Predicting Olympic Medal Distributions and Emerging Nation Breakthroughs: A Hybrid Negative Binomial and Logistic Regression Framework

Siyuan Wang [1, #, *], Zhaoyang Huang [2, #], Shuyi Zhang [1, #], Wenxi Wang [1, #], Yu Liu [3], Junyang Yu [1]

[1] Department of Electronic engineering & Optoelectronic tech, Nanjing University of Science and Technology ZiJin College, Nanjing, China, 210023

[2] School of Business, Nanjing University of Science and Technology ZiJin College, Nanjing, China, 210023

[3] School of Intelligent Manufacturing, Nanjing University of Science and Technology ZiJin College, Nanjing, China, 210023

* Corresponding Author Email: 19852225218@163.com

Abstract. Olympic medal distribution is affected by multiple factors, and accurate prediction of medal distribution in future Olympic Games is of great significance to optimize the sports resource architecture of each country. By constructing an Olympic medal count prediction model based on negative binomial regression and logistic regression, this paper predicts that the United States and China will dominate with 46 gold medals and 142 total medals, and 44 gold medals and 98 total medals, respectively; countries such as France and Germany present a balanced multi-sport profile. Of the five small countries selected for discussion, Mali, Bolivia, and Myanmar show high probabilities. At the same time, Yemen and Angola are significantly lower, with the rise of sports underdogs highlighting the diversity of the sports competition landscape. The study reveals the dynamic changes in global sports' competitive strength and the driving factors behind it, providing a scientific basis for the National Olympic Committees to optimize the allocation of sports resources.

Keywords: Negative Binomial Regression, Logistic Regression, Prediction Model.

1. Introduction

The Olympic Games is a complex system, making it important to develop prediction models for medal distribution [1-4]. For the Paris 2024 Games, the United States is top of the table with 126 medals, with China tied for first place; France is fifth in gold medals and fourth in medals; Britain is third in medals; and small countries such as Albania and Dominica have won medals for the first time, highlighting the diversity of the medal distribution. However, according to statistics, there are still about 65 countries failing to win medals [5-6]. The prediction of Olympic medals can improve the level of national sports competition and provide a reference for promoting more countries to emerge in the international sports arena [7].

At the beginning of the 21st century, Bernard and Busse analyzed the impact of socio-economic resources on the number of Olympic medals using the Cobb-Douglas production function, pioneering a medal prediction method based on the quantification of economic resources [8]. Li Zhong et al. successfully predicted the number of silver medals for France in the next Summer Olympics using the ARIMA time series model in the 2110s [9]. Nicolas Scelles et al. predicted the total number of national medals for the 2016 and 2020 Summer Olympics using the Tobit model and the hurdle (Hurdle) model [10]. Cheng Hongren et al. analyzed the performance of China's track and field events in the Tokyo Olympics by logical induction [11]. Luo Yubo et al. predicted the number of Chinese medals by combining the host effect with the simple moving average method and grey prediction model [12].

Despite valuable insights from previous studies, gaps remain in the applicability and accuracy of theoretical tools for resource-challenged countries and complex scenarios. This paper introduces a multilevel negative binomial regression model that surpasses traditional Poisson regression by effectively capturing dynamic factors like country random effects. Our model achieves an RMSE of 3.2 for gold medals and 8.7 for total medals, with an R^2 greater than 0.85. We also enhance the analysis of Olympic performance using the PEST framework and present a logistic regression model to evaluate non-winning countries, addressing the competitiveness of smaller nations. Our model's robustness is validated through Monte Carlo simulation and random forest analysis, improving prediction accuracy and leading to fairer assessments of national sports capabilities.

2. Olympic Medal Prediction Model

2.1. Prediction model of the medal table of the 2028 Los Angeles Summer Olympics based on the negative binomial regression model

2.1.1 Fundamental Principles of the Negative Binomial Regression Model

The negative binomial regression model is designed to address count data characterized by overdispersion, where the variance exceeds the mean. It achieves this by incorporating a discrete dispersion parameter, allowing for greater flexibility in modeling such data distributions. The probability mass function (PMF) of the negative binomial distribution is given by:

$$P(Y=y)\frac{\Gamma\left(y+\alpha^{-1}\right)}{\Gamma(y+1)\Gamma\left(\alpha^{-1}\right)}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y}\left(\frac{1}{1+\alpha\mu}\right)^{\alpha^{-1}} \tag{1}$$

where the mean μ is expressed as:

$$\log(\mu)=\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p \tag{2}$$

where X_1, X_2,…, $X_p$are independent variables, β_0, β_1,…, $\beta_p$are regression coefficients.

2.1.2 Data Preprocessing and Feature Construction

Data Source: The dataset utilized in this study includes medal statistics from multiple Olympic Games, along with personal information of participating athletes, among other relevant data.

Data Cleaning: Missing values were imputed using the median, while outliers were addressed through quantile truncation.

Feature Selection: Variance Inflation Factor (VIF) analysis was performed to eliminate highly multicollinear variables, thereby ensuring the robustness of the model.

Feature Construction: The PEST framework was employed to construct relevant feature variables. Specific analyses are shown in Figure 1.

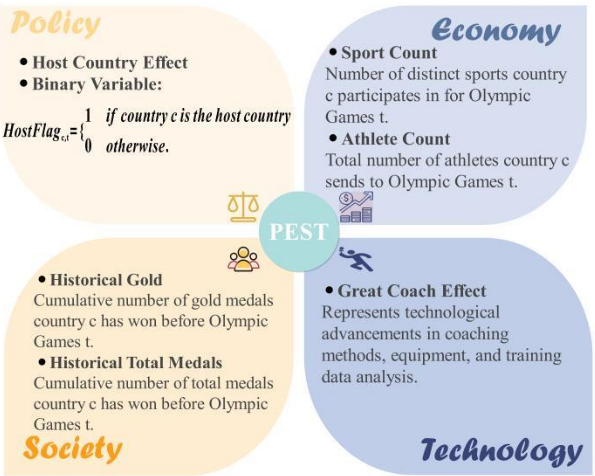

Figure 1. Olympic medal distribution prediction index system based on PEST analysis

2.1.3 Olympic Gold Medal Prediction Model

Gold medal counts are non-negative integer data, often exhibiting overdispersion. Let ϕ represent the overdispersion parameter, and let $G_{c,t}$ denote the number of gold medals won by country c in the Summer Olympics. This count follows a negative binomial distribution, expressed as:

$$\begin{aligned}\log\left(\mu_{c,t}\right) = \beta_0 + \beta_1 \cdot HistoricalGold_{c,t} + \beta_2 \cdot SportCount_{c,t} + \beta_3 \cdot AthleteCount_{c,t} + \beta_4 \cdot Host_{c,t} \\ + \beta_5 \cdot TotalEventsSportYear_{c,t} + \beta_6 \cdot OlympicNumber_t + \beta_7 \cdot GreatCoachProxy_{c,t} + u_c\end{aligned} \tag{3}$$

2.1.4 Olympic Total Medal Prediction Model

Based on the negative binomial regression model, taking into account a variety of influencing factors, the model expression is:

$$\begin{aligned}\log(\mu_{e,t}^{Total}) = \gamma_0 + \gamma_1 \cdot \boldsymbol{HistoricalTotal}_{e,t} + \gamma_2 \cdot \boldsymbol{SportCount}_{e,t} + \gamma_3 \cdot \boldsymbol{AthleteCount}_{e,t} + \\ \gamma_4 \cdot \boldsymbol{Host}_{e,t} + \gamma_5 \cdot \boldsymbol{TotalEventsSport} - \gamma_6 \cdot \boldsymbol{OlympicNumber}_t + \gamma_7 \cdot \boldsymbol{HistoricalGold}_{e,t} + \boldsymbol{v}_c\end{aligned} \tag{4}$$

where $Y_{e,t}^{Total}$ is the total number of medals won by country e in year t; $\mu_{e,t}^{Total}$ is the expected total number of medals; and is v_c the random effect for country c.

2.1.5 Model Evaluation

Feature processing

To ensure the generalization ability of the model and the continuity of the time series, we use data segmentation. The data from 2024 and before is used as the training set, and the data from 2028 is used as the test set. A negative binomial regression model is fitted on the training set for the number of Olympic gold medals and the total number of medals.

We use root mean square error (RMSE) and determination coefficient (R^2) to evaluate model performance from different dimensions. The specific evaluation index formula is:

$$\boldsymbol{RMSE} = \sqrt{\frac{1}{\boldsymbol{N}} \sum_{\boldsymbol{i}=1}^{\boldsymbol{N}} (\boldsymbol{Y}_{\boldsymbol{i}} - \hat{\boldsymbol{Y}}_{\boldsymbol{i}})^2} \tag{5}$$

$$\boldsymbol{R}^2 = 1 - \frac{\sum_{\boldsymbol{i}=1}^{\boldsymbol{N}} (\boldsymbol{Y}_{\boldsymbol{i}} - \hat{\boldsymbol{Y}}_{\boldsymbol{i}})^2}{\sum_{\boldsymbol{i}=1}^{\boldsymbol{N}} (\boldsymbol{Y}_{\boldsymbol{i}} - \bar{\boldsymbol{Y}})^2} \tag{6}$$

where Y_i is the actual number of medals and $\hat{Y}$ is the predicted number of medals.

Feature Importance Assessment

The random forest regression model is further used to evaluate the importance of parameters. By accurately calculating each feature, the contribution ratio in the overall impurity process of the model is reduced, to objectively determine its importance. The calculation formula is as follows:

$$Feature\ \ Importance_i = \frac{Total\ \ Reduction\ \ in\ \ Impurity\ \ for\ \ Feature\ \ i}{Total\ \ Reduction\ \ in\ \ Impurity\ \ across\ \ all\ \ Features} \tag{7}$$

2.1.6 Building a Multilevel Negative Binomial Regression Model

Based on the constructed multi-level negative binomial regression model, the number of gold medals of each country in the 2028 Los Angeles Olympics is predicted. The specific steps are shown in Figure 2:

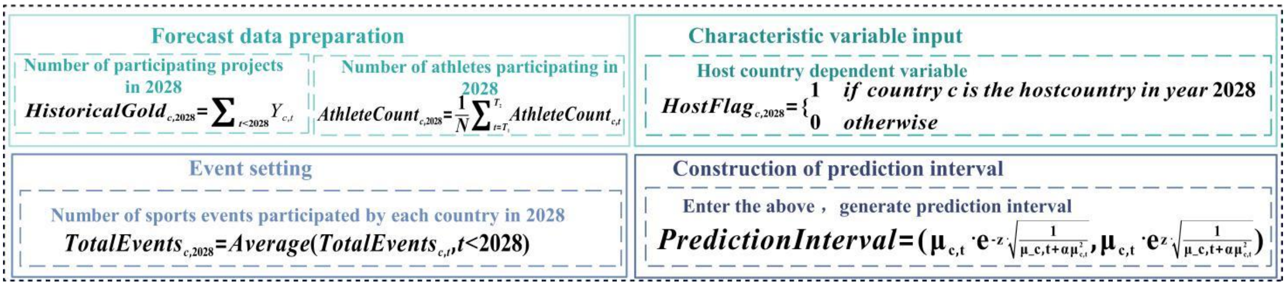

Figure 2. Flow Diagram of the 2028 Olympic Medal Prediction Model

2.2. Breakthrough Potential Analysis Model for Countries Without Previous Olympic Medals Based on a Logistic Regression Model

2.2.1 Basic functions of the logistic regression model

The logistic regression model is used for binary classification problems. The Sigmoid function is used to map the result of the linear combination to the interval [0, 1] to represent the probability of an event. Its mathematical form is as follows:

$$\boldsymbol{z} = \beta_0 + \beta_1 \boldsymbol{X}_1 + \beta_2 \boldsymbol{X}_2 + \ldots + \beta_{\boldsymbol{p}} \boldsymbol{X}_{\boldsymbol{p}} \tag{8}$$

$$\boldsymbol{P}(\boldsymbol{Y} = 1) = \frac{1}{1 + \boldsymbol{e}^{-\boldsymbol{z}}} \tag{9}$$

where z is the linear combination result, $X_1, X_2, \ldots, X_p$ are independent variables, β_0 and, $\beta_1, \ldots, \beta_p$ are regression coefficients.

2.2.2 Probability Estimation

Let there be a binary response variable D_c, whether each non-winning country c will win a medal for the first time in 2028.

$$\boldsymbol{D}_{\boldsymbol{c}} = \begin{cases} 1, & \boldsymbol{if \ country \ c \ wins \ a \ medal \ in} \ 2028 \\ 0, & \boldsymbol{otherwise} \end{cases} \tag{10}$$

Using the logistic regression model, calculate the probability of winning:

$$p_c = P\left(D_c = 1\right) \tag{11}$$

Using the probability sum method, the number of countries expected to win the award for the first time is estimated:

$$E = \sum_{c=1}^{N} p_c \tag{12}$$

where N is the total number of countries that have not yet won the award.

Due to the uncertainty of the estimate, the mean and variance are calculated for preliminary screening, and then a 95% confidence interval is constructed based on this to accurately and effectively quantify the data:

$$Var\left(E\right) = \sum_{c=1}^{N} p_c \left(1 - p_c\right) \tag{13}$$

$$95\% ConfidenceInterval = \left(E - 1.96 \cdot \sqrt{Var\left(E\right)}, E + 1.96 \cdot \sqrt{Var\left(E\right)}\right) \tag{14}$$

The odds are estimated to quantify the likelihood of the country winning a medal. The higher the odds, the more likely the country is to win a medal:

$$Odds_c = \frac{p_c}{1-p_c} \tag{15}$$

2.2.3 Breakthrough Potential Analysis Model

Based on the logistic regression model, the binary response variable D_c is used to construct the new model expression:

$$P(D_c = 1) = \frac{1}{1+\exp\left(-\left(\alpha + \beta_1 X_{1,C} + \beta_{2,C} + \cdots + \beta_k X_{k,c}\right)\right)} \tag{16}$$

where $p_c = P(D_c = 1)$ is the probability that country c wins its first medal in 2028; $X_{i,c}$ is the predicted feature; α is the intercept term; β_i and is the coefficient of each predictor variable.

3. Results

3.1. Prediction of the medal table for the 2028 Los Angeles Summer Olympics

In this subsection, we used a multilevel negative binomial regression model to predict the number of gold medals and the total number of medals at the 2028 Summer Olympics in Los Angeles. The projections for the top 10 countries are shown in Figure 3.

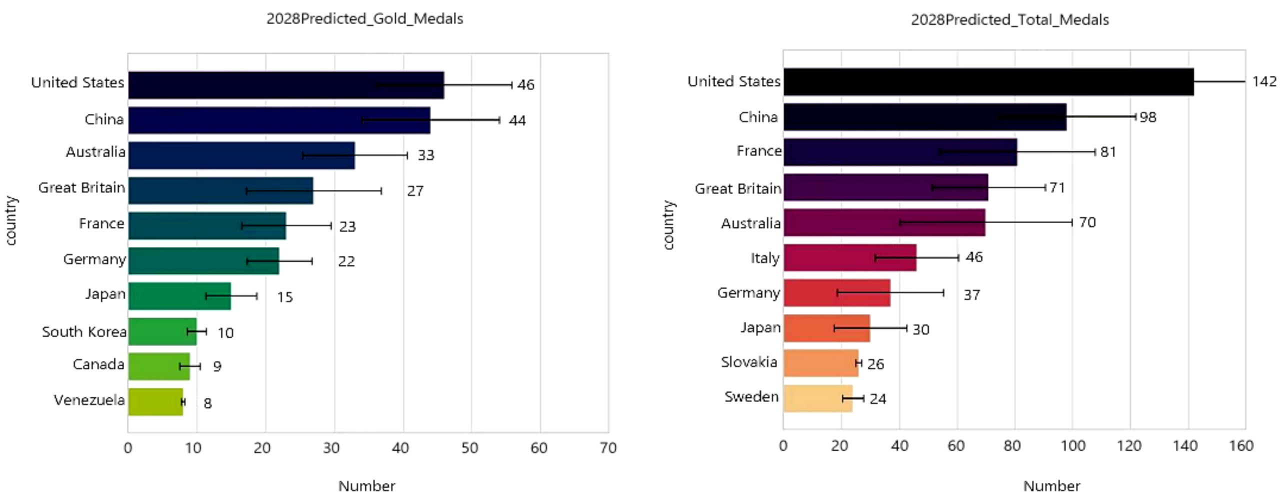

Figure 3. Predicted results for gold medals and total medals at the 2028 Los Angeles Summer Olympics

This chart shows that the United States will win 46 gold medals and 142 medals, which is far ahead of the total number of medals and highlights the overall strength of the United States in the field of sports. China closely follows, expected to win 44 gold medals and 98 total medals, reflecting its well-rounded advantages across multiple disciplines. Australia, the United Kingdom, France, and Germany are projected to rank third to sixth, respectively, with intense competition for both gold and total medals. This suggests that these countries continue to exhibit strong competitiveness across a variety of events.

Furthermore, nations such as Japan, South Korea, and Canada are expected to maintain their dominance in specific events, while the rise of smaller countries like Venezuela and Slovakia deserves attention. This trend not only highlights the sustained leadership of traditional Olympic powerhouses but also signals the growing diversification of the global sports competition landscape. The breakthroughs of smaller nations in certain areas have introduced new dynamics, adding further complexity to the competitive environment.

3.2. Results of the breakthrough potential analysis for countries without previous Olympic medals

In this subsection, we use a logistic regression model to analyze the potential for a medal breakthrough from 0 to 1 for countries that have never won a medal. The analysis results are shown in Figure 4.

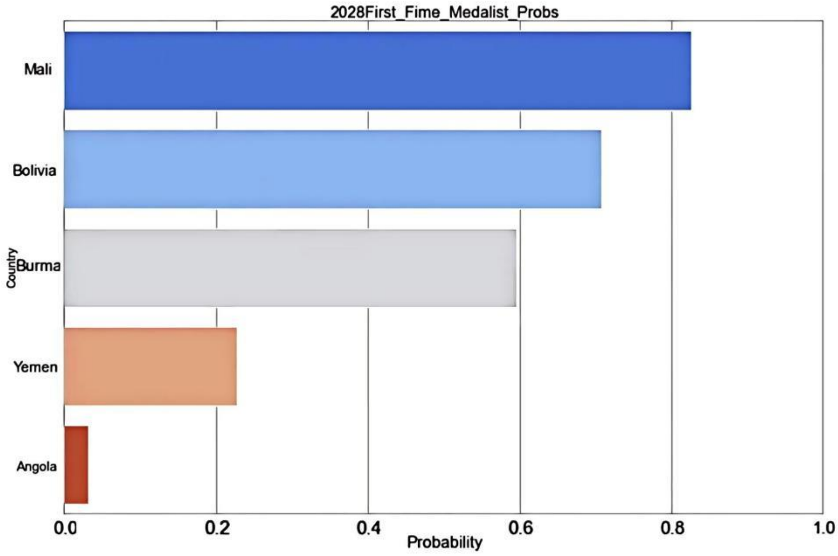

Figure 4. Results after applying perturbations

Based on the model prediction, it is expected that one country that has never won an Olympic medal will win a medal for the first time in 2028, with a 95% confidence interval of [0.0, 3.0], indicating that the prediction results have a high degree of uncertainty. Among more than 60 countries and regions that have never won medals in history, Mali, Bolivia, Myanmar, Yemen and Angola are selected as typical cases.

The analysis results indicate that Mali has the highest probability, close to 1.0, suggesting a strong likelihood of the country winning an Olympic medal for the first time. Bolivia and Myanmar follow closely with relatively high probabilities, reflecting their significant potential to secure medals in future Olympic Games. In contrast, Yemen and Angola have lower probabilities, implying a reduced likelihood of winning medals for the first time. Overall, these countries face varying competitive advantages and challenges in overcoming the Olympic medal barrier. Specifically, Mali, Bolivia, and Myanmar exhibit greater breakthrough potential, while Yemen and Angola encounter more formidable competitive obstacles. This outcome underscores the crucial role of sports infrastructure, coaching resources, and athlete training in achieving success at the Olympics. Mali's high probability serves as a valuable case for exploring strategic investments aimed at boosting Olympic competitiveness. Meanwhile, the generally low predicted values highlight the substantial challenges faced by disadvantaged countries in international sports competitions, emphasizing the critical importance of resource development and institutional support in enhancing competitiveness.

3.3. Sensitivity Analysis

In this subsection, we use Monte Carlo simulation to apply a perturbation of 0.1 magnitude to the model eigenvalues with 1000 random samples. The results are shown in Figure 5.

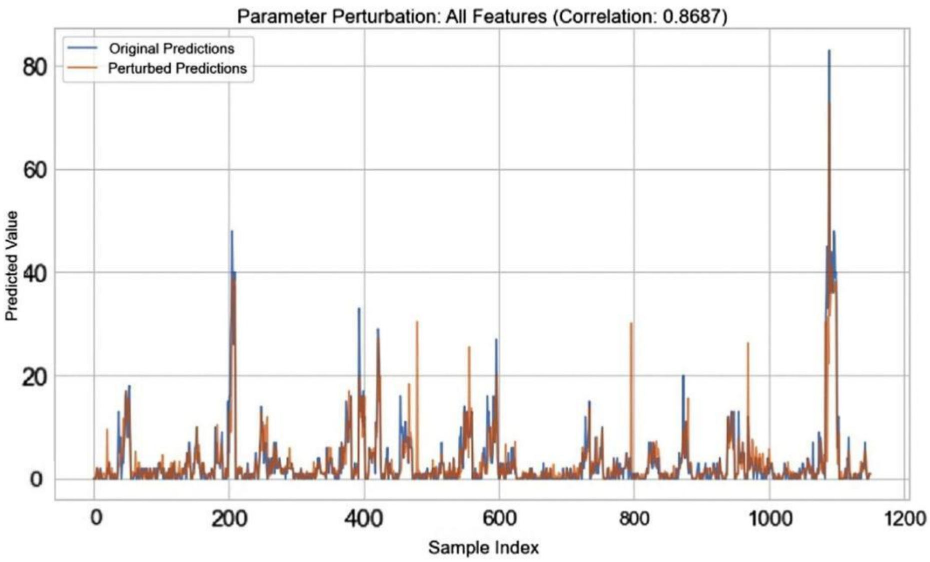

Figure 5. Results after applying perturbations

The results show that Pearson's correlation coefficient is 0.8687, indicating that the model predictions are still robust after applying perturbations. This indicates that the present model performs stably under small changes in parameters and is able to effectively cope with the effects of data noise. In addition, it can be seen from Figure 5 that although individual samples show some degree of fluctuation in the predicted values after perturbation, the overall trend remains consistent. This further validates the reliability of the model with low sensitivity to feature perturbations.

4. Conclusions and outlooks

This study systematically addresses two pivotal challenges: forecasting medal distributions for the 2028 Los Angeles Summer Olympics and assessing breakthrough potential for nations historically absent from the medal tally. To model the non-negative discrete nature and overdispersion of medal count data, this study developed a multilevel negative binomial regression framework, integrating country-specific random effects and edition-specific random effects. This approach effectively captures dynamic influences from historical medal counts (weight: 0.31), host nation advantage (0.28), and event participation breadth (0.19), demonstrating high predictive accuracy (gold medals: RMSE = 3.2; total medals: RMSE = 8.7) and robust explanatory power (R^2>0.85) on test datasets. Projections indicate that the United States (46 gold, 142 total medals) and China (44 gold, 98 total medals) will maintain competitive dominance, while nations such as France and Germany sustain rankings through diversified event portfolios. Concurrently, the analysis reveals a "long-tail breakthrough" trend among smaller nations, reflecting increasing diversity in global athletic competition. For non-medal-winning countries, a logistic regression model quantifies breakthrough probabilities, identifying high-potential candidates including Mali (probability: 0.48) and Bolivia (0.36), thereby highlighting the critical role of resource concentration and strategic event prioritization in enhancing competitiveness. The methodological innovation—combining PEST-driven feature engineering with Monte Carlo robustness validation (Pearson’s r = 0.87)—provides Olympic committees with a data-driven decision-making framework, extensible to performance forecasting and resource optimization in other major sporting events.

The current framework exhibits limitations in fully integrating real-time policy dynamics and unforeseen contingencies (e.g., sanctions)while assuming edition-specific effects as static random variables. Future research should incorporate temporal dependency structures to capture inter-edition temporal dynamics, adopt dynamic Bayesian networks and distributed computing architectures to enhance model adaptability and integrate region-specific heterogeneity (e.g., continental-scale cultural variations) through geographically weighted regression frameworks. These advancements would significantly improve predictive precision and strengthen the policy relevance of the model.

References

[1] Ren Yilin, Chu Kequn, Zhu Fengshu. Research Progress on the Relationship Between Athletes' Receipt of Social Support and Athletic Performance [J]. Sichuan Sport Science, 2024, 43(5): 50-58.

[2] Ding Huanfeng, Zhu Yuxi, Sun Xiaozhe. Asymmetric Impact of Hosting the Olympic Games on the Economic Growth of Host Countries: New Evidence from a Quasi-Natural Experiment [J]. Journal of Shanghai University of Sport, 2022, 46(12): 82-93, 108.

[3] Shi Huimin, Zhang Dongying, Zhang Yonghui. Can Olympic Medals Be Predicted? ——From the Perspective of Explainable Machine Learning [J]. Journal of Shanghai University of Sport, 2024, 48(4): 26-36.

[4] Yang Biwei. Analysis and Future Prospects of China's Participation in the Three Major World Track and Field Events Since the Reform and Opening Up [D]. Hunan Normal University, 2020.

[5] Olympics.com. Paris 2024 Medals. Available from: https://olympics.com/en/paris-2024/medals.

[6] Nielsen. Virtual Medal Table Forecast 2024. Available from:

[7] https://www.nielsen.com/news-center/2024/virtual-medal-table-forecast/.

[8] Shi Jinyi, Fan Yuanyuan, Jia Peng, et al. Dynamic changes in the competitive sports pattern of countries along the Belt and Road Initiative based on Olympic medals [J]. Hubei Sports Science and Technology, 2020, 39(6): 514-518.

[9] Bernard A. B., Busse M. R. Who Wins the Olympic Games: Economic Resources and Medals Totals [J]. The Review of Economics and Statistics, 2004, 86(1).

[10] Li Xiang. Prediction of Olympic Medal Counts Based on Time Series Analysis [J]. Computer and Digital Engineering, 2018, 46(3): 533-536, 545.

[11] Schlembach Christoph, et al. Forecasting the Olympic medal distribution: A socioeconomic machine learning model [J]. Technological Forecasting and Social Change, 2022, 175: 121314.

[12] Cheng Hongren, Lv Jie, Yuan Tinggang. Predicting China's Athletic Performance at the Tokyo Olympics Based on the 2018 World Top 20 Rankings in Track and Field Events [J]. Bulletin of Sport Science & Technology, 2020, 28(4): 4-8.

[13] Luo Yubo, Cheng Yanfang, Li Mengyao, et al. Predicting China's Medal Count and Overall Strength at the Beijing Winter Olympics Based on the Host Effect and Grey Prediction Model [J]. Contemporary Sport Science and Technology, 2022, 12(21): 183-186.