Research On Olympic Medal Prediction Based on Random Forest-ARIMA Combined Model

Yanru Lin [1], Dingcheng Ye [1], Xingang Jia [2, *]

[1] Chien-Shiung Wu College of SEU, Southeast University, Nanjing, China, 210096

[2] School of Mathematics, Southeast University, Nanjing, China, 210096

* Corresponding Author Email: 101010344@seu.edu.cn

Abstract. The prediction of Olympic medals is complex due to the multi-dimensional influencing factors and the non-linear feature relationships. Therefore, a combined model capable of handling both non-linear feature relationships and capturing time series trends is required. Specifically, this study predicts the medal table for the 2028 Los Angeles Olympics based on a Random Forest-ARIMA combined model. After cleaning and integrating past data using non-negative least squares in a "property division" manner, the United States, China, and Japan are predicted to rank first, second, and third respectively, with a total of 92, 61, and 33 medals, and 37, 28, and 14 gold medals respectively. The United States, China, and Brazil are expected to continue their upward trends, while the United Kingdom will maintain a stable number of medals. Australia, Italy, and Japan have experienced greater medal fluctuations in recent Olympics and are expected to continue this trend. SAM is most likely to win its first medal, with a probability of approximately 72.3%, and three countries are expected to win their first medals.

Keywords: Random Forest-ARIMA Hybrid Model, Medal Tally Prediction, Non-negative Least Squares Method, "Dividing Property-style" Organization.

1. Introduction

As the most significant global multi-sport event, the Olympic Games' medal distribution pattern reflects the comprehensive sports strength of various countries. Accurately predicting the medal performance of countries in future Olympic Games is of great significance. Traditional prediction methods mainly rely on simple statistical analysis or expert experience, which are difficult to comprehensively consider the combined effects of multiple influencing factors. With the rapid development of machine learning technology, new paths have been provided for building more scientific prediction models. However, Olympic data has its unique complexity: on the one hand, historical data contains special circumstances such as country splits and mergers, changes in Olympic Committee codes, and medals awarded to refugees and independent athletes, which affect data consistency; on the other hand, the advantages of various countries in different events evolve over time and are influenced by multiple factors such as the host country effect, gender balance, and new events. These characteristics require prediction models to be capable of handling complex multi-dimensional feature relationships and capturing trend changes in long-term time series, which is often difficult for a single model to achieve simultaneously [1,2].

To more accurately predict the Olympic medal tally, this study established a random forest-ARIMA integrated model to meet the above two requirements [3]. To clean and integrate historical Olympic data, this study used non-negative least squares to correct the historical data of various countries based on their data before and after changes due to various factors that prevented them from continuing to participate in the Olympic Games, ensuring the consistency and continuity of the data. In addition, medals awarded for special reasons were ignored as they were not representative. After solving historical and political legacy issues, this study established a random forest-ARIMA combined model based on given historical data [4,5]. This model conducted multi-dimensional fitting to obtain as accurate a conclusion as possible. Additionally, the model also predicted the countries that might win their first medal at the 2028 Los Angeles Olympics.

2. Olympic Historical Data Cleaning and Integration

The data in this article is sourced from https://www.comap.com/contests/mcm-icm.

2.1. Integration of NOC

Upon reviewing the data, this study found that many participants who actually belonged to the same country were assigned to different teams. Additionally, many countries changed their National Olympic Committees (NOCs) due to various historical reasons, political factors, etc. Given the continuity of these countries, regions, and participants, this study uniformly uses NOCs as the classification basis for the participating teams. For the changes in the same country's NOCs, this study uniformly modifies the previous NOCs to the latest ones, such as MAL, NBO, and MAS representing different NOCs of the same country. This approach can reduce the influence of outliers and maintain the consistency of the data.

2.2. Integration of Individuals and Refugees Competing

Upon reviewing the data, this study found that there were many individual and refugee participants whose National Olympic Committees (NOCs) included AIN, IOA, ROT, etc. The awards won by these participants were not included in the statistics of the medal table, and the number of winners was relatively small. Therefore, this study deleted these data and did not use them for model learning and calculation. For individual participating delegations such as EUN and ROC, their participants still competed in the name of the original country as individual athletes, and there were a large number of participants and a considerable number of medals won. Therefore, this study included them in the original countries for model learning and calculation, in order to avoid the influence of outliers and maintain data consistency.

2.3. Dissolution and Consolidation of Nations

During the process of reviewing the data, this study found that over the 128 years, among the 33 Olympic Games, many powerful sports nations either merged or split, leaving only the historical records. However, upon a closer examination of these historical records and the data of the newly formed countries, this study discovered a close correlation between them. For the merged countries, such as East Germany and West Germany, the correlation before and after the merger was obvious, and the equivalent historical data of the merged Germany could be obtained simply by adding up:

$$y_{c,t} = y_{c1,t} + y_{c2,t} \tag{1}$$

$$p_{c,t} = p_{c1,t} + p_{c2,t} \tag{2}$$

For divided countries such as the Soviet Union and Yugoslavia, it cannot be simply explained by simple addition. This study uses a non-negative least squares model to handle the situation.

Obviously, the number of medals won by each country is a non-negative number. However, because the new countries formed by the division and disintegration "divide the property", the number of medals won by each country calculated by the model is a virtual number. Therefore, in this study, it is believed that the medal numbers and participants of the original countries inherited by the new countries can be decimal numbers. Thus, this study introduces the non-negative least squares model [6]:

$$\min \left\| \begin{pmatrix} M_{ci,t} \\ W_{ci,t} \end{pmatrix} - t * \begin{pmatrix} \beta_{Mti} \\ \beta_{Wti} \end{pmatrix} \right\|^2 \tag{3}$$

$$\min \left\| \begin{pmatrix} G_{ci,t} \\ S_{ci,t} \\ B_{ci,t} \end{pmatrix} - \begin{pmatrix} \beta_{GMi} & \beta_{GWi} & \beta_{Gti} \\ \beta_{SMi} & \beta_{SWi} & \beta_{Sti} \\ \beta_{BMi} & \beta_{BWi} & \beta_{Bti} \end{pmatrix} * \begin{pmatrix} M_{ci,t} \\ W_{ci,t} \\ t \end{pmatrix} \right\|^2 \tag{4}$$

$$\begin{cases} \begin{cases} \begin{pmatrix} M_{ci,t} \\ W_{ci,t} \end{pmatrix} = t * \begin{pmatrix} \beta_{Mti} \\ \beta_{Wti} \end{pmatrix} \\ \begin{pmatrix} M_{co,t} \\ W_{co,t} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{x} M_{ci,t} \\ \sum_{j=1}^{x} W_{cj,t} \end{pmatrix} \\ M_{ci,t} \geq 0 \\ W_{ci,t} \geq 0 \end{cases} \\ \begin{cases} \begin{pmatrix} G_{ci,t} \\ S_{ci,t} \\ B_{ci,t} \end{pmatrix} = \begin{pmatrix} \beta_{GMi} & \beta_{GWi} & \beta_{Gti} \\ \beta_{SMi} & \beta_{SWi} & \beta_{Sti} \\ \beta_{BMi} & \beta_{BWi} & \beta_{Bti} \end{pmatrix} * \begin{pmatrix} M_{ci,t} \\ W_{ci,t} \\ t \end{pmatrix} \\ \begin{pmatrix} G_{co,t} \\ S_{co,t} \\ B_{co,t} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{x} G_{ci,t} \\ \sum_{j=1}^{x} S_{cj,t} \\ \sum_{k=1}^{x} B_{ck,t} \end{pmatrix} \\ G_{ci,t} \geq 0 \\ S_{ci,t} \geq 0 \\ B_{ci,t} \geq 0 \end{cases} \end{cases} \tag{5}$$

Among them, $M_{ci,t}.W_{ci,t}.G_{ci,t}.S_{ci,t}.B_{ci,t}$ represent the number of male and female athletes participating in the competition, the number of gold medals, silver medals, and bronze medals that the i-th newly formed country should receive, respectively. t represents the number of the Olympic Games, and β is the corresponding coefficient.

Therefore, this study can obtain the corresponding non-negative least squares regression coefficients, as well as the number of male and female athletes participating in the competition, the number of gold medals, silver medals, and bronze medals for each country. As shown in Figure 1, after the disintegration of the Soviet Union, the "property-sharing" integrated data of the main countries regarding male and female athletes from the Soviet Union, as well as the number of gold, silver and bronze medals, were obtained. These were the results distributed to each country.

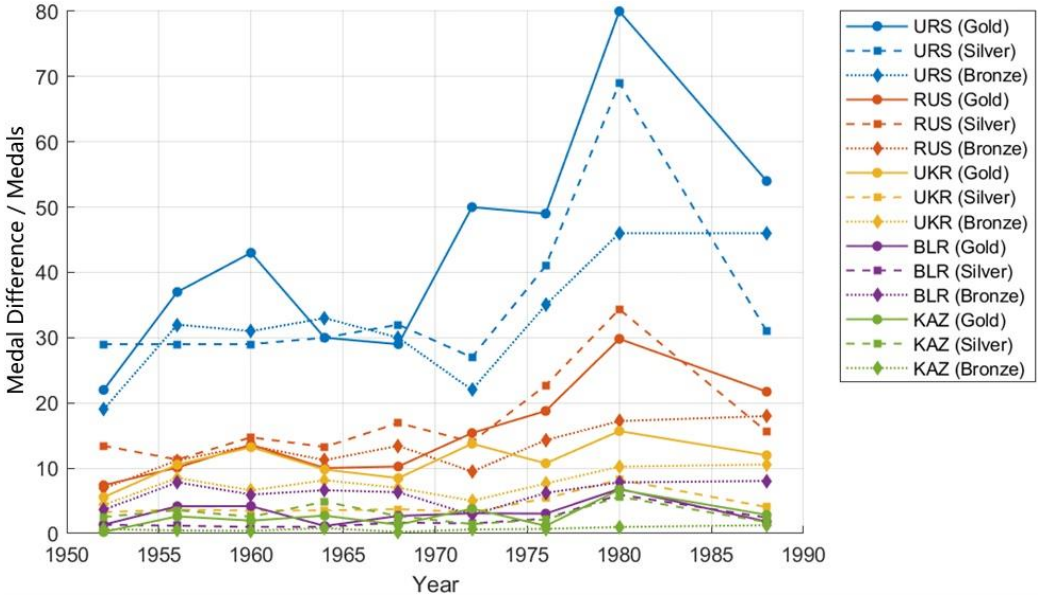

Figure 1: Medal Difference by Year (Soviet and Related Countries)

2.4. Interim Olympics and Special Awards

During the process of reviewing the data, this study found that there was an informal Olympic Games - an inter-Olympic Games - between the 3rd and 4th Olympic Games. This Olympic Games had no medal list and did not follow the regular pattern of being held every four years. Therefore, it was not taken into consideration. Additionally, the Aeronautics project, which awards to those who

have made significant contributions to the aviation field, and the Alpinism project, which awards to those who have conquered high mountains, etc., were also not considered due to their accidental nature. This was done to reduce the influence of outliers and maintain the consistency of the data.

3. Establishment and Validation of Medal Prediction Model

In order to effectively predict the medal table of the 2028 Los Angeles Olympics, this study established a random forest-ARIMA integrated model based on the previously "property-based" cleaned and integrated data and the released information related to the Los Angeles Olympics. This model takes into account historical medal factors, historical number of participating athletes' factors, time factors, as well as event type and quantity factors, gender factors, national conditions factors, and host country effects. Moreover, the model also predicts the countries that may win the first medals at the 2028 Los Angeles Olympics. The specific contents of the model and the results will be elaborated below.

3.1. Construction of Random Forest-ARIMA Combined Model

The ARIMA model is a widely used time series analysis method, suitable for handling continuous data with temporal correlations. In this paper, the ARIMA model is employed to observe the data change trends on the overall timeline of various countries, compensating for the deficiency of macro time parameters in the random forest model. By adjusting the values of the AR term and MA term orders, the autocorrelation and random fluctuations in the time series are captured. Then, combined with the predicted values calculated by the random forest and the original data samples, the ARIMA model is trained, and the change curves of the number of medals won by various countries over the years and in the coming several sessions are obtained for prediction and evaluation [7,8].

3.2. Estimation of Participant Numbers for the Los Angeles Olympics

Table.1. Sports for the Los Angeles Olympics

Major Events	Major Events	Major Events	Major Events	Major Events
3x3 Basketball	BMX Racing	Flag Football	Rowing	Swimming
Artistic Gymnastics	Canoe Slalom	Football (Soccer)	Rhythmic Gymnastics	Table Tennis
Archery	Canoe Sprint	Golf	Rugby Sevens	Taekwondo
Artistic Swimming	Coastal Rowing	Handball	Sailing	Trampoline Gymnastics
Athletics	Cricket	Hockey	Shooting	Tennis
Badminton	Cycling Road	Judo	Skateboarding	Triathlon
Baseball	Cycling Track	Lacrosse Sixes	Softball	Volleyball
Basketball	Diving	Squash	Sport Climbing	Water Polo
Beach Volleyball	Equestrian	Modern Pentathlon	Marathon Swimming	Weightlifting
BMX Freestyle	Fencing	Mountain Bike	Surfing	Wrestling

According to the official website of the Los Angeles Olympics https://la28.org, this study reveals the names of the 50 major events that have been announced for the Los Angeles Olympics, as shown in Table 1:

In comparison to the overall disciplines of the 2024 Paris Olympics, this study revealed that during the Los Angeles Olympics, there was no mention of boxing or breakdancing being added. Instead, seven different disciplines were introduced: Baseball, Coastal Rowing, Cricket, Flag Football, Lacrosse Sixes, Softball, and Squash. Among these, Baseball and Softball had previously been part of the Olympic program, whereas the remaining five disciplines were entirely new additions with no prior reference data available. Consequently, this study concludes that these five newly introduced events

align with the host country's strengths. For the other events, the non-negative least squares method (as outlined in the data preprocessing phase) was employed to estimate participant numbers, as depicted in Figure 2. This dataset was then utilized as input for training the predictive model.

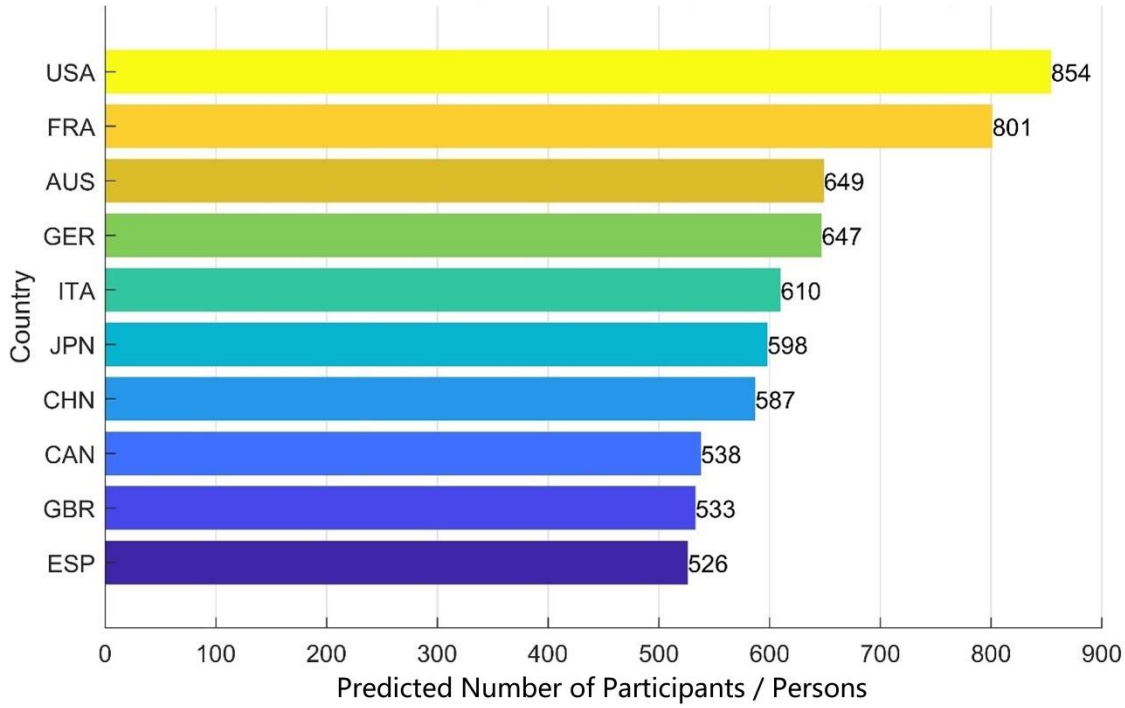

Figure 2: Example of Participation Projections by Country for the Olympic Games Los Angeles

3.3. Prediction of the Medal Standings for the Los Angeles Olympics

Based on the Random Forest-ARIMA integrated model, this study made predictions for the number of gold medals, silver medals, bronze medals, and total medals won by various countries at the 2028 Los Angeles Olympics. The prediction steps are as follows:

Firstly, feature input was conducted. Specifically, the study utilized preprocessed data along with the estimated number of participants in the Los Angeles Olympics to form the dataset. For model training, feature vectors such as NOC, YEAR, SPORT, $G_{c,t}.S_{c,t}.B_{c,t}.M_{c,t}.W_{c,t}.P_{c,t}$, ISHOST, etc., were employed.

Secondly, decision tree prediction and statistical analysis were performed. After training the Random Forest-ARIMA integrated model, the optimal number of decision trees was determined to be $T = 297$. The predicted medal counts from each tree were then used to construct a density distribution graph.

Lastly, the prediction interval was calculated. By leveraging the density distribution, the study derived an 80% confidence interval, which provided prediction ranges for the number of gold, silver, bronze, and total medals won by different countries [9,10].

As illustrated in Table 2, this study predicts the medal ranges for the top 20 medal-winning countries at the 2028 Los Angeles Olympics. Among them, the United States, China, and Japan are expected to rank in the top three positions, followed by Australia, the United Kingdom, and France.

Table.2. Medal counts' prediction interval for the Olympic Games Los Angeles 2028

NOC	Gold/ Medals	Silver/ Medals	Bronze/ Medals	Medal/ Medals
USA	36.79147	31.67453	24.05849	92.52449
CHN	27.55482	19.02684	15.10564	61.68729
JPN	14.00402	8.762364	10.35219	33.11857
AUS	12.51338	13.52703	13.10012	39.14053
GBR	8.667564	12.74023	14.33416	35.74196
FRA	8.344827	11.68006	12.15263	32.17752
ITA	6.726288	10.15928	10.25296	27.13853
CAN	6.396238	6.704027	8.87295	21.97322
NED	6.102924	5.939134	7.466516	19.50857
GER	6.057879	9.820379	7.782224	23.66048
ESP	4.79842	5.072316	5.628714	15.49945
BRA	4.032925	5.217965	5.881435	15.13232
KOR	3.944384	4.098832	4.270631	12.31385
KEN	3.744386	3.188104	3.218698	10.15119
NZL	3.080452	3.029678	3.070729	9.180859
POL	2.792702	3.292281	3.840268	9.925251
HUN	2.791537	2.931449	3.257167	8.980153
BEL	2.596773	2.652657	3.070136	8.319566
CZE	2.03709	2.242317	2.532008	6.811416
ROU	1.820914	2.1971	1.663216	5.68123

As shown in Figures 3 and 4, the predicted intervals for gold medals and total medals of each country are presented. As shown in Figures 5 and 6, the probability density distribution graphs of the predicted number of gold medals for China and the United States in 2028 are depicted. From the figures, it can be seen that the predicted number of gold medals for China and the United States should be 28 and 37 respectively, and the highest probability density value is 37 and 46.

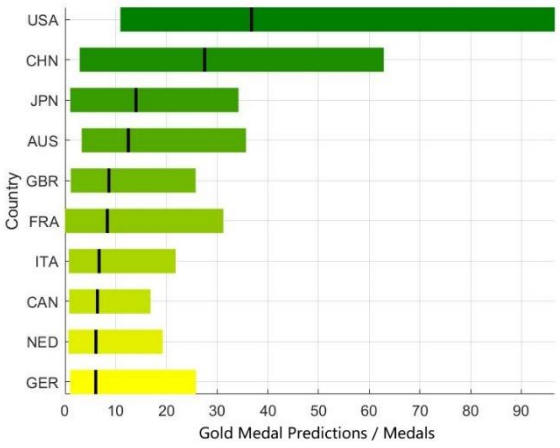

Figure 3: Gold Medal Prediction Interval

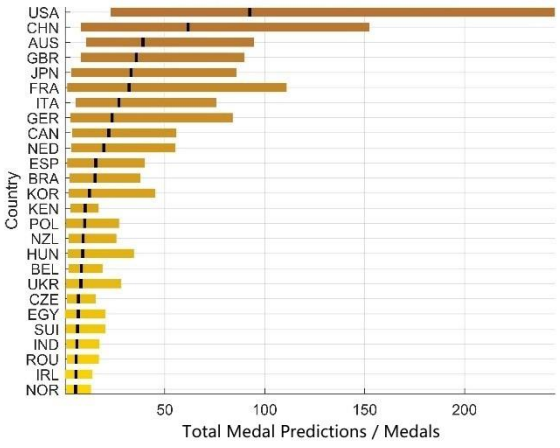

Figure 4: Total Medal Prediction Interval

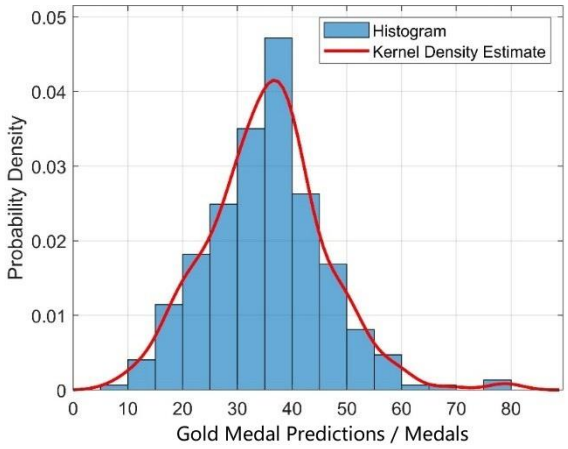

Figure 5: Probability Distribution of Chinese Gold Medals Predictions

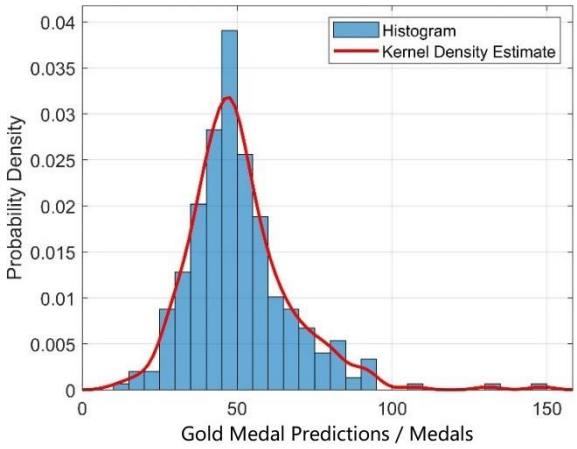

Figure 6: Probability Distribution of USA Gold Medal Predictions

3.4. Prediction of National Performance in the Los Angeles Olympics

In the process of training the integrated random forest-ARIMA model for predicting medal rankings at the Los Angeles Olympics, this study generated line graphs illustrating the fluctuations in gold medal counts for each country. Based on these visualizations, assessments were made regarding the advancements and declines in performance.

As depicted in Figure 7, nine countries with consistently strong overall performances in recent years were selected based on their total medal tallies. It is evident that the ARIMA model successfully identified the trends in medal variations for these nations across recent Olympic Games. The predictions indicate that Brazil is likely to sustain an upward trajectory, while the United Kingdom will maintain a relatively stable medal count. In contrast, Australia, Italy, and Japan have exhibited significant fluctuations in their medal counts during recent Olympics, a pattern expected to persist in the future. Both China and the United States demonstrated steady growth in their medal achievements. France, the previous host nation, may see its medal count revert to a more typical level following the loss of hosting privileges. In summary, the United States, China, Japan, and Brazil are anticipated to show improvement, whereas France, Italy, and Australia might experience a decline in performance.

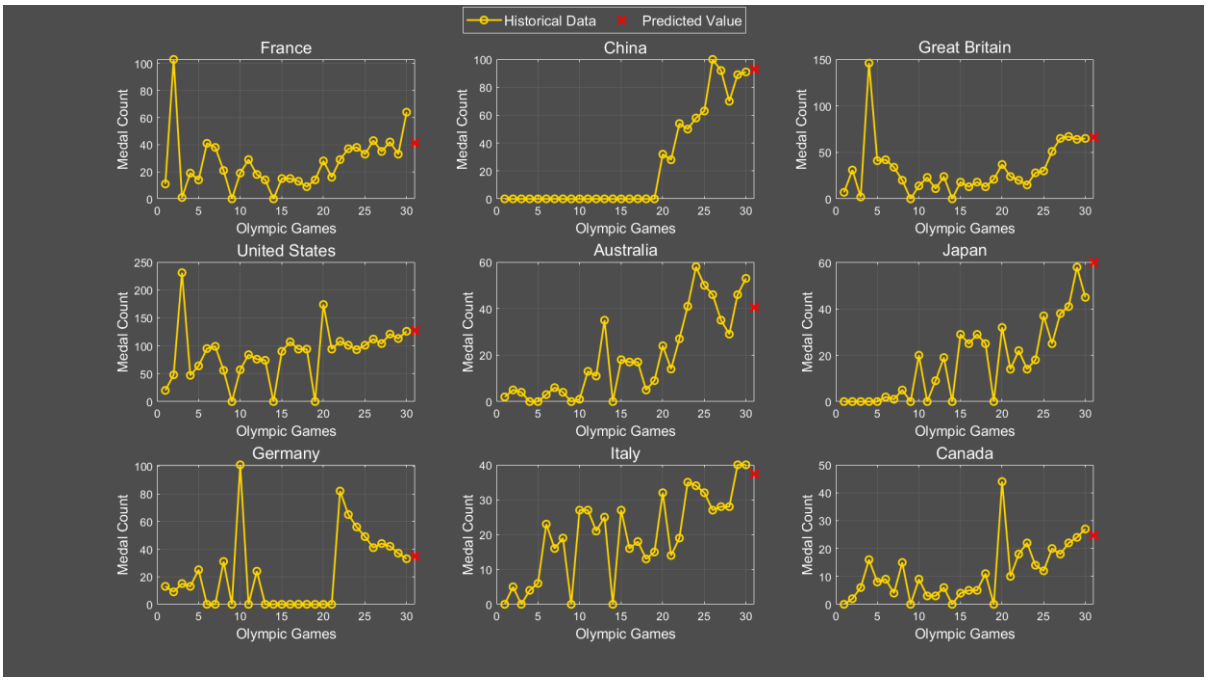

Figure 7: The TOP 9 Countries' Olympics Medal Counts

3.5. Prediction of First Medal Acquisition

Based on the predicted range of the Olympic medals awarded in Los Angeles obtained from the above, this study has derived the total medal count range for countries that have not yet won any medals. At this point, this study uses the Newton bisection method to bound the probability density functions of the awards received by various countries. Eventually, the maximum probability density obtained represents the probability of winning an award for that country. and obtains Table 3 showing the probability and odds of winning medals for countries with a higher probability:

Table.3. Probabilities and odds for first medal acquisition

NOC	Probability/%	Odd	NOC	Probability/%	Odd
SAM	72.30%	2.6101	LBN	16.70%	0.2005
ANG	42.71%	0.7455	PLE	15.43%	0.1825
MLI	26.22%	0.3554	ESA	13.95%	0.1621
GUI	22.41%	0.2888	GBS	11.21%	0.1263
PNG	19.03%	0.2350	NCA	10.78%	0.1208
GUM	18.39%	0.2253	NEP	8.46%	0.0924

An analysis was conducted on the probability of winning medals for all teams that have not previously won any medals. This study predicts that three teams will win their first medal [11].

4. Conclusions

This study cleaned and integrated the historical data of the Olympic Games using the non-negative least squares method, and then, based on the obtained data, used the random forest-ARIMA combined model to predict the medal table of the 2028 Los Angeles Olympics. The following conclusions were drawn: The United States, China, and Japan ranked in the top three, with a total of 84, 74, and 37 medals respectively; American Samoa is most likely to win its first medal, with an approximately 72.3% probability. Approximately three countries are expected to win their first-ever medal. However, due to the failure to consider the actual national conditions of each country and other objective factors such as the international situation, as well as the lack of comparison with other deep learning models, the data predicted in this study may be inaccurate. In the future, this research will combine other deep learning models and consider more objective factors to further improve the prediction results.

This paper provides a research idea and framework for application in related fields such as sports statistics, proving the feasibility of the random forest-ARIMA combined model for predicting the Olympic medal table.

References

[1] Zhu M N, Peng T, Chen K. Mathematical Analysis of the Influencing Factors of the Olympic Medal Table[J]. Contemporary Sports Technology, 2017, 7(27): 239-243.

[2] Shi H M, Zhang D Y, Zhang Y H. Can Olympic medals be predicted? --An interpretable machine learning perspective[J]. Journal of Shanghai University of Physical Education and Sports, 2024, 48(04): 26-36.

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

[4] Wang M, Pan J, Li X, et al. ARIMA and ARIMA-ERNN models for prediction of pertussis incidence in mainland China from 2004 to 2021[J]. BMC Public Health, 2022, 22(1): 1447.

[5] Schaffer A L, Dobbins T A, Pearson S. Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: a guide for evaluating large-scale health interventions[J]. BMC Med Res Methodol, 2021, 21(1): 58.

[6] Liang Y Z, Liu L, Peng L, et al. Research on Weighted Bayesian Inversion Algorithm with Non-Negative Least Squares Constraints[J]. Acta Photonica Sinica, 2020, 49(10): 216-226.

[7] Liu H. Sea surface height prediction based on the joint ICEEMDAN-ARIMA[J]. BEIJING SURVEYING AND MAPPING, 2025, 39(04): 436-442.

[8] Hoarau A, Martin A, Dubois J, et al. Evidential Random Forests[J]. Expert Systems with Applications, 2023, 230:120652.

[9] Xu Y, Li H, Lin C, et al. CellBRF: a feature selection method for single-cell clustering using cell balance and random forest[J]. Bioinformatics, 2023, 39(39 Suppl 1): i368-i376.

[10] Yu W L, Gao J, Wang R T. Internet of Things traffic classification based on lightweight random Forest algorithm[J]. Computer Engineering and Design, 2024, 45(12): 3553-3559.

[11] Csurilla G, Ferto I. How to Win the First Olympic Medal? and the Second[J]. SOCIAL SCIENCE QUARTERLY, 2024, 105:1544-1564.