Olympic Competitiveness Assessment and Medal Prediction Based on TOPSIS Entropy Weight-Random Forest Model

Shanfei Zhang[#], Qian Yang[*,#]

School of Business, Beijing Language and Culture University, Beijing, China, 100083
[#]These Authors Contributed Equally.
**Corresponding author: blcu2022yq@163.com*

Abstract: *With the continuous refinement of the Olympic system, existing medal prediction methods based on single indicators have become inadequate for assessing multidimensional competitiveness. Specifically, this study aims to establish a comprehensive evaluation system for evaluating medal-winning potential at the national level. Using the entropy method and PCA-TOPSIS approach, four core dimensions—ability, changes, competitiveness, and strategy—are extracted from twelve initial indicators to form a multidimensional evaluation index system. On this foundation, a Random Forest Regression model is employed to conduct a predictive analysis of Olympic medal distribution. The research findings indicate that the United States is expected to perform exceptionally well in both gold medals and total medals, maintaining its leading position. The model evaluation results, with a value of 0.9437 and an MSE of 0.0117, demonstrate a high degree of goodness-of-fit. Furthermore, in simulating predictions for the medal tally of the 2024 Olympic Games, the model's predictions deviate by less than 5% from actual data, further validating the model's effectiveness. The study reveals that this comprehensive evaluation system effectively overcomes the limitations of single indicators, providing a more scientific and comprehensive analytical approach to Olympic medal prediction. This research not only broadens the theoretical dimensions of sports competitiveness assessment but also offers a quantitative reference for countries to formulate Olympic strategies, thereby possessing significant theoretical and practical value.*

Keywords: *TOPSIS Entropy Weight, Random Forest Model, Medal Potential Evaluation System, Competitiveness Assessment*

1. Introduction

As the most authoritative and influential comprehensive sporting event globally, the Olympic Games serve as a vital platform for demonstrating the sporting prowess and overall national strength of various countries. Establishing scientific and accurate medal prediction models to uncover the patterns of medal winning and provide inspiration and assistance to more countries holds profound practical significance. Currently, some scholars primarily rely on time series prediction models and simple empirical models for medal predictions. These methods primarily focus on the application of single-factor or linear models, such as the ARIMA model[1] and linear regression[2]. Although these methods have, to some extent, revealed the key factors influencing medal attainment, they often overlook the complex interrelationships among the multidimensional factors that affect Olympic competitiveness. With the continuous improvement of the Olympic system, the gradual standardization of competition rules, and the increasingly fierce competition, traditional statistical models are no longer sufficient for accurately predicting medal distributions. Consequently, scholars have begun to explore the use of intelligent prediction models, including machine learning[3], grey models[4], neural networks[5], and others. These models can comprehensively consider multidimensional indicators, more accurately reflecting the current status of national sporting strength and predicting future trends. For instance, Cao Jingbo (2021)[6] employed a multiple stepwise regression model, a grey model, and a combined model of the two to predict the medals of a specific Olympic Games, further validating the effectiveness of multidimensional factors in medal prediction. Additionally, to address the remaining shortcomings of intelligent prediction models, Tian Hui et al. (2021)[7] used a logistic regression model to predict the number of medals for Chinese athletes at a specific Olympic Games. However, this model's assumptions are overly stringent and primarily focus on predicting the medal count of the host country, with limited predictive ability for other countries' medal situations.

Through the review of past literature, this study argues that a medal prediction model should not only accurately reflect the sporting strength of various countries but also possess the capability to predict future trends, providing support for countries to formulate scientific and reasonable training plans and competition strategies, thereby promoting the prosperous development of global sports. Therefore, to comprehensively consider the factors influencing medals, this study, building on the research of Jiang Lei et al. (2022)[8], adopts the entropy method and PCA-TOPSIS method to construct an evaluation framework for assessing countries' medal-winning potential. This indicator system encompasses four core dimensions and 12 specific indicators. In terms of research methods, this study refers to the literature by Schlembach et al. (2022)[9], utilizing the Random Forest Regression model to predict the medal distribution at the 2028 Olympic Games and employing Mean Squared Error (MSE) and R² values to assess the model's performance, ensuring the reliability of the prediction results..

2. Data Sources and Preprocessing

2.1 Data Sources

The data for this study is sourced from the website of The Consortium for Mathematics and its Applications (www.comap.com). Due to the limited timeliness and comparability of Olympic Games data from excessively distant past, this study utilizes data from a total of eight Olympic Games editions spanning from 1996 to 2024.

2.2 Select and Process Indicators

Through in-depth mining and analysis of historical data, this study finds that the factors affecting the number of Olympic medals in various countries are complex and diverse. To more accurately reveal the inherent patterns of these variables and quantitatively characterize the competitive advantages of countries or athletes in the Olympic Games, this study has developed a probability assessment system consisting of four aspects: Athlete award ability, National sports competitiveness, National participation strategy, and Changes in competition system. The specific division is shown in Figure 1.

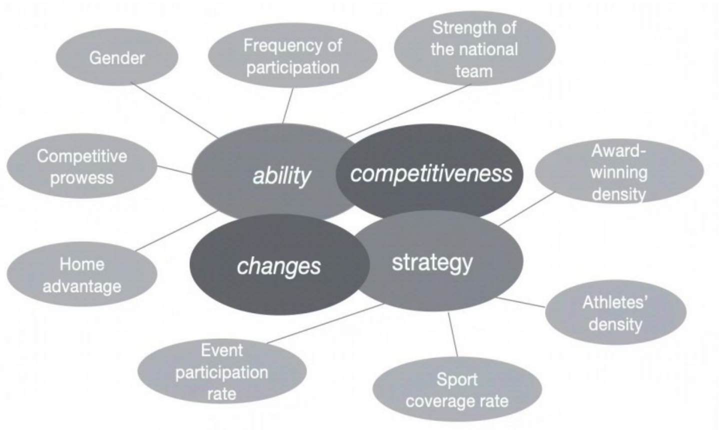


Figure 1 Probability Assessment System

2.2.1 Athlete Award Ability (ability)

To accurately assess athlete performance and predict their prospects for winning awards, this study proposes an evaluation model based on the TOPSIS Entropy Weight Method. This method combines TOPSIS with entropy to objectively determine the weights of evaluation indicators, effectively reducing the influence of subjective factors on the results. The model structure comprises steps such as data collection and preprocessing, entropy calculation for weight determination, and performance evaluation using the TOPSIS technique based on weighted indicators.

What's more, each athlete has significant variations in their style, personal level, and other individual factors. Additionally, each game is influenced by multiple complex factors and is subject to frequent changes. Therefore, this study considers both internal and external factors that influence athlete awards. Internal influencing factors include gender, participation frequency, participation experience, and competition competitiveness, while external influencing factors include the strength of the national representative team and whether they are athletes from the host country.

Instead of analyzing the performance of all athletes in all games using the same model, this study chooses to measure the individual award-winning abilities of athletes first, and then determine the average level of athletes in a country. Rather than directly analyzing all athletes in a country using a model, this method provides higher accuracy and relevance. The indicators and their properties are shown in Table 1.

(1) Gender: Indicating athletes' physiological and psychological competitive advantages in specific events, with a significant influence on medal distribution;

(2) Frequency of participation: Gauging athletes' versatility in competition, where a higher frequency of involvement translates to more opportunities to demonstrate their skills and secure awards;

(3) Participating experience: Reflecting athletes' familiarity with the Olympic stage, where extensive experience enhances the likelihood of medal success;

(4) Strength of the national team: Quantifying the overall competitiveness of a country's sports, as strong national teams typically exhibit higher medal-winning efficiency;

(5) Competitive prowess: Assessed based on athletes' award points, with higher points indicating greater strength and medal potential;

(6) Home advantage: Granting athletes from the host country psychological, environmental, and other favorable conditions, which may positively impact their medal performance.

Table 1 Athlete Award Ability

First-class indicators	Second-class indicators	Third-class indicators	Indicator property
ability	Internal influencing factors	Gender	+
		Frequency of participation	+
		Participating experience	+
		Strength of the national team	+
	External influencing factors	Competitive prowess	+
		Home advantage	+

Based on the entropy method, this study calculates the award-winning capability index for each athlete from different countries, and then divides it by the total number of participating athletes from that country to obtain the per capita award-winning capability index.

2.2.2 National Sports Competitiveness (competitiveness)

This study measures national sports competitiveness through historical medal counts. The total number of medals, especially gold medals, serves as an important and intuitive indicator for evaluating a country's overall strength in international sports events such as the Olympics. The more medals a country wins, the more it reflects its competitiveness and advantage on the international sports stage.

2.2.3 National Participation Strategy (strategy)

To have a specific indicator to access the National participation strategy, this study employs Principal Component Analysis (PCA) to analyze the dataset and reduce dimensionality. Subsequently, this study determines the weights for each indicator and ranks the proximity of the evaluated objects' numerical distance to the ideal target. This method combines the advantages of both sub-models and enables more accurate estimating of influences of nations' participation strategy of awarding medals.

Table 2 National participation strategy

First-class indicators	Second-class indicators	Third-class indicators	Weight
strategy	Participation breadth	Event participation rate	0.372
		Sport coverage rate	0.218
	Participation depth	Athletes' density	0.262
		Award-winning density	0.146

This study measures a country's participation strategy from two dimensions: participation breadth and participation depth, as shown in Table 2. Each of them contains two third-class indicators respectively. Participation breadth includes: the event participation rate, which is measured by the proportion of events a country has participated in out of the total number of events in previous Olympic Games; the sport coverage rate, which is measured by the proportion of sports a country has participated in out of the total number of sports. Participation depth includes: the athletes' density, which is measured by the average proportion of athletes a country has sent to each event out of the total number of athletes participating in that event; the award-winning density, which refers to the success rate or efficiency of winning medals among the participating athletes.

2.2.4 Changes in Competition Format (changes)

This study considers the potential impact of the addition or removal of events in each Olympic Games on the distribution of medals. Apart from some events being canceled due to policy adjustments by the International Olympic Committee (IOC), the host country often adjusts the Olympic program according to its own sports culture, to showcase its iconic sports to the world. This may result in the addition of advantageous events or the elimination of disadvantageous ones, thereby leading to changes in the distribution of medals. This study quantifies the impact of changes in the events from one Olympic Games to the next on the medal distribution through analysis and uses this as an important feature variable for predicting the number of Olympic medals.

2.3 Variable Correlation Analysis

This study uses variable correlation analysis to explore the degree of association between various variables, aiming to uncover potential relationships among them and reveal any possible synergistic or antagonistic effects that may exist between the variables. The relationship between the variables is shown in Figure 2.

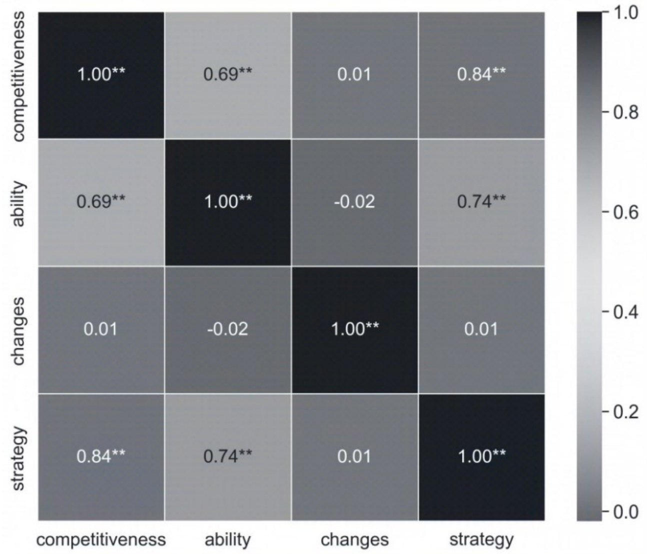


Figure 2 Feature correlation Matrix with Significance

3. Construction of Olympic Medal Influence Factors and Assessment System

This study selects the Random Forest model to predict the 2028 Olympics. The reason for choosing the Random Forest model is its ability to handle a large number of feature variables, effectively avoid overfitting, and automatically assess the importance of feature variables. This study establishes the model as follows:

$$f(x,\{\theta_n\}) = \frac{1}{N}\sum_{n=1}^{N} f_n(x,\{\theta_n\}) \tag{1}$$

Where $\{\theta_n\}$ is the nth regression tree; x is the influencing factor; $f(x,\{\theta_n\})$ is the overall passenger flow prediction result; $f_n(x, \{\theta_n\})$ is the prediction result of the nth tree; N is the number of trees.

The three basic steps to construct a Random Forest model are as follows, as shown in Figure 3:

(1) Randomly select N samples with replacements from the original training dataset to form a training subset, resulting in N subsets $\{\theta_1, \theta_2, \dots, \theta_n\}$. Construct a decision tree using each subset, yielding N decision trees $\{(x, \{\theta_1\}), (x, \{\theta_2\}), \dots, (x, \{\theta_n\})\}$.

(2) During the construction of each decision tree, randomly select m (where $m \ll M$) variables from the M feature variables as candidate branch variables, and choose the best one from the m features as the

splitting node for each tree.

(3) After obtaining the required number of decision trees, the prediction result of each decision tree model is denoted as $f_n(x, \{\theta_n\})$, and the final prediction result is the average of these decision tree prediction results.

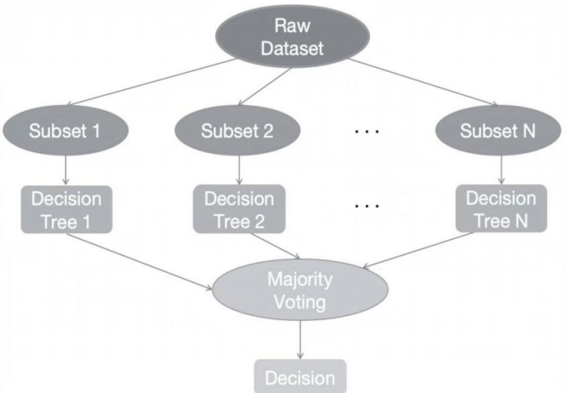


Figure 3 Schematic diagram of random forest model

Subsequently, this study used grid search to optimize the number of decision trees and maximum depth in the random forest algorithm. The random seed number was set to a fixed value to ensure the reproducibility of the results, and the optimal parameter values were selected to improve the predictive performance of the model. Next, this study inputs the feature variables and divides the dataset into a training set and a testing set in a ratio of 7:3 for training. Finally, the trained model is used to predict the Olympic medal situation in 2028.

4. Application of Random Forest Model in Olympic Medal Prediction

4.1 Model implementation details

In this study, entropy method and PCA-TOPSIS method are used to establish a random forest regression model based on the four core dimensions of ability, changes, competitiveness and strategy, and the medal prediction model is implemented in Python.

4.2 Analysis of experimental results

Figure 4 presents the confidence intervals (CI) for the prediction results of various countries. By calculating the confidence intervals, this study can assess the range of uncertainty in the model's predictions, further validating the robustness of the model. The results show that the confidence intervals for the predicted values of most countries are narrow, indicating the high reliability of the model's predictions. However, the confidence intervals for a few countries are relatively wide, possibly reflecting the variability inherent in the data or limitations in sample size.

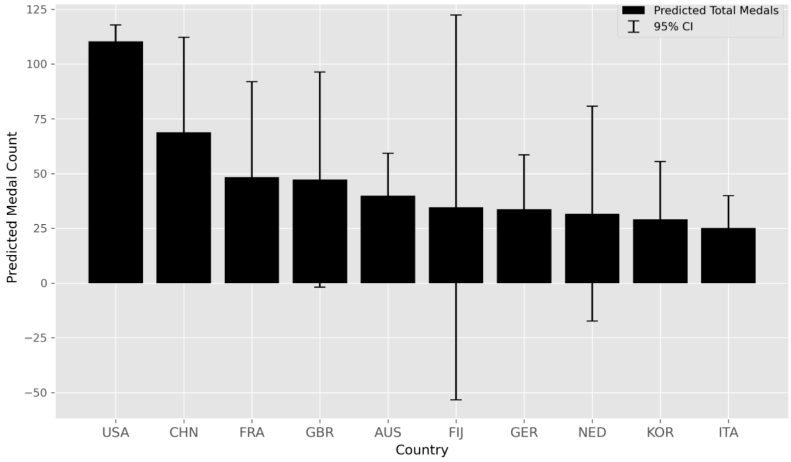


Figure 4 Predicted results with 95% confidence intervals

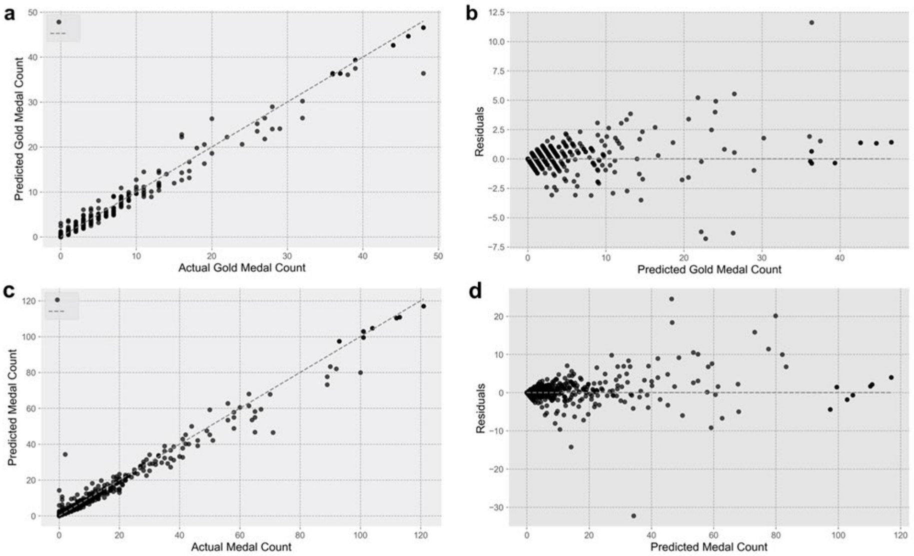
To evaluate the accuracy and reliability of the model, this study adopted Mean Squared Error (MSE) as the primary evaluation metric, with the results presented in Table 3. The analysis indicates that the model exhibits a low MSE value, suggesting a small average level of prediction error and high prediction accuracy. The model's R^2 value is 0.9437, indicating a good fit of the model to the data. Additionally, this study simulated the 2024 Olympic medal counts, and the results showed an error margin of less than 5% compared to the actual outcomes. This close alignment further validates the model's reliability and provides confidence in its predictions for the 2028 Olympics.

Based on the Random Forest model, this study has predicted the number of gold medals and total medals for each country at the 2024 Olympics. The results are shown in Table 3. In addition, by comparing the predicted values with the gold and total medals won in 2024, this study has assessed the progress and regression of different countries, as shown in Figure 4. This study has rounded the data for simplicity, where blue shading in the figure indicates a potential regression in 2028, while yellow shading suggests potential progress. Based on the model's results, this study believes that the US, France, and Great Britain are likely to make progress.

Table 3 Medal Prediction Results

2028 Rank	Country	2024		2028		Differences		MSE	R^2
		Gold	Total	Gold	Total	Gold	Total	Total	Total
1	United States	40	126	41	122	1	-4	0.0117	0.9437
2	China	40	91	39	90	-1	-1		
7	Great Britain	14	65	18	65	4	0		
5	France	16	64	21	64	5	0		
4	Australia	18	53	18	50	0	-3		
3	Japan	20	45	20	44	0	-1		
9	Italy	12	40	10	41	-2	1		
6	Netherlands	15	34	13	33	-2	-1		
10	Germany	12	33	10	33	-2	0		
8	South Korea	13	32	13	32	0	0		

Among them, the US is expected to perform exceptionally well in both gold medals and total medals, continuing to rank first, with China closely following in second place overall. Due to a large number of countries, this study has only listed the top ten countries in the overall ranking and the residual plot for the medal training set results of the United States, which ranks first, as shown in Figure 5.



(a) Gold Medal Prediction; (b) Residuals of Gold Medal Prediction; (c) Total Medal Prediction; (d) Residuals of Total Medal Prediction.

Figure 5 Medal Prediction of the US.

4.3 Sensitivity Test

To evaluate the robustness and reliability of the model, this study conducted a sensitivity test by introducing random noise to the input data. Specifically, this study added ±5% and ±10% random noise to the input data and observed the resulting changes in the model's R^2 score and predicted medal counts, with the results presented in Figure 6. The Random Forest model achieved an R^2 score of 0.9437, indicating a high level of predictive accuracy. When ±5% noise was introduced, the R^2 score decreased slightly to 0.9370, representing a variation of only 0.67%. Similarly, with ±10% noise, the R^2 score was 0.9398, showing a minimal variation of 0.40%. These results demonstrate that the model's performance remains stable even under significant data perturbations, highlighting the robustness of the feature engineering process and the model's ability to handle noise in the input data.

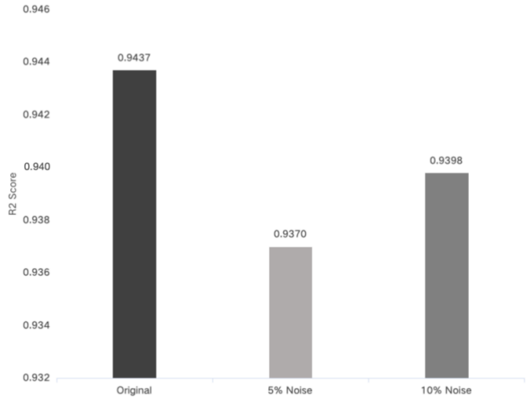


Figure 6 Model Stability under Random Noise

Furthermore, the predicted medal counts exhibited less than a 2% variation across all noise levels, reinforcing the model's stability. This low sensitivity to noise suggests that the model's predictions are reliable and not overly dependent on minor fluctuations in the input data.

The results provide strong evidence that the model is well-suited for real-world applications, as it can maintain high accuracy and reliability even when faced with imperfect or noisy data.

5. Conclusions

This study establishes a multidimensional evaluation system for Olympic medal prediction by integrating the TOPSIS entropy weight method and Random Forest regression. Through dimensionality reduction of twelve initial indicators into four core dimensions—athlete award ability, national sports competitiveness, participation strategy, and competition format changes—the model achieves high predictive accuracy (R^2=0.9437, MSE=0.0117) and demonstrates strong robustness under sensitivity tests. The simulation results for the 2024 Olympics show deviations of less than 5% from actual outcomes, with the United States and China predicted to maintain the dominance in gold and total medals. These findings validate the effectiveness of combining multidimensional indicators and machine learning in addressing the complexity of the Olympic competitiveness assessment. However, this study has limitations. The model relies on historical data, which may not fully capture future trends influenced by geopolitical shifts or technological advancements in sports. Future research could incorporate real-time dynamic data and refine clustering methods to better represent the nuanced interactions in Olympic competitions.

This study provides a research framework and analytical approach applicable to sports statistics and management. By integrating entropy-based weighting for objective indicator evaluation and Random Forest for nonlinear pattern recognition, the methodology effectively addresses the multidimensional nature of competitiveness assessment. The high predictive accuracy and stability under noise perturbations demonstrate the feasibility of this model in solving complex problems such as strategic resource allocation, performance optimization, and policy formulation in sports. Specifically, the framework can help national sports agencies quantitatively evaluate strengths and weaknesses, optimize training programs, and design targeted strategies to enhance Olympic competitiveness, thereby advancing the scientific management of global sports development.

References

[1] Liu Longxiang. Research on the Prediction of Olympic Gold Medals in 2020 Based on Data Mining Model[J]. Information Recording Materials, 2018, 19(05): 203-205.

[2] Zhang Yuhua. Prediction of the medal count for the 31st Olympic Games in China based on a linear regression dynamic model[J]. Journal of Henan Normal University (Natural Science Edition), 2013, 41(02): 24-26+60.

[3] Shi Huimin, Zhang Dongying, Zhang Yonghui. Can Olympic medals be predicted? From the perspective of interpretable machine learning[J]. Journal of Shanghai University of Sport, 2024, 48(04): 26-36.

[4] Luo Yubo, Cheng Yanfang, Li Mengyao, Xie Xinru, Huang Shuoshuo. Prediction of the number of Chinese medals and overall strength in the Beijing Winter Olympics based on the host effect and the grey prediction model [J]. Contemporary Sports Technology, 2022, 12(21): 183-186.

[5] Wang Fang Prediction of medal results of the 2020 Olympic Games based on neural networks [J]. Statistics and Decision, 2019, 35(05): 89-91.

[6] Cao Jingbo Prediction and Preparation Strategy for China's Medals in the 2022 Beijing Winter Olympics Based on the Host Effect [C]. Abstracts of the 2021 Academic Conference on Cultural Resources Assisting Cultural Communication and Cultural Heritage Development of the Beijing Winter Olympics Sports Culture Development Center of the General Administration of Sport of China, 2021: 73-74.

[7] Tian Hui, He Yiman, Wang Min, et al. Prediction of Chinese athletes' medals and participation strategies for the 2022 Beijing Winter Olympics - based on the analysis of the home advantage effect of the Olympic Games [J]. Sports Science, 2021, 41(02): 3-13+22.

[8] Jiang Lei, Zhang Youyin, Zhang Jingquan, Liu Mengqiao, Xu Heng. Evaluation and difference analysis of provincial leisure sports competitiveness based on entropy weight TOPSIS method [J]. Journal of Shaanxi Normal University(Natural Science Edition), 2022, 50(06): 113-123.

[9] Christoph Schlembach, Sascha L. Schmidt, Dominik Schreyer, Linus Wunderlich. Forecasting the Olympic medal distribution: A socioeconomic machine learning model[J]. Technological Forecasting and Social Change, 2022, 175:121314.