efficiency perspective. Therefore, this paper combines DEA and negative binomial regression approaches to predict Olympics outcomes. The method provides valuable new insights for performance prediction and Olympic Games predictions.

3 Problem and Methodology

In Section 3.1, we describe the problem setting of this paper, and in Section 3.2, we develop the mathematical model. To simplify the performance prediction problem, we narrow the topic by predicting the number of medals to be won in the next Olympic Games based on data from previous Olympic Games.

3.1 Problem setting

For the focal country (the one whose results we will predict), we expect that country to participate in *n* sports in the next Olympic Games. Because each Olympic sport requires a certain investment, we believe all sports will be allocated some funds. For a prespecified data sample, there are q Olympic Games recorded for the country, with the p-th (p=1,...,q) Olympic Game being characterized by invested funds $m^p \geq 0$ and outcomes x_r^p (r=1,...,s). Besides, each sport j (j=1,… ,n) obtains output $x_{rj}^p \geq 0$ within the total $m_j^p \geq 0$ in the p-th (p$\in P_j$) Olympic Game.

After the last Olympic Games, each country prepares for the next. Each country will allocate funds to the sports in which they will participate in the next Olympic Games and predict the outcomes. The total outcomes of all sports can then be translated into the number of Olympic medals won. The number of medals won at the next Olympic Games can be maximized with a proper approach. The problem, therefore, arises about how to allocate funds appropriately among the various sports and plan the outcomes of these sports to maximize the number of medals won at the next Olympic Games.

Since the Olympics are held once every four years, we assume that a country starts preparing for the next Olympics after the last one. Due to the GDP limit, the country invests a certain percentage of its GDP in the Olympics ($m_0 = \alpha\text{GDP}$). Thus, the total Olympic investment can be allocated among the individual sports. By summing the individual sports' outcomes, the entire country's total quantified outcomes can be predicted. In addition, based on the relationship between the total Olympic outcomes and the number of medals, it is possible to predict the number of medals that will be won by the country in the next Olympic Games.

Decision variables:

$m_j \geq 0$, the funds allocated to sport j, $m_0 \geq m_j \geq 0$,

$x_{rj} \geq 0$, the expected outcomes of sport j, $r = 1,...,s$,

$\hat{x}_r \geq 0$, the total expected outcomes for the focal country such that $\hat{x}_r = \sum_{j=1}^{n} x_{rj}$.

Parameters:

m_0, the total funds to be allocated to the sports by the focal country,

j, index of sports, $j = 1,...,n$,

r, index of outputs, $r = 1,...,s$,

P_j, the set of Olympic records for sport j ($|P_j| = q_j$),

p, the p-th Olympic record of a country in the sample, or the p-th record of a sport in the sample ($p \in P_j$),

$\mathrm{m}^p > 0$, the total funds a country can allocate to the p-th Olympic Games,

$\mathrm{x}_\mathrm{r}^\mathrm{p} \geq 0$, the p-th output achieved by a country in the p-th Olympic Games,

$\mathrm{x}_\mathrm{rj}^\mathrm{p} \geq 0$, the p-th output achieved by sport j in the p-th Olympic Games,

$m_j^\mathrm{p} \geq 0$, the funds allocated to sport j in the p-th Olympic Games.

In summary, we will maximize a certain objective function f by efficiently allocating funds. Note that the objective function f is determined according to the objective of the analysis, so it can have various formulas. This basic forecasting model can be formulated as model (1).

$$Max \quad f = f(m_1,...,m_n)$$

$$\#s.t.(1)\sum_{j=1}^{n} m_j = m_0 \ \#\#\#\#\#\#$$

$$m_j \geq 0, \forall j = 1,...,n$$

3.2 Methodology

3.2.1 Regression

We achieve the prediction in two main steps. First, the production function between the number of medals and various Olympic outputs is estimated by multivariate statistical regression analysis. Because the number of medals is a count variable, we use negative binomial regression to fit the relationship between the number of Olympic medals and various Olympic outcomes. The probability mass function (PMF) of the negative binomial can be represented by model (2).

$$P(Y=y)=\frac{\Gamma\left(y+\theta^{-1}\right)}{\Gamma(y+1)\Gamma\left(\theta^{-1}\right)}\left(\frac{\theta^{-1}}{\theta^{-1}+\lambda}\right)^{\theta^{-1}}\left(\frac{\lambda}{\theta^{-1}+\lambda}\right)^{y}\#(2)$$

where θ represents the dispersion parameter, and Γ represents the gamma function. The mean of the negative binomial distribution is λ, and the variance is $\lambda(1+\theta\lambda)$, so the negative binomial distribution allows the variance to exceed the mean (Zhang et al., 2012). The relationship between the number of Olympic medals won by a country and the various outcomes is shown in model (3).

$$Y=E_{\mathrm{xp}}(\beta_0+\sum_{\mathrm{r}=1}^{s}\beta_r x_r+\varepsilon)\#(3)$$

Many estimation methods can be used to estimate β and ε in the above equation, but Maximum Likelihood Estimation (MLE) is typically used to estimate these parameters. Assuming that the estimated parameters are (β_0^*, β_1^*,...,β_s^*), maximizing the number of medals can therefore be equated to maximizing the following linear equation.

$$\hat{\mathrm{f}}=\beta_0^*+\beta_1^*x_1+\cdots+\beta_s^*x_s\#(4)$$

3.2.2 Efficiency analysis for sports and the whole country

To achieve Olympic predictions, we need to specify a possible efficiency or inefficiency status during the predicted period. Therefore, we should first evaluate the country's efficiency and its efficiency in individual sports. Many methods are used for efficiency evaluation, most notably stochastic frontier analysis (SFA) and DEA. Here we use a method developed by Li et al. (2021) based on DEA, a method that has been successfully applied to NBA predictions. This method uses historical data to construct an efficiency frontier onto which all DMUs are projected. Consequently, the real units are compared with these projections to evaluate their relative efficiencies.

In this paper, we use the same model to evaluate the focal country's Olympic performance overall and in individual sports. Taking sport j as an example, P_{j} denotes the dataset of the sport in the past years, i.e., the reference set, and $\mathrm{q_j} = |P_{\mathrm{j}}|$ denotes the number of games in this dataset. Next, we can use the observations in the dataset to evaluate the relative efficiency of sport j in the *o*-th Olympic Games. We consider constructing frontier surfaces at the individual sport and country levels from historical data, subject to the total funding constraint.

$$\begin{gathered} Max\sum_{r=1}^{s} s_{rj}^{o} \\ s.t. \sum_{k\ \in\ P_j} \lambda_k m_j^k \le m_j^o \\ \sum_{k\ \in\ P_j} \lambda_k x_{rj}^k \ge x_{rj}^o + s_{rj}^o, \forall r = 1,\ldots,s \\ \sum_{k\ \in\ P_j} \lambda_k = 1 \\ \lambda_k, s_{rj}^o \ge 0, \forall k\ \in\ P_j; r = 1,\ldots,s \end{gathered} \#(5)$$

Supposing the optimal solution of model (5) is ($\lambda^{o}{}_{\mathrm{k}}^{*}, s^{o}{}_{rj}^{*}$), the inefficiency ratio of each sport can be calculated according to the following equation.

$$\rho^{o}{}_{rj}^{*} = \frac{s^{o}{}_{rj}^{*}}{x_{rj}^{o}} (r = 1,\ldots,s) \#(6)$$

We argue that the inefficiency states of countries and individual sports during the prediction period are unknown and fraught with uncertainty, but for the evaluated country, its past inefficiency states can be viewed as potential proxies for future inefficiencies. A natural idea is that a weighted average of past inefficiency states can represent future inefficiency states. Assume that the dataset of a country's game records in sport j is C_{j}. Then, based on the results in the equation (6), we can add up its inefficiency ratios and use the funds allocated to each sport as weights. With equation (7), we can obtain the weighted average inefficiency for each measurement.

$$\rho^{+}{}_{\mathrm{rj}}^{*} = \frac{\sum_{o \in C_j} m_j^o \rho^{o}{}_{rj}^{*}}{\sum_{o \in C_j} m_j^o} (r = 1,...,m) \#(7)$$

3.2.3 Prediction model

For DEA-based performance prediction, we consider the average inefficiency ratio to be a good proxy for the inefficiency state in the prediction period. A country's dominance in certain sports is expected to continue; it has performed better than other countries in the past and, therefore, is likely to perform better in the future also (Waguespack and Salomon, 2016). Therefore, we have the prediction model shown as model (8).

$$\begin{aligned}
Max \quad \hat{f} = \beta_0^* + \beta_1^* \hat{x}_1 + \cdots + \beta_s^* \hat{x}_s \\
s.t. \sum_{j=1}^{n} m_j = m_0 \\
\sum_{k \in P_j} \lambda_{kj} m_j^k \le m_j, \forall j = 1,...,n \\
\sum_{k \in P_j} \lambda_{kj} x_{rj}^k \ge x_{rj} + \rho_{rj}^* x_{rj}, \forall r = 1,...,s; j = 1,...,n \\
\sum_{k \in P_j} \lambda_{kj} = 1, \forall j = 1,...,n \\
\sum_{j=1}^{n} x_{rj} = \hat{x}_r, \forall r = 1,...,s \\
\sum_{l=1}^{q} \lambda_l x_r^l \ge \hat{x}_r + \rho_r^* \hat{x}_r, \forall r = 1,...,s \\
\sum_{l=1}^{q} \lambda_l = 1 \\
0 \le \mathrm{m}_j \le m_0, \forall j = 1,...,n \\
\lambda_{kj}, \lambda_l \ge 0, \forall j = 1,...,n; \forall k \in P_j; l = 1,...,q
\end{aligned} \#(8)$$

In model (8), m_j is the decision variable to allocate the optimal investment capital to each sport, and x_{rj} is the corresponding optimal output. The intensity variable λ_{kj} is used to construct an efficiency frontier for each sport j $(\mathrm{j} = 1,...,n)$. The first constraint requires that the funds allocated to all sports must add up to the country's total investment in the Olympics. The next three constraints are designed to ensure that each sport's planned inputs and outputs lie within the production possibility set during the prediction period, based on the variable returns to scale (VRS) assumption. Here the product of the planned input/output and the weighted average inefficiency ratio is the inefficiency state, or inefficiency slack, over the prediction period. The constraint $\sum_{j=1}^{n} x_{rj} = \hat{x}_r$ requires that the total output of the whole country in that Olympic Games is the sum of individual sport outputs in the prediction period. The remaining three constraints ensure that the expected output is within the production possibility set for the country's overall performance.

Assuming that the optimal solutions of the above model are m_j^* $(j = 1,\ldots,n)$ and $\hat{x}_r^*$ $(r = 1,\ldots,s)$, the optimal objective function is $\hat{f}^* = \beta_0^* + \beta_1^* \hat{x}_1^* + \cdots + \beta_s^* \hat{x}_s^*$. Thus, this function can obtain a prediction of the number of medals for the country during the prediction period.

4 Application to China's 2020 Olympic Performance

In this section, we apply the DEA-based data-driven approach proposed in Section 3 to predict China's Olympic performance in 2020. We obtain the prediction results and conduct several experiments to discuss and analyze the corresponding prediction results in more depth.

4.1 Data description

In this section, we use data from the 2008, 2012, and 2016 Summer Olympics to predict China's ideal performance in the 2020 Olympics. We use the average past performance of each country in each sport, i.e., the average inefficiency ratio, to represent future performance. Therefore, we excluded sports that have appeared only once in the Olympics, such as 3×3 basketball, skateboarding, and surfing, because they have no Olympic history and so we cannot get their average inefficiency ratios from historical data. In addition, we consider only the top thirty countries for the 2020 Olympics. The Olympic Games game data came from National Olympic Committees (https://olympics.com), and the GDP data was obtained from the World Bank website (https://data.worldbank.org/indicator/NY.GDP.MKTP.CD). We excluded observations with null values in the dataset. The descriptive statistics are shown in Tables 1 and 2.

In our predictive model, the input variable is a country's investment in sports. Since we do not have access to all countries' specific Olympic investments, we use a certain percentage of each country's GDP as that country's investment in their Olympic efforts, and here we use 0.004% of GDP. The output variables are the number of participants in a given sport and the rank score achieved (we use the reciprocal of the ranking in the sport as the score, e.g., 1 point for first place, 1/2 points for second place, and in general, 1/*j* points for *j*-th place). We believe that the more a country participates in the competition, the better its chances of winning an Olympic medal. A higher score indicates that the country has a greater advantage in the sport and a better chance of winning a medal.

Table 1 Statistics of the previous three Olympics (*N* = 90).

	Investment	Rank score	Number of players(#Players)
Max	2756000000.00	104.55	763
Min	1845932.00	4.89	44
Mean	262941500.00	27.62	297.1
SD	483142500.00	22.61	186.40

Table 2 Individual sport data in the previous three Olympics.

Sport name	Game count	Funding	Rank score	#Players
Archery	66	12579827.88	0.58	5.53
Artistic Gymnastics	55	14157497.49	1.94	35.24
Artistic Swimming	48	14884257.47	0.37	5.81
Athletics	88	9797709.02	3.81	45.19
Badminton	62	13208194.12	0.70	5.26
Basketball	33	16578252.69	0.46	18.33
Beach Volleyball	40	16604523.52	0.52	5.00
Boxing	77	10786204.46	1.33	5.61
Canoeing	78	10650466.79	1.68	12.42
Cycling	83	10345499.87	1.96	16.94
Diving	54	14644914.83	1.25	7.87
Equestrianism	60	13262015.52	1.00	13.73
Fencing	61	12920059.39	1.50	12.98

Football	37	16827472.22	0.36	22.84
Handball	34	7393564.516	0.39	21.26
Hockey	35	18963995.91	0.43	26.37
Judo	79	10708047.35	1.53	7.94
Modern Pentathlon	57	13653130.81	0.35	2.98
Rhythmic Gymnastics	42	13472525.06	0.41	4.71
Rowing	82	10368658.28	1.36	16.28
Sailing	72	11811291.31	1.26	10.07
Shooting	81	10388561.27	1.66	12.22
Swimming	90	9619662.438	3.96	37.08
Table Tennis	69	11692803.51	0.57	5.91
Taekwondo	66	12234495.89	0.71	2.67
Tennis	72	11438508.51	0.72	8.00
Trampoline Gymnastics	22	16624499.16	0.72	2.64
Triathlon	66	12621352.37	0.32	3.61
Volleyball	33	19222724.11	0.49	17.73
Water Polo	32	17194159.35	0.45	19.63
Weightlifting	72	11183357.81	0.94	4.53
Wrestling	74	10707716.02	1.73	7.68

4.2 Preliminary prediction results

First, we use the Olympic data from 2008, 2012, and 2016 to perform multiple statistical regressions; the results are shown in Table 3. The table shows that our multivariate regression statistical model can fit the number of medals very well. The results show that we can use the estimated regression equations to estimate the quantitative relationship between the various outcomes and the number of medals won. Thus, the regression equation used in the prediction model is shown below.

$$2.065178 + 0.0008013 \times \#\text{players} + 0.0243367 \times \text{rank score}$$

Note that the rank score is the most important item here, and the coefficients of both the #players and rank score are significantly positive, which is consistent with the expected hypothesis. This indicates that the more participants and the higher rank a country has, the more likely it is to win an Olympic medal.

Table 3 Regression results.

Variable	Coefficients	Std. Error	z-Statistic
β_0	2.065178***	0.0729995	28.29

β_1-#players	0.0008013**	0.0003496	2.29
β_2-rank score	0.0243367***	0.003314	7.34
	Pseudo R^2	0.2469	

Note: *, **, and *** denote significant levels of 0.1, 0.05, and 0.01, respectively.

The Summer Olympic Games are the largest event in the world and are held every four years. It seems likely that professional or world-class athletes would maintain good athletic performance during the time between Olympic years, so the earlier performance can be used to predict later performance. Research shows a significant relationship exists between the number of medals won in previous Olympic Games and the number of medals won in the current Olympic Games (Waguespack and Salomon, 2016). A natural assumption is that the most recent Olympic Games recorded data can reflect a country's Olympic efficiency. Therefore, we used data from the 2008, 2012, and 2016 Olympics to calculate China's inefficiencies at the national level and at the individual sport level by solving model (5). The inefficiency ratio is calculated by model (6), and the average inefficiency ratio can be obtained by model (7). Later, we will discuss the case of using different datasets in Section 4.3.

In Olympics in which it is not the host nation, China can only qualify for the Olympics in each sport through, for example, the World Championships before the Olympics competition or qualifying competitions for each sport. Therefore, China may not be able to participate in all sports in the 2020 Olympics. We calculated the average inefficiency ratio for each sport by model (7) after excluding the sports in which China did not participate in the 2020 Tokyo Olympics; the results are shown in Table 4. We optimized the allocation of Olympic funds and obtained the most promising performance of China in the 2020 Olympics by substituting the average inefficiency ratio into model (8). The results are shown in Table 5.

Table 4 China's average inefficiency ratio for each sport and whole country.

Sports	#Players	Rank score	Sports	#Players	Rank score
Archery	0.00	3.37	Modern Pentathlon	0.00	3.70
Artistic Gymnastics	0.33	1.83	Rhythmic Gymnastics	2.53	18.84

Artistic Swimming	0.00	1.33	Rowing	1.54	3.61
Athletics	1.27	2.15	Sailing	1.00	4.35
Badminton	0.13	0.27	Shooting	0.10	0.41
Basketball	0.00	5.63	Swimming	0.27	1.52
Beach Volleyball	1.97	12.03	Table Tennis	0.17	0.12
Boxing	0.09	1.09	Taekwondo	0.39	0.52
Canoeing	0.82	5.00	Tennis	1.24	7.05
Cycling	1.53	1.27	Trampoline Gymnastics	0.00	0.37
Diving	0.00	0.04	Triathlon	1.86	46.98
Equestrianism	19.39	73.01	Volleyball	0.86	2.68
Fencing	0.41	1.64	Water Polo	0.87	6.32
Football	1.06	9.49	Weightlifting	0.00	0.00
Handball	0.03	2.50	Wrestling	0.74	4.84
Hockey	0.85	9.50	-	-	-
Judo	0.65	3.05	Country	0.45	0.27

Table 5 Prediction results for China in the 2020 Olympics.

Sports	#Players	Rank score	Funding
Archery	12.00	1.07	1262905004
Artistic Gymnastics	46.75	4.25	13187500
Artistic Swimming	9.58	0.86	11314285.71
Athletics	64.42	8.66	91866666.67
Badminton	15.89	5.50	32179310.34
Basketball	24.00	0.30	73793103.45
Beach Volleyball	2.69	0.18	73793103.45
Boxing	10.11	2.53	879254.3871
Canoeing	14.82	1.14	21496296.3
Cycling	13.44	4.89	19618181.82
Diving	16.00	8.40	13187500
Fencing	21.91	2.41	11314285.71
Football	16.99	0.14	21496296.3
Hockey	17.81	0.14	7946666.67
Judo	8.50	1.60	27760000
Modern Pentathlon	4.00	0.28	16106666.67
Rhythmic Gymnastics	2.27	0.13	8600000
Rowing	18.48	1.47	13587096.77
Sailing	9.02	1.05	19618181.82
Shooting	25.37	5.82	13187500
Swimming	78.13	10.90	91866666.67
Table Tennis	10.23	5.08	13187500
Taekwondo	2.88	2.63	6580000
Tennis	8.94	0.53	77290322.58
Trampoline Gymnastics	4.00	1.70	13187500

Volleyball	12.88	0.41	73793103.45
Water Polo	13.91	0.17	1051111.11
Weightlifting	10.00	6.25	32179310.34
Wrestling	10.32	1.54	5027586.21
Country	505.35	80.04	2068000000
	Predicted Number of Medals		82.92

The results predict that China could win 82.92 medals at the 2020 Olympics by allocating the optimal amount of sports funding to each sport. There are some basic assumptions behind the predicted results. First, for this study, we assume that in previous Olympic Games, countries around the world used a fixed percentage of GDP as the Olympic investment and that these Olympic investments were equally distributed to each sport. Although it is impossible for a country to distribute the Olympic investment exactly equally, the prediction model forecasts future Olympic performance well, so we believe that our results still have significant value despite this first assumption. Second, we believe that the average inefficiency ratio in the past can represent a country's inefficiency during the forecast period. Therefore, in the prediction process, we eliminated those sports that appeared for the first time in the 2020 Tokyo Olympics because we could not get historical results for these sports, such as 3×3 basketball, skateboarding, and surfing.

We compared the prediction results obtained from model (8) with other predictions. Here we use two predictions that were more widely disseminated before the start of the Olympics by Gracenote and the Financial Times. Gracenote (www.gracenote.com) is a world-renowned data company. Before the Olympic Games started, the company published its predictions for the Games, the virtual medal table. The company's predictions were based on statistical models and data obtained since the 2016 Summer Olympics for various major sports, such as World Championships, World Cups, and qualifiers. Over time, the company continuously updates its virtual medal table for the Olympics based on the latest information available.

The Financial Times (https://ig.ft.com/tokyo-olympics-alternative-medal-table/) also released its predicted Olympic medal table before the start of the Olympic Games, showing a predicted number of medals for each country. Similar to previous studies, the medal table is primarily based on an economic forecasting model that uses population size, gross domestic product per capita, and past performance as predictors to predict future Olympic performance.

We do not use the final number of medals actually won by the Chinese national team at the Tokyo 2020 Olympics as our final comparative value. According to our assumptions, we have excluded those sports that were to appear for the first time in Tokyo 2020. Therefore, we subtracted the number of medals China won in those sports from the total number of medals won as the final total for the Chinese national team. At the Tokyo 2020 Olympics, the Chinese national team won 88 medals in all sports, and after subtracting the three medals won in sports appearing for the first time, the Chinese national team won 85 medals. Similarly, since the other companies were predicting the overall outcome of the Olympics, we preprocessed their predictions equally by subtracting three medals from their predictions. The final results are shown in Table 6.

Table 6 Comparison of predicted results.

Indicators	Our results	Gracenote	Financial Times
Actual medals	85	85	85
Predicted medals	82.92	63	71
Differences	2.08	22.00	14.00
Accuracy	97.56%	74.12%	83.53%

The difference between Gracenote's predicted results and the actual results is 22 medals, which reflects an accuracy rate of 74.12% (1-|85-63|/85). The inaccuracy could be explained by Gracenote's use of past results of major tournaments to make predictions since, due to the COVID-19 epidemic, China missed many tournaments in 2020. Thus, the prediction accuracy rate based on such tournaments was lower. Compared with Gracenote's prediction, Financial Times' prediction was slightly better. Financial Times predicted China would win 71 medals at the 2020 Tokyo Olympics, representing an accuracy rate of 83.53% (1-|85-71|/85). By optimizing the allocation of funds, this study predicts that China should have won 82.0047 medals in the 2020 Tokyo Olympics, with an accuracy rate of 97.56% (1-|85-82.92|/85). Comparing the three predictions, it is clear that this article's new method has higher prediction accuracy than the previous two methods. Therefore, it can be concluded that the prediction method based on data-driven DEA better predicts the performance of the Chinese team in the 2020 Tokyo Olympics.

4.3 Discussion

In the previous section, we used the Chinese performance in the 2020 Tokyo Olympics as an example to illustrate the effectiveness of our proposed data-driven approach in predicting future Olympic performance. In the following, we will further discuss the method, such as performing sensitivity tests on the inefficiency ratio, using different datasets for prediction, and applying the method to other countries.

In predicting Olympic Games performance, we assume that all outputs' inefficiency ratios are constant. However, the inefficiency ratios may change for each sport and country. For example, the country's performance in some sports worsens because star athletes retire and the younger athletes do not reach the higher level of the previous star athletes. In contrast, when a remarkably skilled athlete appears, the country may suddenly emerge as a stronger contender in that athlete's sport. Therefore, sensitivity tests, i.e., percentage changes in the inefficiency ratio, were conducted separately for each sport, for each country, and for each sport and country pair to check for impacts on the final prediction. First, we make a percentage change in the inefficiency ratio for all sports and calculate the optimal solution by substituting the revised inefficiency into the model (8). The variation of the optimal solution is shown as the blue line in Figure 1.

Similarly, we perform sensitivity tests at both the country-level inefficiency ratio and the inefficiency ratio of both—the final results appear as the red and green lines in Figure 1. When we changed the inefficiency ratio at the national and individual sport level, the final results were sensitive. When we change the inefficiency ratio for country and individual sport separately, the final number of medals obtained is more sensitive to the change in inefficiency ratio for individual sports than for countries. Throughout the Olympic Games, a single sport's efficiency can determine whether a sport wins a medal, and the medals won by a country are the cumulative result of the individual sports' results. Therefore, when a country prepares for the next Olympic Games, the optimal path involves allocating its limited funds rationally to individual sports to maximize each sport's efficiency and maximize the medals won in the next Olympic Games.

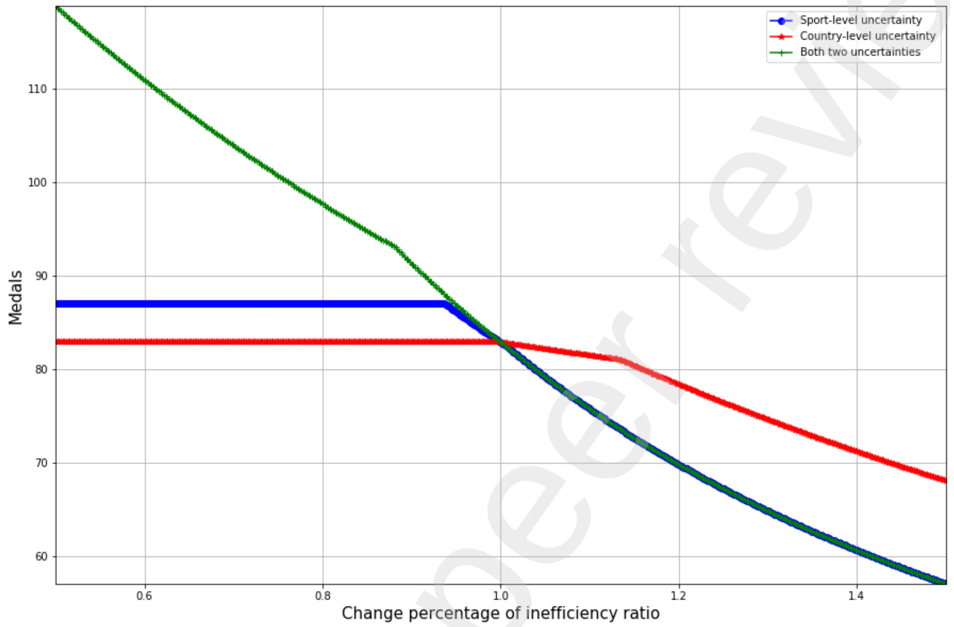

Figure 1 Sensitivity test results

The prediction results in Table 6 were obtained based on historical data from the 2008, 2012, and 2016 Olympics, and the results may differ when we use data from different periods. Therefore, a natural idea is to use historical data from just the previous Olympic Games (2016) and from just the previous two (2016 and 2012) to see how the resulting predictions differ from the prediction using data from all three Olympics Games. When we use data from just the 2016 Olympics, the predicted medal count for China is 77.51, giving an accuracy rate of 91.19% (1-|85-77.51|/85). When we use data from the 2012 and 2016 events, the predicted medal count for China is 86.84, representing an accuracy rate of 97.84% (1-|85-86.84|/85), which is slightly better than the three-event prediction's 97.56% (Table 6).

The prediction results we obtained using the three datasets are very close, and all three are better than the other two methods' results shown in Table 6. All results can predict China's performance in the 2020 Olympics very well, indicating that our method has good robustness. As can be seen, the lowest predicted medal count is obtained using data from the 2016 Olympics, which is consistent with our previous assumptions. We believe that the past efficiency state can represent the future efficiency state, and the Chinese team performed worse at the 2016 Olympics compared to 2012 and 2008, so using data from the 2016 Olympics is likely to underestimate the performance of the Chinese team at the 2020 Olympics.

Table 7 Prediction results with different datasets.

	2016	2016 and 2012	2016, 2012, and 2008
#Players	417.71	445.31	477.03
Rank score	65.67	69.99	57.08
Medals	77.51	86.84	82.92
Accuracy	91.19%	97.84%	97.56%

In the study above, we use historical data to successfully predict the best performance of the Chinese team at the 2020 Tokyo Olympics. We next extend the method to predict other countries. Consistent with the previous idea, we first calculate the average inefficiency ratio of each country using Models (5), (6), and (7), and then use Model (8) to obtain the predicted best performance of each country in the Tokyo 2020 Olympics. The results are shown in Table 8. Then we compare the results with Gracenote and Financial Times. Since we could not obtain Gracenote’s predicted values for Cuba, Croatia, Belarus, South Africa, the Czech Republic, and Azerbaijan, we exclude these countries in Table 9. Notably, the US (United States) is greatly overestimated in the forecast, which may be because the US has been number one in the medal table in past Olympic competitions. It has won far more medals than any other country. Our use of DEA is to forecast the best performance for that outcome, so we may overestimate the performance of the US when we predict the number of medals for the Olympics. Using our method proposed in Section 3, the two optimal output results for the US in 2020 are 644.90 and 104.55, while the actual results for the two outputs of the US at the Tokyo 2020 Olympics are 613 and 96.87, respectively. Therefore, we exclude the extreme values of the US. The final comparison results are shown in Table 9. The RMSE of the prediction using the method in this paper is 7.3658, which is better than the other prediction results. As can be seen from Figure 2, the prediction results follow the same trend as the real results. Our method can predict how other countries will perform at the 2020 Tokyo Olympics, which indicates that our approach is robust.

Table 8 Prediction results for other countries.

Country	Player number	Rank score	Medals
AUS	477.03	57.08	46.36
AZE	33.00	11.48	10.71
BLR	127.39	25.40	16.21
BRA	310.38	24.13	18.20
CAN	355.72	30.08	21.81
CHN	505.35	80.04	82.92
CRO	74.55	11.06	10.96
CUB	129.71	24.65	15.94
CZE	139.08	20.71	14.60
DEN	119.37	19.70	14.02
ESP	329.04	27.56	20.08
FRA	452.12	48.00	36.44
GBR	470.10	63.20	53.51
GER	505.61	56.78	47.09
HUN	195.03	39.02	23.83
ITA	379.63	38.27	27.13
JAM	67.91	11.90	11.13
JPN	408.49	39.19	28.40
KAZ	127.97	16.49	13.05
KEN	87.92	13.59	11.78
KOR	262.46	35.28	22.97
NED	237.80	27.50	18.63
NZL	178.90	29.38	18.61
POL	237.76	26.59	18.22
RSA	142.98	10.97	11.55
RUS	405.66	66.22	54.69
SWE	129.58	12.64	11.90
UKR	206.04	34.04	21.30
USA	644.90	104.55	168.41
UZB	57.62	10.78	10.74

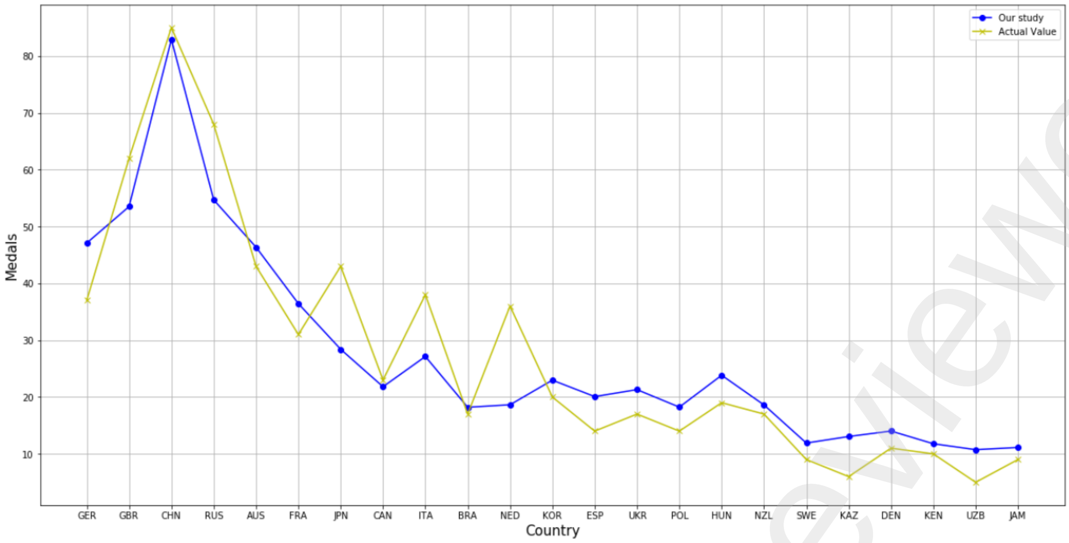

Figure 2 Prediction results for countries in the 2020 Olympics

Table 9 Comparison of predicted results for the 2020 Olympic Games.

Country	Our study	Gracenote	Financial Times
AUS	54.49	40	29
BRA	31.01	24	22
CAN	34.27	21	22
CHN	77.51	66	74
DEN	12.91	13	15
ESP	22.07	23	17
FRA	48.68	42	55
GBR	89.29	52	68
GER	90.99	35	44
HUN	17.29	28	16
ITA	34.27	41	29
JAM	10.89	11	10
JPN	39.45	60	57
KAZ	13.06	15	20
KEN	11.63	15	14
KOR	22.62	20	22
NED	24.00	48	21
NZL	15.91	15	18
POL	17.68	18	14
RUS	59.44	68	54
SWE	13.21	13	11
UKR	19.83	21	13
UZB	11.31	11	16
RMES	7.37	7.96	9.54

5 Conclusions, limitations and future directions

In this paper, by using a data-driven approach based on DEA, we have accurately predicted the Olympic performances by China and other countries in 2020. The method performs the prediction using two main steps. First, negative binomial regression is used to estimate the production function of the number of medals and various Olympic outcomes. Then we use DEA to allocate Olympic funds to obtain the optimal Olympic outcomes. In our model, we assume that a country's dominance in certain sports will continue, at least in the short term. That is, past performance can be used to predict future performance. The final prediction shows China winning 82.92 medals in the 2020 Tokyo Olympics. By comparing this result with other prediction results, we find that the data-driven prediction approach based on DEA can predict future Olympic performance very well.

Although this study yields more precise predictions than previous methods, we suggest four ways to extend it. First, the prediction might be more accurate if we could acquire the actual investment data for individual sports. Second, the total number of medals is fixed in the Olympic scenario, and we omitted this fixed-sum property in the methodology. Future research can consider the fact that the total number of medals is fixed. Third, this paper focused on Olympics, but the method could be applied to other sporting events, such as the soccer World Cup, various tournaments, and NFL Major League Football. Fourth, this paper only considered score values and Olympic delegation size as outputs. Additional input and output metrics could be considered in future studies to optimize the final prediction results.

References

Allen, S. V., T. J. Vandenbogaerde, D. B. Pyne and W. G. Hopkins (2015). "Predicting a Nation's Olympic-Qualifying Swimmers." International Journal of Sports Physiology and Performance **10**(4): 431-435.

Bernard, A. B. and M. R. Busse (2004). "Who wins the Olympic Games: Economic resources and medal totals." Review of Economics and Statistics **86**(1): 413-417.

Blais-Morisset, P., V. Boucher and B. Fortin (2017). "L'impact des dépenses publiques consacrées au sport sur les médailles olympiques." Revue économique **Vol. 68**(4): 623-642.

Charnes, A., W. W. Cooper and E. Rhodes (1978). "Measuring the efficiency of decision making units." European Journal of Operational Research **2**(6): 429-444.

Chen, C. and H. Yan (2011). "Network DEA model for supply chain performance evaluation." European Journal of Operational Research **213**(1): 147-155.

Condon, E. M., B. L. Golden and E. A. Wasil (1999). "Predicting the success of nations at the Summer Olympics using neural networks." Computers & Operations Research **26**(13): 1243-1265.

Cook, W. D. and J. Zhu (2008). "CAR-DEA: Context-dependent assurance regions in DEA." Operations Research **56**(1): 69-78.

Fazlollahi, P., A. Afarineshkhaki and R. Nikbakhsh (2020). "Predicting the Medals of the Countries Participating in the Tokyo 2020 Olympic Games Using the Test of Networks of Multilayer Perceptron (MLP)." Annals of Applied Sport Science **8**(4).

Forrest, D., I. G. McHale, I. Sanz and J. D. Tena (2017). "An analysis of country medal shares in individual sports at the Olympics." European Sport Management Quarterly **17**(2): 117-131.

Forrest, D., I. Sanz and J. D. Tena (2010). "Forecasting national team medal totals at the Summer Olympic Games." International Journal of Forecasting **26**(3): 576-588.

Geng, Z., J. Dong, Y. Han and Q. Zhu (2017). "Energy and environment efficiency analysis based on an improved environment DEA cross-model: Case study of complex chemical processes." Applied Energy **205**: 465-476.

Hallmann, K., C. Breuer and B. Kuehnreich (2013). "Happiness, pride and elite sporting success: What population segments gain most from national athletic achievements?" Sport Management Review **16**(2): 226-235.

Holub, M., A. Stanula, A. Baron, W. Glyk, H. Rosemann and B. Knechtle (2021). "Predicting Breaststroke and Butterfly Stroke Results in Swimming Based on Olympics History." International Journal of Environmental Research and Public Health **18**(12).

Huang Chang, M., H. Shen Wei and C. Xiao Xiao (2012). "Research on construction and application of the GM(1,1) forecast model of Olympics track and field achievements." Grey Systems: Theory and Application **2**(2): 178-196.

Johnson, D. K. N. and A. Ali (2004). "A tale of two seasons: Participation and medal counts at the Summer and Winter Olympic Games." Social Science Quarterly **85**(4): 974-993.

Kavetsos, G. and S. Szymanski (2010). "National well-being and international sports events." Journal of Economic Psychology **31**(2): 158-171.

Khushalani, J. and Y. A. Ozcan (2017). "Are hospitals producing quality care efficiently? An analysis using Dynamic Network Data Envelopment Analysis (DEA)." Socio-Economic Planning Sciences **60**: 15-23.

Lei, X., Y. Li, Q. Xie and L. Liang (2015). "Measuring Olympics achievements based on a parallel DEA approach." Annals of Operations Research **226**(1): 379-396.

Li, Y., X. Lei, Q. Dai and L. Liang (2015). "Performance evaluation of participating nations at the 2012 London Summer Olympics by a two-stage data envelopment analysis." European Journal of Operational Research **243**(3): 964-973.

Li, Y., L. Liang, Y. Chen and H. Morita (2008). "Models for measuring and benchmarking olympics achievements." Omega-International Journal of Management Science **36**(6): 933-940.

Li, Y., L. Wang and F. Li (2021). "A data-driven prediction approach for sports team performance and its application to National Basketball Association." Omega-International Journal of Management Science **98**.

Lins, M. P. E., E. G. Gomes, J. de Mello and A. de Mello (2003). "Olympic ranking based on a zero sum gains DEA model." European Journal of Operational Research **148**(2): 312-322.

Mengjie, J., Z. Yue, C. Furong, Z. Bofeng and K. Yoshigoe (2020). A Random Forest Regression Model Predicting the Winners of Summer Olympic Events.

Otamendi, J. and L. M. Doncel (2014). "Medal Shares in Winter Olympic Games by Sport: Socioeconomic Analysis After Vancouver 2010." Social Science Quarterly **95**(2): 598-614.

Ouenniche, J. and S. Carrales (2018). "Assessing efficiency profiles of UK commercial banks: a DEA analysis with regression-based feedback." Annals of Operations Research **266**(1-2): 551-587.

Pfau and W. Donald (2006). "PREDICTING THE MEDAL WINS AT THE 2006 WINTER OLYMPICS: AN ECONOMETRICS APPROACH." The Korean Economic Review **22**(2): 233-247.

Tsui, W. H. K., H. O. Balli, A. Gilbey and H. Gow (2014). "Operational efficiency of Asia-Pacific airports." Journal of Air Transport Management **40**: 16-24.

Vagenas, G. and E. Vlachokyriakou (2012). "Olympic medals and demo-economic factors: Novel predictors, the ex-host effect, the exact role of team size, and the "population-GDP" model revisited." Sport Management Review **15**(2): 211-217.

Van Tuyckom, C. and K. G. Jöreskog (2010). "Going for gold! Welfare characteristics and Olympic success: an application of the structural equation approach." Quality & Quantity **46**(1): 189-205.

Waguespack, D. M. and R. Salomon (2016). "Quality, Subjectivity, and Sustained Superior Performance at the Olympic Games." Management Science **62**(1): 286-300.

Wu, J., L. Liang and Y. Chen (2009). "DEA game cross-efficiency approach to Olympic rankings." Omega-International Journal of Management Science **37**(4): 909-918.

Wu, J., Z. X. Zhou and L. A. Liang (2010). "Measuring the Performance of Nations at Beijing Summer Olympics Using Integer-Valued DEA Model." Journal of Sports Economics **11**(5): 549-566.

Yang, F., D. D. Wu, L. Liang and L. O'Neill (2011). "Competition strategy and efficiency evaluation for decision making units with fixed-sum outputs." European Journal of Operational Research **212**(3): 560-569.

Yang, S., L. Luo, B. Tan and H.-Y. Kung (2021). "Research on Sports Performance Prediction Based on BP Neural Network." Mobile Information Systems **2021**: 1-8.

Zhang, X., Y. Lei, D. Cai and F. Liu (2012). "Predicting tree recruitment with negative binomial mixture models." Forest Ecology and Management **270**: 209-215.