Vol. 14 (2026): Proceedings of the 2nd International Conference on Artificial Intelligence, Modern Engineering and Environmental Sustainability (IC-AIMEES 2026)
E-ISSN: 3105-0522 P-ISSN：3105-0514
DOI: https://doi.org/10.70267/ ic-aimees.202602105123

Unveiling the Future of Olympic Glory

Anlan Zheng[1,*,†] and Yiran Tang[2,†]

[1]*College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, Zhejiang 310023, China*

[2]*College of Information Engineering, Zhejiang University of Technology, Hangzhou, Zhejiang 310023, China*

**Corresponding author: Anlan Zheng*

[†]*These authors contributed equally to this work and share first authorship*

Abstract

As the 2028 Los Angeles Olympics approach, accurately predicting the distribution of medals has become a critical issue in the fields of sports science and decision-making. This study addresses the challenges traditional models face in capturing the cyclical nature of Olympic data and the influence of political factors by proposing a multi-model integration forecasting framework. The aim is to improve the accuracy of predictions for the 2028 medal tally by quantifying historical trends, external political variables, and coaching effects, while identifying countries with the potential to achieve a "breakthrough" in their medal count. The study employs Seasonal and Trend Decomposition using Loess (STD-Loess) and Seasonal Differencing to construct an ARIMAX-based Seasonal Empirical Forecasting Model that incorporates participation frequency for predicting the total number of medals won by each country. It integrates Markov Chains and Bayesian updating to construct the First Medal Prediction Model for forecasting the first medal won. A Tobit model and the PageRank algorithm are used to evaluate the contribution of specific sports disciplines (SDE), and segmented regression is employed to quantify the "great coach" effect. The Seasonal Empirical Forecasting Model outperformed XGBoost in terms of prediction accuracy. Forecasts indicate that the United States (98 medals, range [37, 145]) and China (86 medals, range [43, 130]) will lead the medal count, while the Czech Republic is projected to achieve a 120% increase in medals. Nepal is deemed the most likely country to win its first medal in 2028, with odds of 1.23. Furthermore, the study found that men's basketball accounts for as much as 57% of the U.S. team's total medal count and confirmed the critical role of coaching intervention in achieving a breakthrough in gold medals. Through an innovative model integration approach, this study provides National Olympic Committees with a scientific basis for resource allocation and recommendations for coach recruitment, while profoundly revealing the centralization of modern Olympic medal distribution and the trend toward international competition.

Keywords

medal predictions, ARIMAX, Markov Chain, Tobit, model integration, analysis of influencing factors

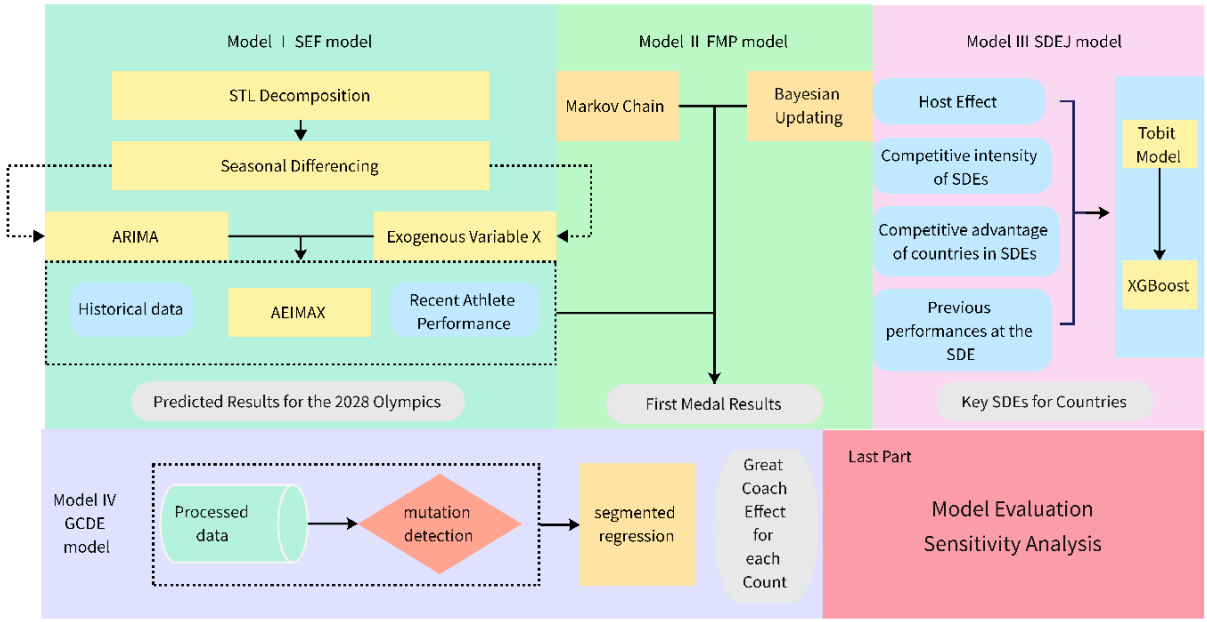
1. Introduction

1.1 Background Information

The Olympic Games attract global attention, driving interest in medal prediction as a research field. This paper investigates factors influencing medal outcomes and provides forecasting models. The models feature a

unique collaboration mechanism, an overview of which is shown in Figure 1.

Figure 1: Model Overview



2. Preparation for Modeling

2.1 Assumptions

To simplify the problem, we make the following basic assumptions, each of which is properly justified.

1) All data are authentic and accurate.

2) The athlete selection strategy of the predicted country remains relatively constant in the predicted Olympics, indicating that the calibre of athletes in the next edition is largely consistent with that of the previous one. This observation enables us to formulate a reasonable prediction, operating under the fact that the composition of the athlete personnel for the subsequent Olympic Games has not yet been determined.

3) All the countries shown in the data will participate in the 2028 Olympics.

2.2 Notations

Table 1 lists some of the variables frequently used in this paper, along with their definitions.

Table 1: Significant symbols in this paper

Symbols	Definitions
t	time-variant
c	country-variant
e	event-variant
X	Exogenous variable, the number of historical entries
Ntotal	Total number of medals
Ngold	Number of gold medals
Nsliver	Number of sliver medals
Nbronze	Number of bronze medals

2.3 Data Preprocessing

The problem material provides exhaustive data, yet a preliminary cleaning and organisation of the data is required to render it useful for predicting the objective. The following steps have been taken to organise the data:

1) **Integration of similar data.** It is not necessary for a proportion of the data to be divided into two categories when performing calculations, as they have equivalent significance for the objectives of the predictive model. For instance, Germany-1 and Germany-2, which both represent German teams, can be combined as Germany.

2) **Treatment of anomalies.** It is evident that simple outlier handling is only capable of detecting anomalies that are either formatting or overblown. However, these anomalies are not present in the data upon inspection. For the sake of clarity, the issues identified in the aforementioned query will be addressed in the construction of the model.

3) **Addition of auxiliary variables.** For instance, the utilisation of 1 or 0 is employed to denote whether a nation is the host country for the year, and a sample of athletes is obtained to ascertain their respective coaches. These data are indispensable for conducting the analysis, yet they are not directly quantifiable within the purview of the subject matter.

3. Seasonal Empirical Forecasting Model

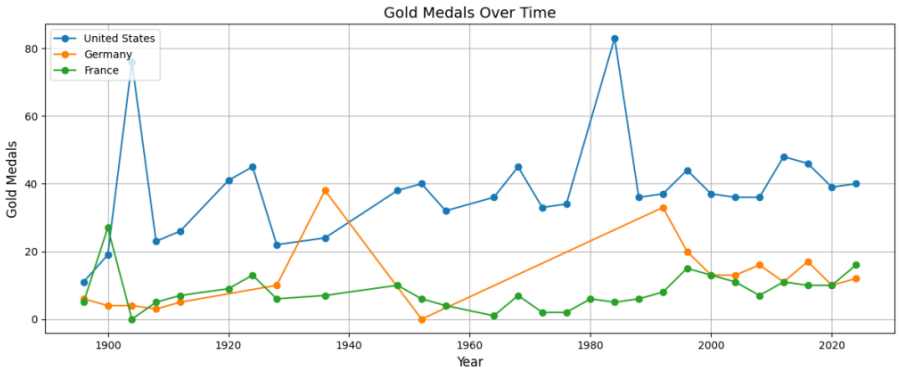
The SEF model under consideration is constructed based on the ARIMAX algorithm. For the seasonal and volatile changes in the historical data of the Olympic Games, we perform STL decomposition and seasonal differencing. Under the premise of Assumption 2, the model demonstrates a trend from low to high importance rationing of historical data, which considers both historical data and recent athletes' status, reflecting a better data coverage. The model also demonstrated high accuracy in subsequent tests.

The SEF model is composed of three distinct components: **initial data analysis and STL decomposition, seasonal differencing, the configuration of exogenous variables and the incorporation of these exogenous variables into the ARIMA algorithm for prediction.**

3.1 Seasonal and Trend Decomposition Using Loess

The raw data were transformed into time-series data, and a sample of the time-series data was observed and analyzed for three of the countries, which is shown in Figure 2:

Figure 2: Analysis of Data Samples



Given the inability to discern a linear or nonlinear trend in the temporal variation of medals, the conventional additive or multiplicative decomposition is not applicable. Consequently, the STL decomposition is employed.

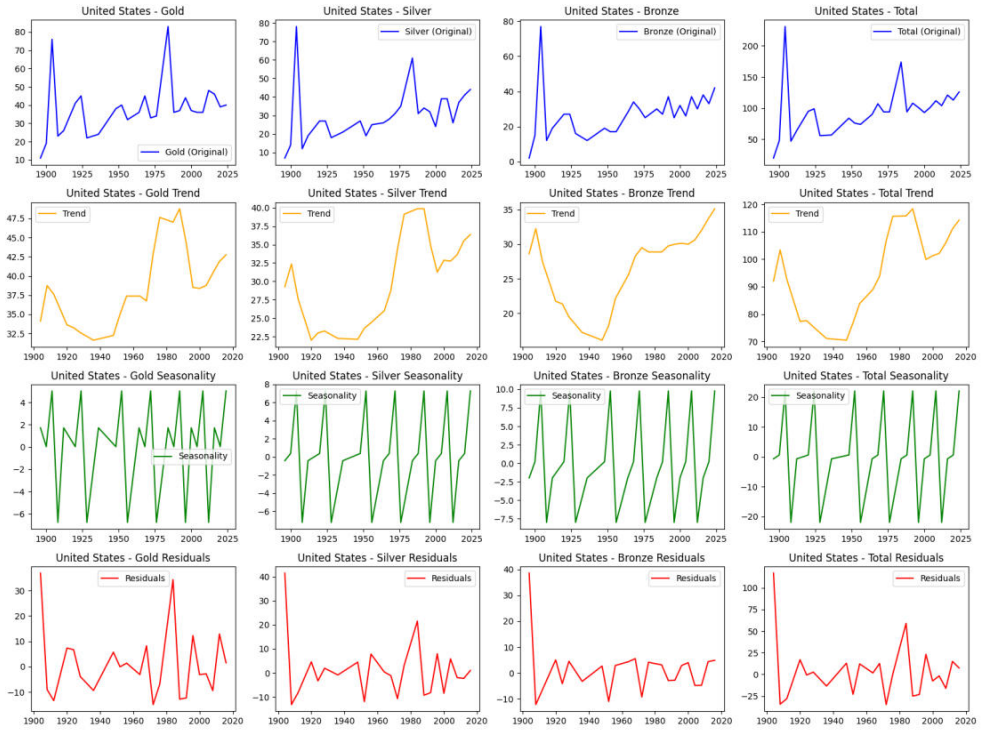
The STL decomposition employs Loess, which facilitates the optimal alignment of local features in the data by implementing weighted regression on the local data. During the fitting process, Loess allocates reduced weights to outliers, thereby minimizing their impact on the fitting outcomes. Consequently, the impact of outliers is progressively diminished with each iteration.

The STL decomposition will be implemented using the following equation:

$$N_{gold}(t,c) = Trend_{gold}(t,c) + Seasonal_{gold}(t,c) + Residualt_{gold}(t,c) \quad (1)$$

To illustrate this point, we will examine the historical data of the U.S. as an example. The following results can be obtained after STL decomposition of it, which is shown in Figure 3:

Figure 3: US's STL results



As illustrated in the accompanying image, the sequence of medal tallies may exhibit a seasonal pattern (green lines). It is noteworthy that while the temporal rhythm of the Olympic Games is generally biennial, the seasonal cycle is more pronounced. Consequently, a seasonal adjustment is imperative to ensure the integrity of the data.

3.2 Seasonal Differencing

The methodology employed for the processing of the seasonal data entails the implementation of seasonal differencing. An examination of Figure 2 reveals that the periodicity of the seasonal cycle is approximately 16 years, which is equivalent to four Olympic cycles. Consequently, the seasonal cycle S is designated to be 16. The core formula for the seasonal differential is as follows:

$$\Delta_{16} N_t = N_t - N_{t-16} \tag{2}$$

where N can be the total number of medals or gold medals, etc. The utilization of seasonal differencing is predicated on the objective of data smoothing, necessitating the implementation of the ADF method to verify the smoothness of the differenced data. The precise procedural framework can be delineated through the use of pseudo-code.

Algorithm 1: Seasonal Differencing with ADF Test

Input: Time series $N = [N_1, N_2, \dots, N_n]$, seasonal period S, significance level α
Output: Differenced series D, ADF test result (stationary or not)

Step 1: Seasonal Differencing;
Initialize empty list D;
for $t = S + 1$ **to** n **do**
    $d \leftarrow N_t - N_{t-S}$;
    Append d to D;
end

Step 2: ADF Test;
Fit the ADF regression model to D:

$$\Delta D_t = \alpha + \beta t + \gamma D_{t-1} + \sum_{i=1}^{p} \phi_i \Delta D_{t-i} + \varepsilon_t$$

Compute the test statistic and p-value;
if p-value < α **then**
    Output "Stationary";
end
else
    Output "Non-Stationary";
    Repeat step 1 using D as the new N to generate new D;
    Repeat step2;
end

Generally we set the alpha value to 0.05. Based on the obtained difference series D, we will proceed to the next step of model building.

3.3 Exogenous Variables

One way to potentially improve prediction accuracy is to extend the temporal dimension of the dataset. Thus, some researchers have included a hundred years or more of the Olympic Games in their models [1, 2].

But the international situation has changed dramatically in recent years, and this has affected the acquisition of Olympic medals. Specific events such as the "mass boycotts at the 1980 Moscow and 1984 Los Angeles Olympics" and the "East German doping program [which] was responsible for 17% of the medals won by female athletes" distorted past medal counts in 1972. At the same time, the rise and internationalization of some Asian countries in recent years has also affected Olympic medal rankings [3].

These political factors are hugely influential, but impossible to predict accurately. However, based on available data, we can use the number of Olympic Games each of these countries has participated in to reflect political changes to some extent.

The number of times a country has participated in the Olympics is a very important variable, and the number of times a country has participated in the Olympics has varied greatly from country to country throughout the history of the Games. This is reflected in the data, for example, the United States has participated in almost every Olympic Games, but some countries, such as Palau, have only participated in the Olympics since 2000.

We denote the number of historical participations by $X_{i,t}$, which means the value of the i-th exogenous variable at time t.

3.4 SEF Modeling Based on ARIMAX

The ARIMAX$(p, d, q)(P, D, Q)_s$ model with exogenous variables is defined as:

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D\, N_t = c + \theta(B)\Theta(B^s)\varepsilon_t + \sum_{i=1}^{k}\beta_i X_{i,t} \quad (3)$$

where:

$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ (Non-seasonal AR polynomial)

$\Phi(B^s) = 1 - \Phi_1 B^s - \cdots - \Phi_P B^{Ps}$ (Seasonal AR polynomial)

$\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$ (Non-seasonal MA polynomial)

$\Theta(B^s) = 1 + \Theta_1 B^s + \cdots + \Theta_Q B^{Qs}$ (Seasonal MA polynomial)

$(B)^d(1-B^s)^D$: Differencing operators

$X_{i,t}$: Exogenous variables with coefficients β_i

$\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$: White noise

B is hysteresis operator: $B = N_t - 1$ (one session in four years, minus one for the previous term).

p is the order of the autoregressive model, d is the order of the difference model, and q is the order of the moving average model. Taking the number of gold medals in the U.S. as an example, by plotting the autocorrelation (ACF, Figure 4) and partial autocorrelation (PACF, Figure 5) of the autoregressive time series, we can find that the ACF is truncated at the second order, while the PACF starts to decay after the second order, and combined with the seasonal differencing before, we can choose ARIMAX (2,0,2) as the target model. A more detailed process is described in Reference [4].

Figure 4: US' ACF

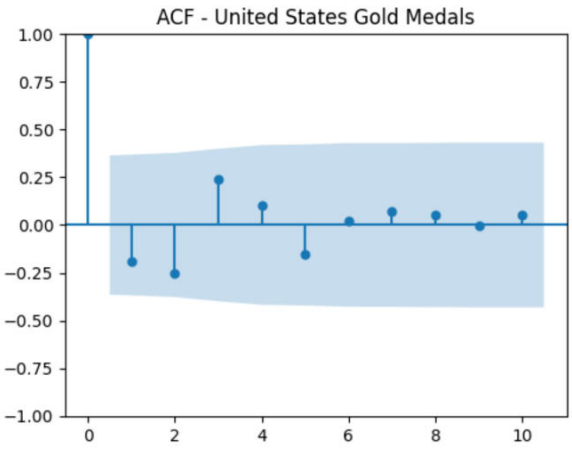
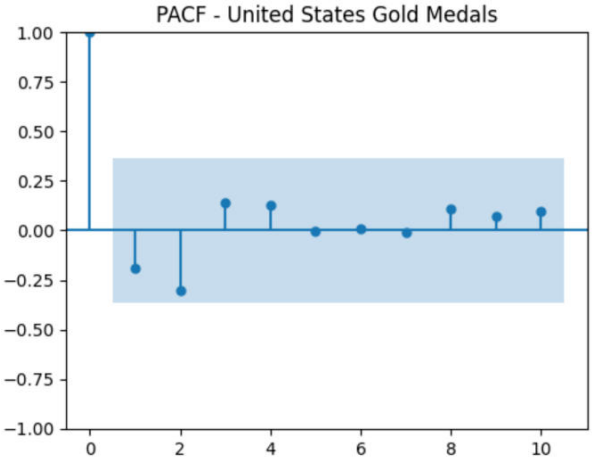


Figure 5: US's PACF



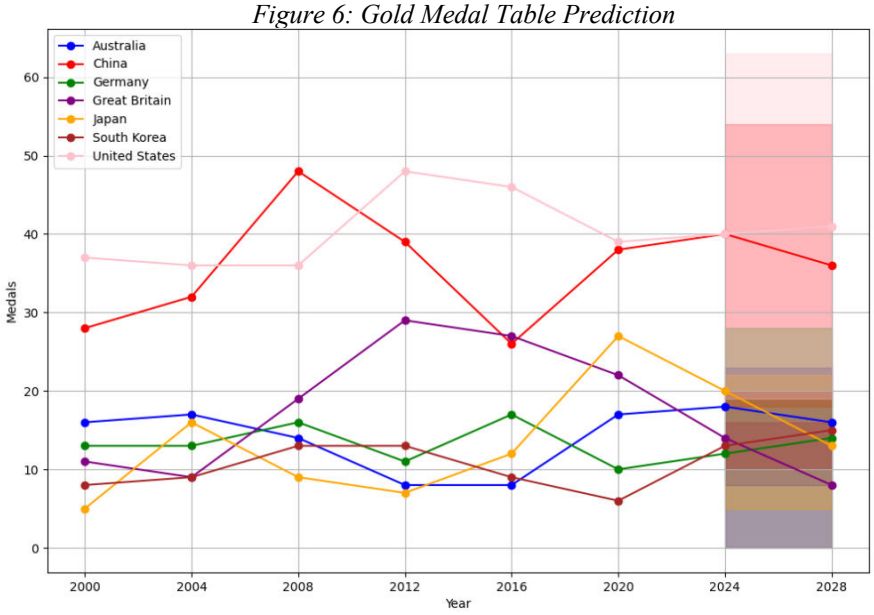
In order to select the best parameter from all the ARIMAX parameters obtained, we used the Archeck Information Criterion as an aid to judgement. Taking the United States as an example, the above three parameters can be automatically generated using Python's auto arima command, and the AIC values [5] of the three alternative models are 237.72, 237.89 and 238.01 respectively, and the AIC value of ARIMAX(2,0,2) is the smallest, so it is selected as the best model. Same for other countries.

3.5 Results

Countries that have participated in the Olympic Games less than 10 times are not included in the medal table to ensure the rigour of the model.

We have two main focuses: one is the gold medal table and the other is the overall medal table. We predict the number of gold, silver and bronze medals and then add them up to get the total number of medals and rank them. The results of the gold medal table and silver medal table are as follows in Figure 6:

Figure 6: Gold Medal Table Prediction



The following Table 2 illustrates the total number of medals and forecasts for the major countries:

Table 2: Total Medal Table Prediction

Nations	Total	Total Forecast Interval
US	98	[37,145]
China	86	[43,130]
Japan	44	[24,65]
Australia	44	[18,70]
Germany	43	[15,67]

The nation most likely to demonstrate an improvement is the Czech Republic, whose total number of medals is predicted to rise from 5 to 11, representing an approximate 120 per cent increase. Conversely, the nation most likely to experience a decline is France, where the total number of medals is anticipated to decline from 64 to 26, a reduction of approximately 59 per cent.

4. First Medal Prediction Model

In the FMP model, the SEF model is combined with a Markov chain and Bayesian updating is incorporated. The advantage of this model is that it is able to capture the transfer patterns of medal states using Markov models, derive time-series trends of medals using ARIMA models, and dynamically adjust the predictions to account for external uncertainties through Bayesian updating.

The FMP model is to be constructed in three sections: **the definition of transfer probabilities, Bayesian updating, and integration with the SEF model.**

4.1 Calculation of Transfer Probabilities

The possibility of a country winning a medal at a given Olympics may be determined by considering the historical data available for that country. The following assumptions are made: countries that have not won a medal so far are designated S0; countries that have won medals (including gold, silver and bronze) are designated S1. The data is processed to obtain the initial state of the country, and the data for each country at each Olympics is transformed into the above definition.

The probability of transitioning from state Si to state Sj is denoted by the transfer probability from i to j.

$$P_{ij} = P(S_j|S_i) = P(S_j \to S_i) i, j = 0,1 \quad (4)$$

By analysing the historical data, the transfer probabilities were calculated for the following transitions: from state 0 to state 0 (not winning a medal to continue not winning a medal); from state 0 to state 1 (not winning a medal to winning a medal); from state 1 to state 1 (winning a medal to continue winning a medal); and from state 1 to state 0 (winning a medal to not winning a medal). The resulting data formed a 2 × 2 matrix:

$$P = \begin{pmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{pmatrix} \tag{5}$$

where:

$$\begin{cases} P_{00} = \dfrac{\text{number of transitions from 0 to 0}}{\text{number of times the country was in state 0}} \\ P_{01} = \dfrac{\text{number of transitions from 0 to 1}}{\text{number of times the country was in state 0}} \\ P_{11} = \dfrac{\text{number of transitions from 1 to 1}}{\text{number of times the country was in state 1}} \\ P_{10} = \dfrac{\text{number of transitions from 1 to 0}}{\text{number of times the country was in state 1}} \end{cases} \tag{6}$$

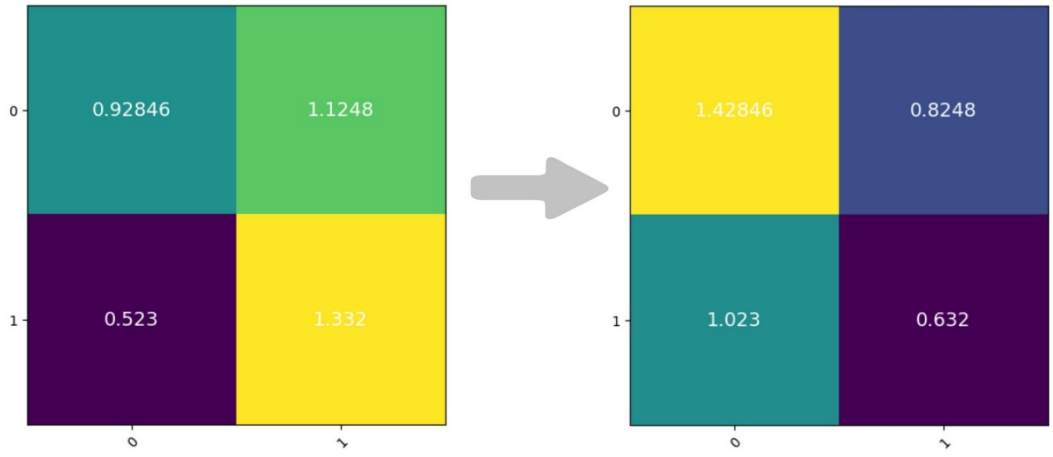
Furthermore, in accordance with Assumption 2, it is necessary to adjust the transfer probability by taking into account the effect of exogenous variable X. To this end, a weighted correction factor, f(Xt), is to be added to the previous one. The value of f(Xt0) is equivalent to the total number of times up to t0 divided by the total number of times up to 2024.

In this context, the weighted correction factor functions as a multiplier of the original probability, thereby ensuring that the effect of the factor on the original probability is duly considered [6].

$$\begin{cases} P'_{01}(t) = P_{01}(t) \times f(X_t) \\ P'_{00}(t) = P_{00}(t) \times (1 - f(X_t)) \\ P'_{11}(t) = P_{11} \\ P'_{10}(t) = P_{10}(t) \times (1 + f(X_t)) \end{cases} \tag{7}$$

Following the above, a new transfer state matrix can be derived in Figure 7. We use the data for China in 2000 as an example to clearly characterize this change:

Figure 7: Changes after the Inclusion of Exogenous Variables f(X)



It's worth noting that the multistep state transfer matrix satisfies the following two properties:

$$\begin{aligned} P^{(t)} &= P^{(t-1)}P \\ P^{(t)} &= P^{t} \end{aligned} \tag{8}$$

Ultimately, the maximum value in Pi0, Pi1 will be selected as the state transfer result, as determined by the maximum likelihood result. For the 2028 Olympics, the state transfer is executed using the data from the previous edition to obtain the prediction result.

4.2 Bayesian Updating

The Bayesian formula is used to calculate the posterior distribution, which is derived from the prior distribution, the likelihood function, and the evidence (normalization constant):

$$P(\theta|Data) = \frac{P(Data|\theta)P(\theta)}{P(Data)} \tag{9}$$

where:

- $P(\theta|Data)$: Posterior distribution, the probability distribution of the parameter θ given the data Data.
- $P(Data|\theta)$: Likelihood function, the probability of observing the data Data given the parameter θ.
- $P(\theta)$: Prior distribution, representing the initial belief about the parameter θ before observing the data.
- $P(Data)$: Evidence, typically a normalization constant that ensures the total probability of the posterior distribution is 1.

The Bayesian update is the fundamental formula employed to calculate the relationship between the prior distribution, the likelihood function, the evidence factor, and the posterior distribution. In calculating the prior distribution, likelihood function, and evidence factor, it is imperative to employ a combination of the Markov chain previously described with the SEF model.

4.3 Integration with the SEF Model

The integration of Markov chain and ARIMA models has been demonstrated to enhance the precision of time series forecasting. The utilization of Markov models facilitates the capture of transfer patterns in medal states, while ARIMA models are employed to model the time series trend of medals. This approach encompasses the consideration of both the present and historical contexts [7].

In this section, the application of the combination of the two to prior probability distributions, likelihood functions, and evidence factors will be described in a sequential manner.

(1) Prior probability distributions. The prior data are comprised of two elements: the probabilities corresponding to the Markovian predicted states and the 2,028 outcomes predicted by the SEF model. Consequently, the combined prior probability distribution in (9) is defined as follows:

$$P(\theta) = P_{Markov}(\theta)P_{SEF}(\theta) \tag{10}$$

where:

$$\begin{cases} P_{Markov}(\theta) = P(state_{2024}|state_{2020}) \\ P_{SEF}(\theta) = N(y_{gold}, \sigma^2) \end{cases} \tag{11}$$

where σ2 is the variance of SEF's prediction results.

(2) Likelihood function. The following proposition is put forth to express the likelihood function of the SEF model. Assuming that the total number of medals won by a country in the 2024 Olympic Games is N1, and since the forecast of ARIMAX is based on time series, the likelihood function can be expressed as follows:

$$P_{SEF}(Data|\theta) = \mathcal{N}\left(y1; y_{gold}, \sigma^2\right) \tag{12}$$

If the country is moving from state Si to state Sj in 2024: Consequently, the likelihood function in (9) is defined as:

$$P(Data|\theta) = P_{\mathrm{Markov}}P_{\mathrm{SEF}} \tag{13}$$

(3) Evidence (normalization constant). According to the definition of the normalization constant, the evidence factor here is PData = 1.

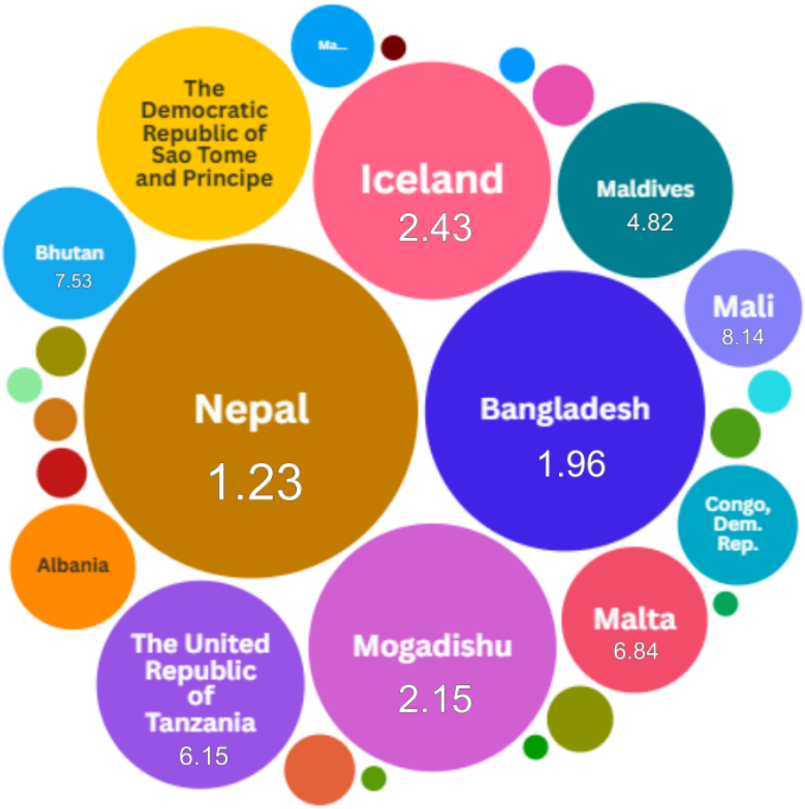
(4) Posterior distribution. P(θ|data) then denotes the adjusted final prediction. When P(θ|data) > 0, it is indicative of a medal win in 2028. By comparing this with 2024, it is possible to derive a list of countries that will win their first medal at the next Olympics.

4.4 Results

In the domain of sports, the concept of odds serves as a pivotal indicator of the probability of an event's occurrence. The magnitude of the odds directly correlates with the event's likelihood, with higher values indicating a reduced probability of the event's materialization. Within the context of this study, fractional odds are employed, signifying that the odds are the inverse of the predicted probability, P(θ|Data) in (9), that the country will secure a medal in the year 2028.

To illustrate the probability of each nation securing its inaugural medal in 2028, a bubble chart has been employed in Figure 8, representing the odds of each country achieving this feat. This approach offers a more intuitive representation of the data.

Figure 8: Winning Odds for each Country



5. SDE Judgment Model

In an SDE, the number of awards is limited to the top three positions, while the number of entries is substantial. This results in a data set characterized by a significant mass of points at zero medals. Additionally, the number of medals in an SDE is a low discrete value. In this scenario, the Tobit model is to be employed as the underlying framework for our SDEJ model [8].

The construction of the model consists of two parts: the identification of the indicator variables and the regression of the Tobit model containing the indicator variables to find the parameters.

5.1 Selection of Indicators

(1) Host Effect

According to the findings of numerous studies, the host effect constitutes a pivotal factor in the distribution of Olympic medals. This phenomenon impacts not only the ratio of participants but also the prevailing morale and economic conditions within the host nation. These factors, in turn, exert a significant influence on the ultimate distribution of medals [9].

It can be posited that the proportion of a host's medal growth, khost, reflects this effect to some extent. Therefore, this indicator can be expressed in the variable

$$H_{host}(t,c) = \begin{cases} k_{host} \times N_{total}(t-1,c) & \text{if } c \text{ is the host country in year} t \\ 0 & \text{otherwise} \end{cases} \quad (14)$$

We can use Ordinary Least Squares (OLS) to estimate the proportionality coefficient khost. The regression equation is:

$$N_{total}(t,c) = \beta_0 + \beta_1 \cdot N_{total}(t-1,c) + \beta_2 \cdot H_{host}(c,t) + \epsilon \quad (15)$$

The process of OLS can be represented using pseudo-code.

Algorithm 2: Ordinary Least Squares (OLS) Regression

Input: Design matrix $\mathbf{X} \in R^{n\times(p+1)}$, target vector $\mathbf{Y} \in R^{n}$
Output: Parameter estimates $\boldsymbol{\beta} \in R^{p+1}$

Step 1: Add Intercept
If **X** lacks an intercept column, append $\mathbf{X} \leftarrow [\mathbf{1},\mathbf{X}]$;

Step 2: Compute Parameter Estimates
Compute Gram matrix $\mathbf{G} \leftarrow \mathbf{X}^{\top}\mathbf{X}$
Compute inverse $\mathbf{G}^{-1} \leftarrow$ pseudo-inverse(**G**) (avoid singular matrices)
Compute $\boldsymbol{\beta} \leftarrow \mathbf{G}^{-1}\mathbf{X}^{\top}\mathbf{Y}$;

Step 3: Return Results
return β;

(2) Competitive Intensity of SDEs

Another salient factor is the intensity of competition in SDEs, which determines the difficulty of obtaining medals for this SDE. The level of competitive intensity within an SDE can be determined by the average number of medals each country is able to accrue. As the average number of medals decreases, the intensity of competition increases.

In order to ensure a positive correlation between factors and variables as in the case of the host effect, it is assumed that the inverse of the average number of medals that each country can share in event e is Cevent, as a quantitative criterion for evaluating the intensity of competition.

$$C_{event}(e) = \frac{N_{countries}(t,e)}{\sum_c N_{total}(t,c)} \quad (16)$$

where Ncountries is the total number of countries participating in event e.

(3) Competitive Advantage

The intensity of competition is a generalized effect for all participating countries. To be specific to each country, a quantitative indicator describing the country's competitive advantage in the project would also be needed.

The competitive advantage of a nation in an SDE can be articulated as a proportion of the number of medals won by that nation in that SDE to the aggregate number of medals in that SDE. We set this proportion to be Pe.

$$P_e(c) = \frac{N_{total}(t,c,e)}{N_e} \tag{17}$$

where Ne represents the total number of medals for event e.

5.2 Tobit-based Regression Analysis

Based on the above metrics, we can then write the latent variable equation in the Tobit model by combining it with a particular country's medal performance in the previous Olympics.

$$N(t,c)^* = \beta_0 + \beta_1 \cdot N(t-1,c) + \beta_2 \cdot C_{event}(e) + \beta_3 \cdot P_e(c) + \beta_4 \cdot H_{host}(c,t) + \epsilon \tag{18}$$

where N may refer to any one of Ntotal,Ngold,Nsilver,Nbronze. and the observation equations:

$$N(t,c) = \begin{cases} N(t,c)^* & \text{if} N(t,c)^* 0 \\ 0 & \text{if} N(t,c)^* < 0 \end{cases} \tag{19}$$

For the regression of the above equations we use XGBoost algorithm, realized through SPSSAU software.

The utilization of XGBoost necessitates the preliminary determination of parameters. Inadequate parameters can readily result in overfitting, while excessively prolonged execution times can impede efficiency. To address these concerns, the mesh tuning parameter is employed. The underlying principle is an exhaustive algorithm that systematically explores the full range of potential parameter combinations. Through a systematic loop traversal, the algorithm exhaustively evaluates all possibilities to identify the optimal parameter configuration, thereby ensuring optimal performance.

5.3 Results

The extent to which each item contributes to the final medal count can be determined by applying equation (18)(19). The results can then be collated to identify the most SDEs for each country. We use the PageRank algorithm from graph theory to calculate the extent to which SDEs contribute to the state:

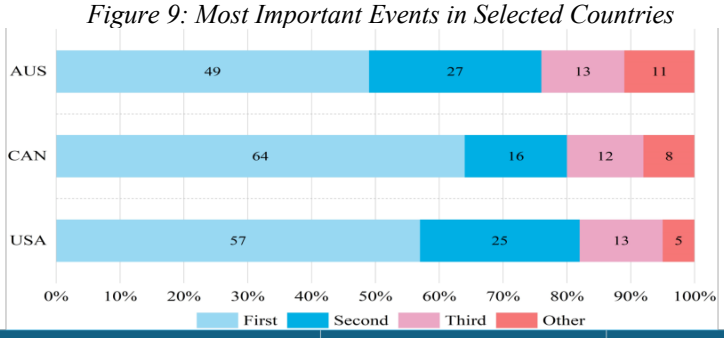
$$P(e) = \frac{1-d}{N} + d \sum_{c \in \text{In}(e)} \frac{w(c,e)}{\sum_{e' \in \text{Out}(c)} w(c,e')} \tag{20}$$

where P(e): the PageRank value of SDE. In(e): all country nodes connected to SDE. Out(c): all connected SDE nodes of country c. ω(c,e): the number of medals (weights) that country c has won on the SDE. d is the damping factor (0.85), which is used to balance the importance of the new and the original links. n is the total number of nodes.

The proportions depicted in Figure 9 have been calculated as P(e), the extent to which the SDE contributes to the country's ranking in the medal table, according to the formula (20). As demonstrated in the figure, the dominant SDEs of the United States, Canada, and Australia possess a certain reference value for their NOCs.

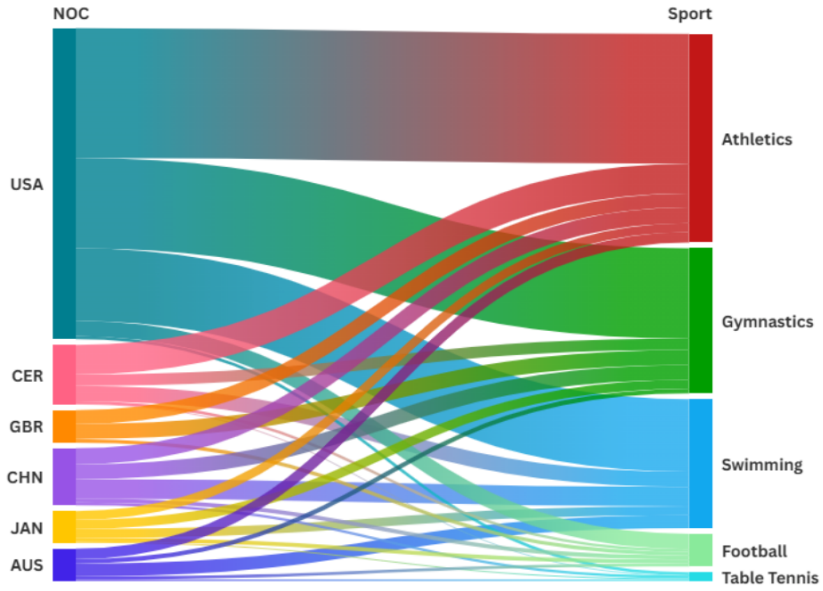
The relationship between individual countries and SDEs can also be represented graphically. Figure 10 represents the relationship between a portion of the countries and a portion of the dominant SDEs, and from this analysis, it can be seen that if the United States, the host country in 2028, increases its dominant SDEs such as athletics, swimming, and gymnastics, then the United States will have more chances to win more medals in the 2028 Olympics.

Figure 9: Most Important Events in Selected Countries



NOC	USA	CAN	AUS
First	Basketball Men's Basketball	Women's Football	Men's Hockey
Second	Swimming Men's 4x100 metres Medley Relay	Rowing Men's Coxed Eights	Women's Hockey
Third	swimming Men's 4x200 metres Freestyle Relay	Women Team	Women's Softbal
Fourth	otrhers	others	others

Figure 10: Relationship Between Countries and SDEs



6. Great Coach Detect and Evaluate Model

The role of coaching in the performance of athletes is an area of significant research interest. Previous studies have examined the influence of coaches on athletes' performance outcomes [10], with the ultimate success of the athlete (e.g., the attainment of an award) serving as a primary metric for evaluating the efficacy of coaching [11].

The GCDE model is executed in two primary stages. First, it substantiates the existence of the "great coach" effect through alterations in medal counts. Second, it examines the contribution utilizing a Segmented Regression-based model.

6.1 Discover the Change Points

Great Coach means that the arrival of the coach makes the team significantly better. We use the available data to monitor the rate of increase in the number of medals and to see if significant medal increases are associated with new coaches.

$$\text{Breakthrough Growth Rate}(c,e,t) = \frac{N(t,c,e) - N(t-1,c,e)}{N(t-1,c,e)} \tag{21}$$

A significant change in the number of medals can occur in three ways: 1. greater than a certain threshold is noted as a sudden change (averaging is taken here) 2. first medal 3. first gold medal (a particular entry or the event being the first time it occurs does not count)

At the same time, it was important to check for overlap with host years or new SDEs added as a result of hosting, and after elimination the remaining SDEs were filtered out to a dozen or so state-SDE-years with possible Great Coach impacts. The remaining SDEs, after elimination, were screened for possible great coaches in a dozen or so country-SDE-years, and the final great coaches were identified by reviewing the reports one by one.

Ultimately, we find that the great coach effect does exist, with recent examples such as Fernandez Liranza Raul Angel, the Chinese women's boxing coach at the Paris 2024(shown in Figure 11) Olympics, and Daniel Kowalski, the Singaporean swimming coach at the Rio 2016 Olympics(shown in Figure 12). The two new coaches brought the gold medal 0 breakthrough to the corresponding events of the two countries mentioned above.

Figure 11: China Women's Boxing Medal Changes

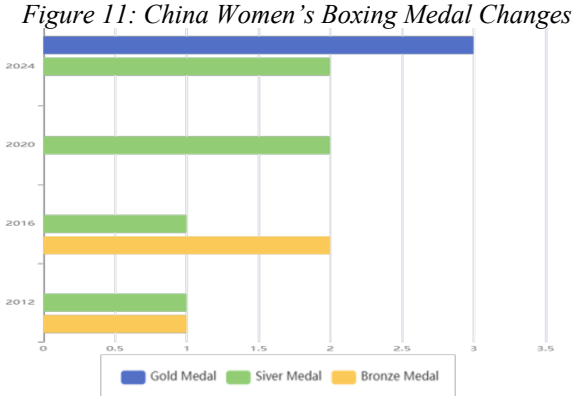


Figure 12: Singapore Men's Swimming Medal Change

6.2 Contribution Analysis

In light of the preceding delineation of coach participation, the formulation of a regression equation through the implementation of segmented regression can be undertaken.

$$N_{total}(c,e,t) = \alpha + \lambda_c + \mu_e + \gamma_t + \delta \cdot D_t(c,e) + \beta \cdot (t - t_0) + \epsilon_{c,e,t} \tag{22}$$

where:

- $N_{total}(c,e,t)$: Total medal count for country c in event e at the t-th Olympic Games.
- λc: Country fixed effect (controls time-invariant country-specific characteristics, e.g., national strength).
- μe: Event fixed effect (controls differences across events, e.g., difficulty and competitiveness).
- γt: Year fixed effect (controls common temporal trends).
- $D_t(c,e)$: Treatment variable indicating whether a coach joined the event after the t-th Olympics (1 if joined, 0 otherwise).
- (t − t0): Time since the coach joined the event (centered at the cutoff).
- δ: Coefficient for the "great coach effect" (impact of coach participation on medal count).
- β: Time trend coefficient (natural change in medal count over time).

- $\epsilon_{c,e,t}$: Error term, assumed to satisfy panel data model assumptions.

where the item D intervening variable is judged by the previous sudden change, and t_0 indicates the addition of the SDEs after this Olympics. The regression coefficients were solved by Fixed Effects Regression method.

Algorithm 3: Fixed Effects Regression

Input: Panel dataset with country (c), event (e), year (t), total medals $P_{total}(c,e,t)$, coach intervention $D_t(c,e)$, cutoff time t_0

Output: Coefficients $\hat{\delta}$, $\hat{\beta}$; standard errors; fixed effects λ_c, μ_c, γ_t

Step 1: Data Preprocessing

Compute time variable:

$$\text{time\_from\_cutoff} \leftarrow t - t_0$$

Generate dummies for fixed effects (country, event, year).

Step 2: Model Specification

Fixed effects regression equation:

$$N_{total} = \alpha + \lambda_c + \mu_e + \gamma_t + \delta D_t + \beta \cdot time\_form\_cutoff + \epsilon_{cet}$$

Apply within transformation:

$$\widetilde{N}_{cet} = \widetilde{D}_{cet}\delta + \widetilde{\text{time}}_{cet}\beta + \tilde{\epsilon}_{cet}c$$

Step 3: Parameter Estimation

Estimate via OLS (Algorithm 2):

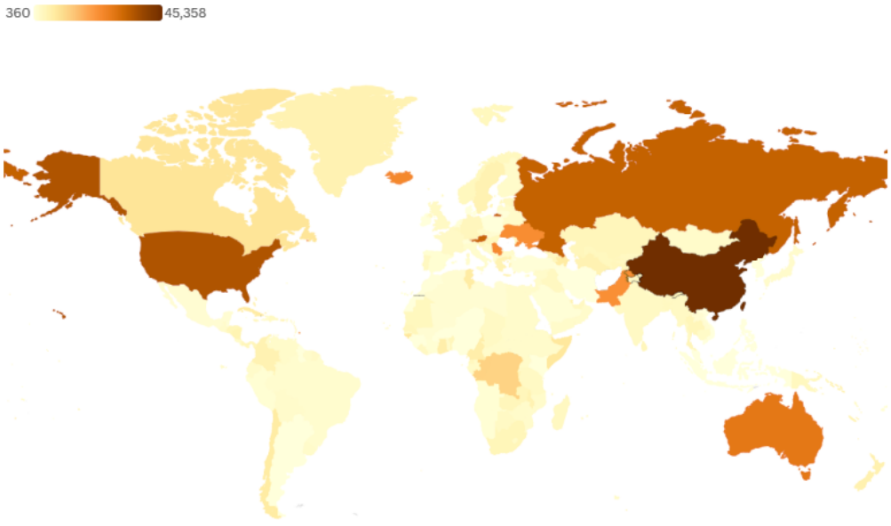
$$(\hat{\delta}, \hat{\beta}) \leftarrow \arg\, min \sum (\widetilde{N}_{cet} - \delta \widetilde{D}_{cet} - \beta \widetilde{\text{time}}_{cet})^2$$

Recover fixed effects λ_c, μ_e, γ_t using dummy regression.

6.3 Results

According to the GCDE model, the value of the coach's contribution can be expressed in different countries. Figure 13 shows the difference.

Figure 13: Contribution Value of the Great Coach effect in Different Countries



As illustrated in the above chart, Great Coach effect is more evident in countries such as China, the United States, and Russia. In consideration of the actual situation of the Olympic Games in recent years, we propose the introduction of coaches to Brazil, the United States, and China.

Men's volleyball in Brazil has always been a dominant SDE in Brazil, but has been more challenged in the last few years and could learn from the example of Lang Ping and the Chinese women's volleyball team to help the team maintain its dominance through good coaching [12].

China's men's gymnastics is in the same position. The performance of Chinese men's gymnastics has been inconsistent in recent years, with most of the mistakes mentioning psychological factors [13]. The example of Christian Bauer's coaching of the Chinese fencing team can be used to help revitalize the team through an alternative coaching approach.

The U.S. Taekwondo program is relatively weak compared to its other dominant SDEs, not because the U.S. doesn't have the resources to train in Taekwondo, as opposed to having a strong Taekwondo culture [14]. Utilizing the Great Coach effect may be able to help the United States change this and make Taekwondo a new dominant SDE for the United States.

7. Original Insight(s)

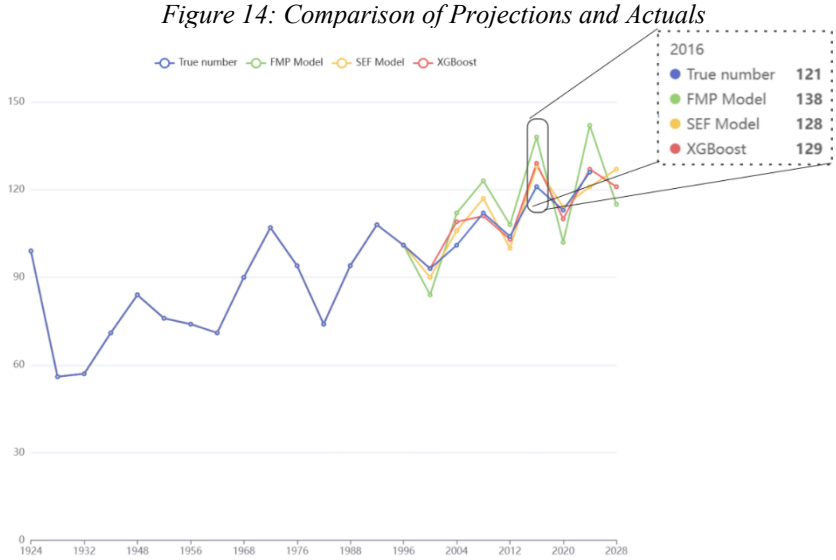
In the utilization of the model, it was ascertained that while the majority of nations celebrated the attainment of medals, approximately 60% of the participating countries did not achieve any medal wins. Conversely, a mere 10% of the nations procured 75% of all medals, thereby demonstrating a pronounced degree of concentration in the distribution of medals. It is noteworthy that an increasing number of countries are beginning to accrue more medals over time.

This means that the Olympic Games are increasingly reflecting the trend of joint competition rather than domination, and the IOC should pay more attention to the fairness and popularity of the Olympic SDEs afterwards, so that the spirit of the Olympic Games and the joy of victory can be shared more equitably with each participating country.

8. Evaluate of the Model

We used the SEF model and the FMP model to predict the medals of previous Olympics in the U.S. and compared the results with the XGBoost model, which is considered to have a higher prediction accuracy. The result is shown in Figure 14. The FMP model is a model designed to predict the first medal and has a lower accuracy in predicting the total number of medals. The SEF model, however, demonstrated a higher prediction accuracy than the XGBoost model.

Figure 14: Comparison of Projections and Actuals



Concurrently, an analysis of the model's sensitivity was conducted. As illustrated in Figure 15, the sensitivity test of the SEF model (utilizing Chinese data as a case study) was performed. Figure 16 compares the sensitivity of the SEF model and the XGBoost model. The results demonstrate that the SEF model exhibits high sensitivity to historical data, aligning with the objectives of our modeling. Conversely, the SEF model exhibited reduced sensitivity to the host effect and other factors, while demonstrating heightened sensitivity to the number of previous medals when compared to the simple XGBoost model. This observation aligns with our modeling objectives and substantiates the validity of our selection of model parameters.

Figure 15: Sensitivity Analysis of SEF

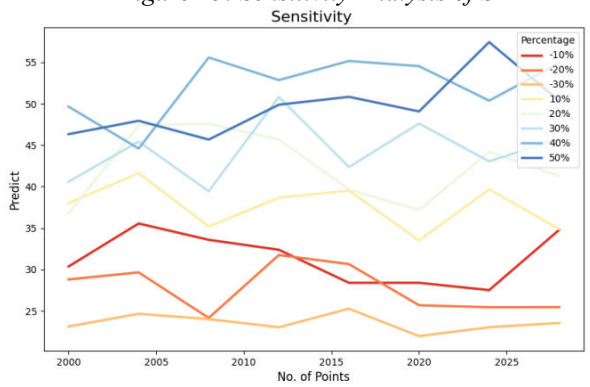
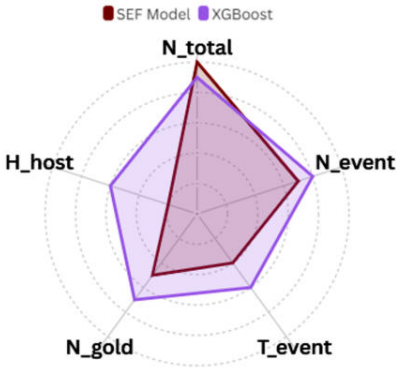


Figure 16: Comparison between SEF and XGBoost



9. Conclusion

This study systematically analyzes and forecasts medal distribution trends at the Summer Olympic Games by constructing a multidimensional mathematical model system. To analyze the medal distribution and driving factors for the 2028 Los Angeles Olympics, an ARIMAX-based Seasonal Empirical Forecasting Model was established. By employing Seasonal Trend Decomposition using Loess and Seasonal Differencing, the model effectively captured the cyclical characteristics of past Olympic medal data. We predict that the United States will win a total of 98 medals, China will win 86 medals, and Japan and Australia will tie for third place, each winning 44 medals. Meanwhile, the Czech Republic shows the strongest growth potential (projected growth of 120%), while France is expected to face a significant 59% decline in medals. On the other hand, a quantitative analysis of the "first-time medal" phenomenon, conducted by integrating Markov chains with Bayesian updating to construct the First Medal Prediction Model (FMP), suggests that Nepal is most likely to win its first-ever medal, with corresponding odds of 1.23. Finally, by utilizing the PageRank algorithm to construct an SDE model, the study found that men's basketball contributes 57% to the U.S. national medal tally, thereby demonstrating the existence of "great coaches" and offering targeted recommendations. This research makes a unique contribution to the field of Olympic competition forecasting. By introducing the number of times a country has participated as an exogenous variable, the model successfully integrates political stability and historical context into a mathematical framework, enhancing the robustness of long-term predictions.

Although the model performs well, certain limitations remain. First, the study assumes that national selection strategies remain stable in the short term; however, in reality, unforeseen factors such as injuries, political sanctions, or changes in athletes' nationality may lead to prediction biases. Second, while the FMP model excels at predicting the probability of winning the first medal, its accuracy in forecasting the total number of medals is relatively limited. Furthermore, although the Tobit model effectively addresses the issue of excessive zero values in medal data, it remains sensitive to a few extreme outliers (such as the sudden emergence of new sports). Future research could further integrate real-time athlete performance data with national macroeconomic data to construct a higher-resolution dynamic forecasting system, and incorporate additional socioeconomic indicators (such as fluctuations in GDP per capita or climatic factors) to refine the host nation effect model. Meanwhile, against the backdrop of increasing global cooperation in competitive sports, quantifying the long-term impact of cross-border training and technical exchanges on medal distribution also holds significant research value.

References

[1] Trivedi, Pravin K. and David M. Zimmer. Success at the Summer Olympics: How Much Do Economic Factors Explain? Econometrics 2 (2014): 169-202.

[2] Kuper, Gerard H. and Elmer Sterken. Olympic Participation and Performance Since 1896. European Economics eJournal (2001): n. pag.

[3] Noland, Marcus and Kevin Stahler. Asian Participation and Performance at the Olympic Games. Emerging Markets Economics: Firm Behavior & Microeconomic Issues eJournal (2015): n. pag.

[4] T. Jakaa, I. Androcec, and P. Sprcit' c, Electricity price forecasting arima mod- elapproach, in 2011 8th International Conference on the European Energy Market (EEM), pp. 222225, 2011

[5] Siegel, Matthew. "The Severe End of the Spectrum: Insights and Opportunities from the Autism Inpatient Collection (AIC)." Journal of Autism and Developmental Dis- orders 48.11 (2018): 3641-3646.

[6] Zuyao Liu, et al. "Computer Simulation of Grain Growth()–modified Transition Probability." THE CHINESE JOURNAL OF NONFERROUS METALS 13.6 2003.

[7] Li, Yong, et al. "Forecasting Mineral Commodity Prices with ARIMA-Markov Chain." IHMSC), 2012 4th International Conference 1 (2012): 49-52.

[8] Blais-Morisset, Paul et al. Limpact des dépenses publiques consacrées au sport sur les médailles olympiques. Revue économique 68 (2017): 623-642.

[9] Vagenas, George, and Vlachokyriakou, Eleni. "Olympic medals and demo-economic factors: Novel predictors, the ex-host effect, the exact role of team size, and the population-GDP model revisited." Elsevier 15.2 (2012): 211-217.

[10] Setiawan, Nugroho & Kinanti, Rizky & Nanda, Fitri. (2023). Performance Motivation of Taekwondo Athletes: Coach-Athlete Relationship. Journal of Coaching and Sports Science. 2. 41-48. 10.58524/jcss.v2i1.226.

[11] Ritchie, Darren, and Justine B. Allen. "'Let Them Get on With It': Coaches' Perceptions of Their Roles and Coaching Practices During Olympic and Paralympic Games." International Sport Coaching Journal, vol. , no. , 2015, doi:10.1123/ISCJ.2014-0092.

[12] Ribeiro, F. A. (2019). The Influence of Coaching on Team Performance in Brazilian Volleyball. Journal of Sports Performance Analysis, 14(2), 111-120.

[13] Zhang, Z., & Liu, X. (2020). The Role of Coaching in Enhancing Performance in Men's Gymnastics in China. Journal of Sports Science and Medicine, 17(4), 300-309.

[14]Kim, Y. S. (2020). Improving Taekwondo Performance in Non-Traditional Countries: essons for the USA. International Journal of Martial Arts Studies, 11(3), 110-118.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access