Problem Chosen	2025	Team Control Number
C	MCM/ICM Summary Sheet	2523614

The Road to Gold: Predicting and Analyzing Olympic Medal Performance

Olympic medals are not only the proof of athletes conquering their own limits, but also a reflection of a nation's willpower, the prediction of the Olympic medals has also become a subject of great interest.

Firstly, the data set was pre-processed, including unified data format coding, missing value filling, and so on. Among them, when predicting the 2028 Olympic medal table, the 49 sports currently announced for the 2028 Los Angeles Olympic Games were used as the object of the purpose study, and the relevant data that did not contain the 2028 sports were excluded.

For problem 1, the Olympic medal prediction model is constructed. Firstly, the countries that have won medals construct the Olympic medal impact indicator set and quantify the host country effect, and substitute the indicator set based on the **gray time series model** prediction into the **SVR** to predict the number of Olympic medals of each country in 2028. Then considering the host country effect. Next, the **Bootstrap method** is used to obtain the prediction interval for the number of medals of each country under the 95% confidence interval, and the top 5 countries are obtained as follows: **USA, CHN, GER, JPN, KOR**, of which the prediction interval for the total number of medals of USA is $[\mathbf{117, 133}]$; Secondly, the award evaluation system is constructed, and **a coupled SVM and Logistic model** is utilized to explore the award-winning situation of non-winning countries in 2028 and the corresponding probability, in which **BAN, IVB, SAM, and VIN** will be awarded medals, with the corresponding probability of **0.679, 0.656, 0.705, and 0.701**;Tirdly, USA, CHA, and GER, which are the top 3 in the predicted medal list, are selected to explore, based on the time series clustering model, the relationship between events/sports and the relationship between the number of medals won by countries, in which USA, CHA, and GER's most dominant sports are, in order, "Basketball Men's Basketball", "Volleyball Women's Volleyball", "Hockey Men's Hockey", and Swimming is the most important for these three countries. Finally, based on the **gray correlation** analysis of the relevant data of USA, FRN and GER, we explore the impact of the setting of the host country events on the total number of medals, and give the top 20 influential events.

For problem 2, the "Great coach" effect was analyzed**.** First**,** the data provided were analyzed by **EDA** to verify the evidence of the Great coach effect on the number of medals. Then, based on **the DID model**, we constructed a data set of control variables to analyze and estimate the impact of the effect by comparing the differences between the experimental and control groups before and after the implementation of the "Great coach" effect. Finally, USA, CHA and GER were selected for the study and recommendations for investing in "Great coach" were given. The contribution of the Great Coach effect to USA sport is 6.5 medals on average and we choose to invest in basketball, which has a medal growth rate of **23.21%**.

For problem 3, based on the model established above, from the perspective of each continent, we explore the trend of the number of medals and athletes in each continent over time, and count the distribution of the total number of medals and gold medals in each continent in 2028, and the results are all **Europe>Asia>North America>Oceania>Africa>South America.**

Finally, this paper carries out sensitivity analysis on all models, and the results show that the performance assessment of the established models are all great, and they can predict the number of Olympic medals well.

Keywords: gray time series model; SVR; Bootstrap; SVM and Logistic; EDA; DID;

Catalogs

1. Introduction

1.1 Problem Background

During the 2024 Summer Olympics in Paris, the overall "medal table" of each country will attract a lot of attention. Each medal is not only a proof of the athletes' conquest, but also a manifestation of the willpower of a nation or a country, which condenses the Olympic spirit and the characteristics of each nation's flavor. At the same time, the prediction of Olympic medals has attracted scholars from all walks of life to study. Among them, the 2028 Olympic Games in Los Angeles, what countries will achieve has also become the focus of general attention.

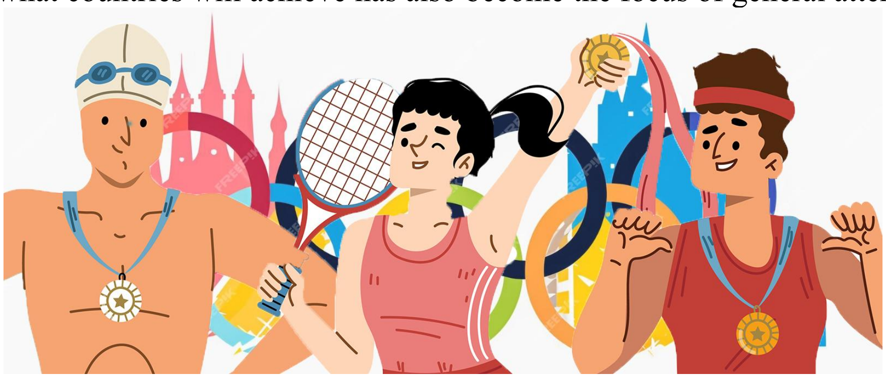

Figure 1 Background diagram

1.2 Problem Restatement

Considering the background information and restricted conditions identified in the problem statements, we need to solve the following problems.

- **Problem1:** Based on the data provided and the requirements, build a corresponding mathematical model to first predict and analyze the medal standings for the 2028 Summer Olympics in Los Angeles, USA, followed by predicting the first-time winning country and its probability, and finally explore the relationship between events/sports and the number of medals won by a country and analyze how the host country chooses the events to affect the number of medals.
- **Problem2:** Examine the data, seek evidence of changes that may be caused by the "great coach" effect, and estimate the impact of this effect on the number of medals won, and finally select three countries to validate the model and give their sports that take into account the "great coaches". Recommendations.
- **Problem3:** Explore in depth the characteristics associated with the number of Olympic medals and explain how these insight(s) can inform country Olympic committees.

1.3 Our Work

In order to better analyze and predict the distribution of Olympic medals based on the data provided, this paper is based on a country-by-country study, and the specific work contains the following elements:

- **Pre-processing of the provided datasets:** data cleaning, data visualization, etc. For example, the "NOC" in each data table corresponds to the name of specific countries, exploring the distribution of basic data characteristics of countries that have won awards and those that have never won awards, and validating the effect of the host country. In particular, when predicting the 2028 Olympic medal list, the 49 sports announced for the 2028 Los Angeles Olympic Games are used as the target research object, and the data related to the sports that do not include the 2028 sports are excluded.

- **Constructing Olympic medal prediction model:** Since the number of Olympic medals is a fluctuating, complex and unpredictable nonlinear time series, the SVR Olympic medal prediction model is constructed. First of all, the countries that have won medals construct the Olympic medal impact indicator set and quantify the host country effect, and substitute the indicator set obtained based on the **gray time series model** prediction into the SVR to predict the number of Olympic medals of each country in 2028. Considering the host country effect, **Bootstrap** method is used to obtain the corresponding medal count prediction interval at 95% confidence interval. Then, an award evaluation system is constructed, and a coupled SVM and logistic model is used to explore the award-winning situation and the corresponding probability for non-winning countries in 2028. Finally, the relationship between event/sports and the number of medals won by countries is investigated using time series clustering and gray correlation model.
- **Analysis of the "Great coach" effect:** The data provided were analyzed by EDA to validate the evidence of the "great coach" effect on the number of medals. Then, based on the DID model, a data set of control variables was constructed to analyze and estimate the impact of the effect by comparing the differences between the experimental and control groups before and after the implementation of the "Great coach" effect. Finally, the top 3 countries in the 2028 forecast list are selected and recommendations for investing in "great coaching" are given.
- **Unique insights into Olympic medal counts:** In order to explore the relevant characteristics of Olympic medal counts in greater depth, based on the model established above, the relationship between medals and continents is explored from the perspective of each continent.

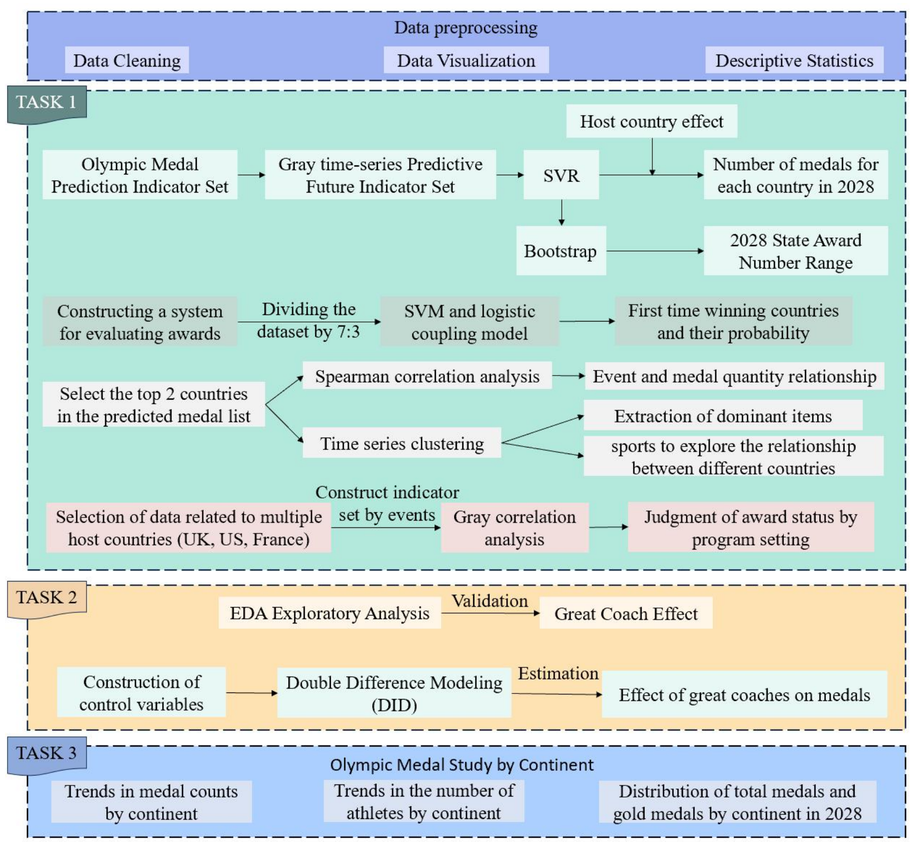

Figure 2 Our work

2. Assumptions and Justifications

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

- **Assumption 1:** It is assumed that the great coaching effect is a time-dependent effect, and that the coaching effect is estimated differently for different years of data.

Justification: In DID, time is needed to make calculations on great coaches with the help of time, so the time factor needs to be controlled.

- **Assumption 2:** Assumption that the great coaching effect only occurs in traditional programs and economically developed countries.

Justification: Simplify the model while taking into account the reality that you have to be an economically powerful country to hire great coaches.

- **Assumption 3:** It is assumed that collecting data with NOC as the label reflects the relationship between the data well.

Justification: To simplify the model and reduce the difficulty of collecting indicator sets.

3. Notations

For convenience, we introduce some important notations below.

Symbol	Description
Me_t	Predicted number of medals
AE	host country effect
$logit(P)$	Probability of winning a prize for the first time
$d_{CORT}(X_T, Y_T)$	similarity measure
$\varepsilon_i(k)$	gray correlation

4. Data Pre-processing and Visualization Analysis

4.1 Data Pre-processing

Firstly, data in the data set with partial encoding errors (suspected use of GBK encoding) were uniformly adjusted to utf-8. Secondly, Russia was not able to compete in 2028, and therefore Russia was not studied.

(1) Visualization of Discipline data in "summerOly_programs"

Combining the relationship between Discipline and Event, we use python software to populate the vacant data in "summerOly_programs" and visualize the number of Disciplines per session in the populated "summerOly_programs".

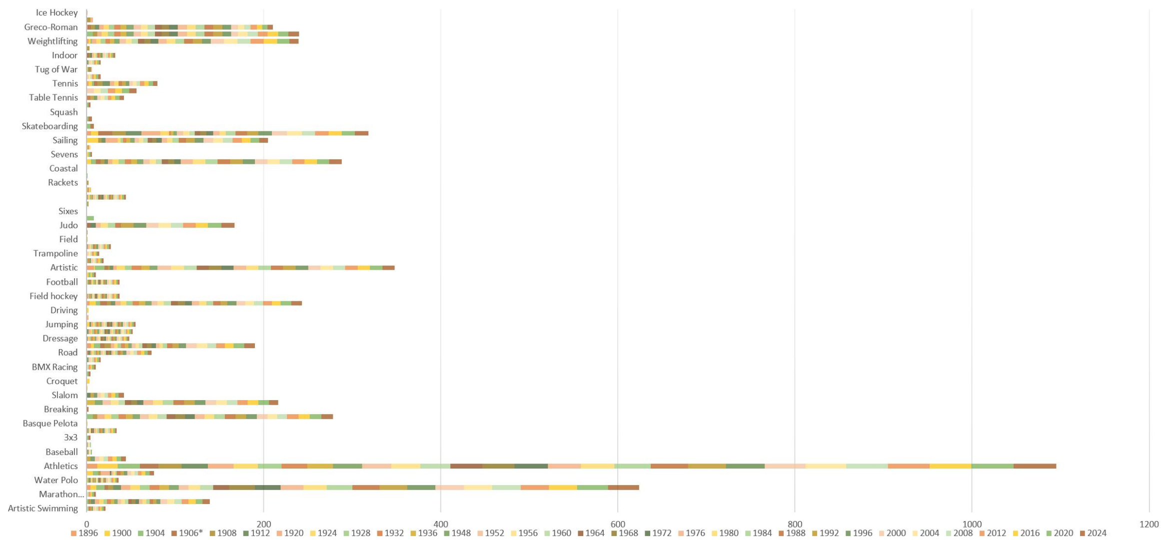

Figure 3 Histogram of the stacking of various types of small projects

(2) NOC corresponds to country code

Since "Team" contains more details than country, "summerOly_athletes" is processed with "NOC" as the country classification. Therefore, "summerOly_athletes" is categorized by "NOC" as the country. Considering the different forms of NOC in different periods, with the help of python software, based on the list of IOC National Codes, the "NOC" in each data table corresponds to the name of a specific country, for example, "AFC" corresponds to "Afghanistan", "CHN" corresponds to "China", "GER" corresponds to "Germany", "USA" corresponds to "America" and so on.

(3) Extraction of data from "summerOly_athletes" for countries that have not yet won a medal

The python software was used to identify countries and their teams in the "summerOly_athletes" data set that have never won a medal, resulting in 75 different country name codes (NOC) [1]. Given the page limit of the text, only the first 30 are shown in this paper:

Table 1 National NOC codes that have never won a medal

NOC	Number of Olympic Games	NOC	Number of Olympic Games	NOC	Number of Olympic Games
AND	13	BOL	16	CRT	1
ANG	11	BRU	7	ESA	14
ANT	12	CHA	14	FSM	7
ARU	10	CAF	12	GAM	11
ASA	10	CAM	11	GBS	8
BAN	11	CAY	12	GEQ	11
BEN	13	CGO	14	GUI	13
BHU	11	COD	12	GUM	10
BIH	9	COK	10	HON	13
BTZ	14	COM	8	IVB	11

(4) Information on athletes from previous Summer Olympics host countries

Using python software, combine "summerOly_hosts" with "summerOly_athletes" in the "summerOly_athletes" to identify the information of athletes of the host country of the previous Summer Olympic Games, to construct a new data set. "filtered_athletes_data_unique" for subsequent construction of the medal prediction model. The distribution of the number of national athletes participating in the host country of the previous Summer Olympics is shown below:

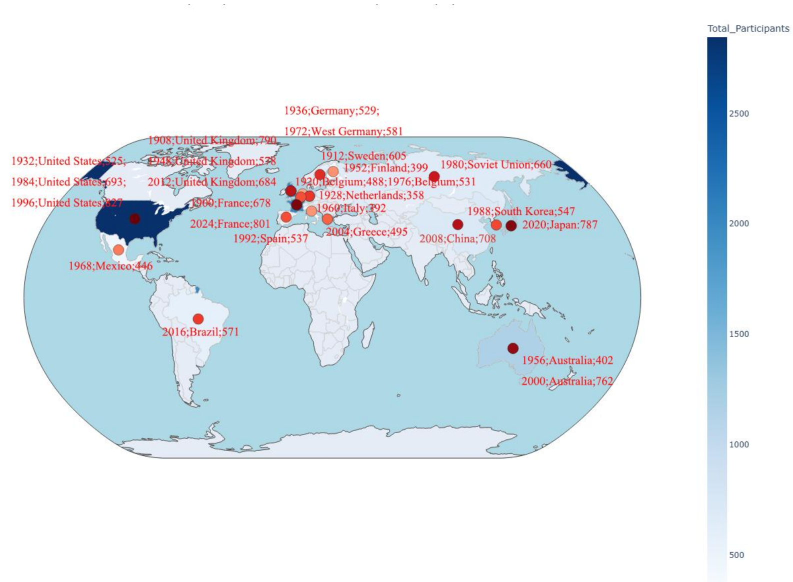

Figure 4 Distribution of the number of participants from the host

4.2 Visualization Analysis

(1) Visual analysis of medal totals by country

Using python software to visualize the total medal counts of countries in the "summerOly_medal_counts" data set from 1896 to 2024, we can get the top 5 countries are "United states, Soviet Union, Great Britain, France, China", and the bottom 5 countries are "Taiwan, Togo, Tonga, Turkmenistan, Virgin islands".

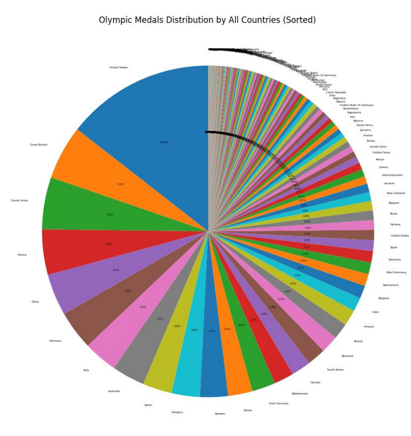

Figure 5 Pie Chart of Total Medal Shares by Country

(2) host country effect

With the home field advantage that both the competition and preparation are in the home country, the host country effect will bring higher probability of medal winning for the home country [2]. In order to verify the impact of the host country effect, this paper uses python software to mark the host country of the Olympic Games in each year, taking the United States and Australia as an example, to carry out the host country effect verification analysis.

Combined with the visualization analysis, the United States' total medal count increased significantly when it hosted the Olympics in 1904, 1932, 1984, and 1966, and Australia's medal count increased in 1956 and 2000 when it hosted the Olympics, while its total medal count ranked in the top 5. It can be seen that the effect of hosting the Olympics can lead to a higher medal winning rate for the country.

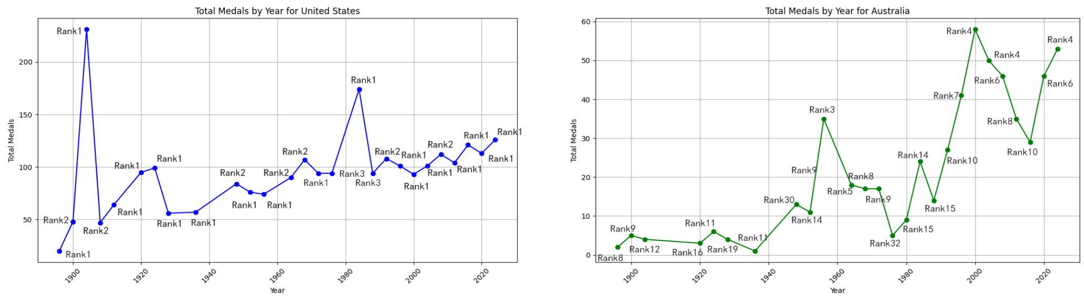

Figure 6 Medal counts and corresponding rankings of the United States (left) and Australia (right) over the years

5. Olympic Medal Prediction Model

Considering the influence of the host country, athletes and the number of Olympic events on the medal table, this paper takes the number of medals as the dependent variable (divided into gold, silver and bronze for consideration), and the set of indicators of Olympic medal influencing factors as the independent variable, to construct the Olympic medal prediction model. The specific process is as follows:

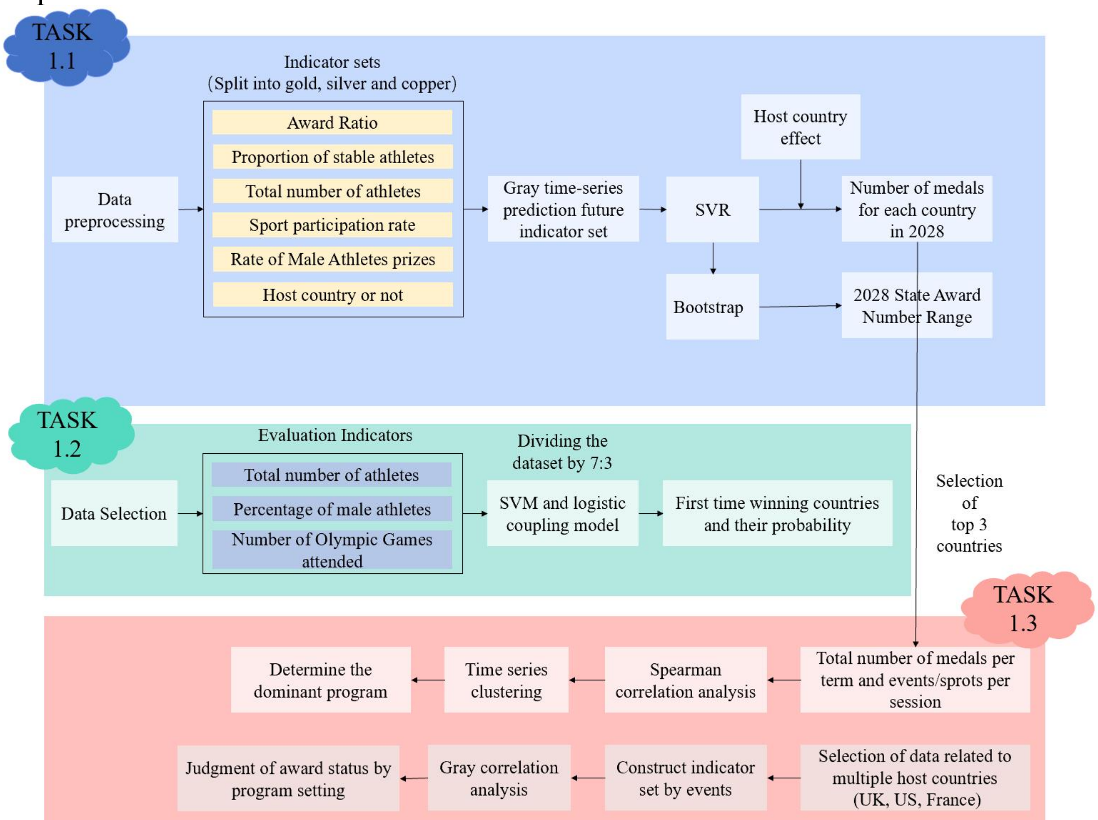

Figure 7 Flowchart of the specific body in this section

Due to the subdivision of "Event" and "Discipline" in the data set, in order to simplify the model, this paper mainly explores the impact of sports on Olympic medals by sport. In order to predict the number of medals for each country in 2028, this paper takes the 49 sports announced for the 2028 Los Angeles Olympic Games as the target research object, and excludes the relevant data of sports that do not include the 2028 sports.

5.1 Construction of the Olympic Medal Impact Indicator Set

Based on the list of announced events in 2028, the three major cleaned datasets were analyzed using python software to construct the Olympic medal impact indicator set $Con = \{c_1, c_2, ..., c_6\}$. The set includes the following six factors: the ratio of the number of corresponding medals won by the country in the first n Olympic Games to the total number of medals won in the current Olympics c_1 , the proportion of stable athletes (in this paper, it refers to the athletes who have won medals for two consecutive years or more) c_2 , the number of total athletes participating c_3 , the ratio of participating sports to the total sports c_4 , the ratio of male athletes who won medals c_5 , and whether or not the country is the host country in the current Olympic Games c_6 (indicated by 0-1 variable, where 0 is no and 1 is yes).

Taking the United States as a representative, the factors obtained from the indicator set are analyzed for correlation by using Spearman correlation coefficient with the help of python. It can be obtained that c_3 and c_6 have strong positive correlation with the number of medals: c_2 has weak correlation with the number of medals.

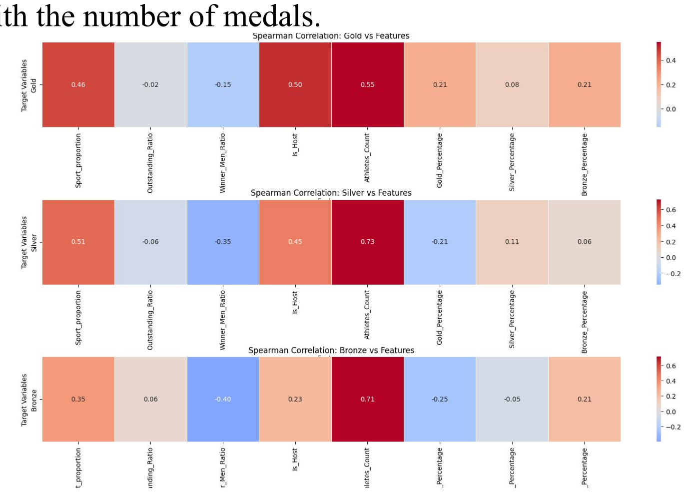

Figure 8 Correlation analysis of each factor with the corresponding number of medals

5.2 Host Country Effect

➢ **Quantification of host country effects**

From this paper 4.2, it can be concluded that with the home advantage of having both the competition and preparation in the home country, the host country effect will play out in a way that will result in a higher probability of winning a medal for the home country. Using the official assignment of points as per the Olympics, 13, 11, and 10 points are assigned to gold, silver, and bronze medals, respectively, resulting in a total score [3]. Due to the fact that the data spans a large number of years, during which individual countries merged or changed over, this paper uses the data under the GDR and FRG to uniformly attribute to Germany.

First of all, we calculate the proportion of medals won by the host countries in the past, and the results are shown in the following table. Among them, X_0 indicates the proportion of the number of gold medals to the total number of gold medals in the current session, and Y_0 indicates the proportion of the combined medal scores to the total medal scores in the current session.

Table 2 Percentage of gold and total medal scores of countries that have hosted the event

Year	Host country	$X_0(\%)$	$Y_0(\%)$
1896	Greece	0.1613	0.0784
1900	France	0.2587	0.0984
......	......	......	......
2020	Japan	0.0850	0.0319
2024	France	0.0713	0.0267

➢ **Prediction of host country effects**

Prediction of the host country effect, taking into account the time trend, after debugging, the paper uses a simple moving average with a step size of 15 (i.e., a time interval of$D = 15$, the first time span of$1896 - 1972$ years) for the 34th session of the data, the Olympic Games were suspended in 1916, 1940 and 1944. By calculating the 14th moving average prediction can be obtained, each time the average move under the gold medals, the combined score host country effect (AE) is calculated as follows, respectively:

$$AE_{gold,i} = \frac{\sum_{n=i}^{D+i-1} (X_0(n) - X_i(n))}{D} \tag{1}$$

$$AE_{overall\ score,i} = \frac{\sum_{n=i}^{D+i-1} (Y_0(n) - Y_i(n))}{D} \tag{2}$$

Among them, n indicates the session of the Summer Olympics; i indicates the i moving average of($i = 1,2......$) ;$X_0(n)$、$Y_0(n)$ indicates the proportion of gold medals (composite score) won by the host country of the *nth* session of the Summer Olympics to the total number of gold medals (total score) of the session;$X_i(n)$、$Y_i(n)$ indicates the average of the proportion of gold medals (composite score) won by the host country of the *nth* session of the Summer Olympics as a non-host country of the other sessions of the Summer Olympics to the total number of gold medals (total score) of the session, as indicated in the$i$ moving average. Similarly, the host effect of the number of silver and bronze medals is calculated using this method.

In this paper, the host country effect after the 14th moving average is used as the predicted value, which yields the values of the host country effect in 2028 as:$AE_{gold} = 9.73\%$,$AE_{silver} = 5.90\%$,$AE_{bronze} = 4.68\%$,$AE_{overall} = 6.77\%$.

5.3 Predicting Future Indicator Sets Based On Gray Time-series Models

Data selection:

In this model, countries that have never won a medal are not considered. In this case, when calculating the indicator set, countries are combined or not according to the existence of countries in 2024.

5.3.1 Gray Forecasting$GM(1,1)$ Model

Assuming that the time series$X^0(t)$ has one observation$X^0 = (x^0(1), x^0(2)), ..., x^0(n)$,a new sequence is generated by accumulating$X^1 = (x^1(1), x^1(2)), ..., x(n)$, where ,$x^{(1)}(t) = \sum_{k=1}^{t} x^{(0)}(t)$ ($t = 1,2, ...,\ n$)α is the developmental gray level and$u$ is the endogenous control gray number, the corresponding differential equation for the$GM(1,1)$ model [4] is:

$$\frac{dx^{(1)}}{dt} + \alpha x^{(1)} = v \tag{3}$$

Let$\alpha^{\wedge}$ be the parameter vector to be estimated, then$\alpha^{\wedge}\begin{bmatrix}\alpha\\ v\end{bmatrix}$, using the least squares method, solves for$\alpha^{\wedge}=(B^TB)^{-1}B^TY_n$, whereB $=\begin{pmatrix}-x(^1(1)+x^1(2))/2 & 1\\ \vdots & \vdots\\ -(x^1(n-1)+x^1(n))/2 & 1\end{pmatrix}$,$Y_n=\begin{bmatrix}x^0(2)\\ \vdots\\ x^0(n)\end{bmatrix}$, and further solving yields: $\widehat{X}^{(1)}(k)=(X^0(1)-\nu/\alpha)e^{-\alpha(t-1)}+\alpha/$ v .Finally, t=1,2,......,n is brought into the above equation and the predicted value is solved.

5.3.2 Improved Gray Prediction $GM(1,1)$ Model

In order to fit the model more accurately, the original $GM(1,1)$ model is improved, i.e., the background value $z^{(1)}(k)=(1/2)\cdot[x^{(1)}(k)+x^{(1)}(k-1)]$ is optimized and improved to:

$$z^{(1)}(k)=\frac{x^{(1)}(k)-x^{(1)}(k-1)}{lnx^{(1)}(k)-lnx^{(1)}(k-1)} \tag{4}$$

where $z^{(1)}(k)=z^{(1)}(k),k=2,3,\dots,n$ when$x^{(1)}(k)=x^{(1)}(k-1)$.

Also the initial conditions in the model were changed to,$\hat{x}^{(1)}(m)=x^{(1)}(m),m=1,2,\dots,n$,m values should be determined by minimizing the average relative error of the predicted values of the actual data. The time corresponding function of the improved $GM(1,1)$ gray model is obtained as:

$$\begin{gathered}\hat{x}^{(1)}(k)=(x^{(1)}(m)-\nu/\alpha)e^{-\alpha(k-m)}+\alpha/\nu\\ \hat{x}^{(0)}(k)=(1-e^{\alpha})(x^{(1)}(m)-\nu/\alpha)e^{-\alpha(k-m)}\\ x^{(1)}(m)=e^{-\alpha}x^{(1)}(m-1)+\frac{b}{a}(1-e^{-\alpha})\end{gathered} \tag{5}$$

where$x^{(0)}(k)=x^{(1)}(k)-x^{(1)}(k-1)$ and$\hat{x}^{(1)}(k)=\hat{x}^{(0)}(k)$

5.3.3 ARIMA-based Time Series Analysis Modeling

ARIMA (p,d,q) where$p$ denotes the order of the autoregressive model,d denotes the order of the difference, and$q$ denotes the order of the moving average model.

$$\begin{cases}\Phi(B)\nabla^d x_i=\theta(B)\varepsilon_t\\ E(\varepsilon_t)=0,Var(\varepsilon_t)=\sigma_t^2\\ E(\varepsilon_t\varepsilon_s)=0,s\neq t\\ E(x,\varepsilon_t)=0,\forall s<t\end{cases} \tag{6}$$

Where,x_i is the corresponding predictor indicator for year$i$,$\nabla^d=(1-B)^d$ is the difference calculation,$\{\varepsilon_t\}$ is the zero-mean white noise series,$\Phi(B)=1-\Phi_1B-\cdots-\Phi_pB$ is the moving smoothing coefficient polynomial of the smooth reversible ARMA(p,q) model,B is the delay operator, and$B^nx_i=x_{i-n}$. Using$max\xi$ to denote the maximum value of the likelihood function and *num* to denote the number of parameters, the AIC and ADF formulas are obtained:

AIC information guidelines set the order:

$$AIC=\frac{-2}{T}ln(max\xi)+\frac{2}{T}(num) \tag{7}$$

ADF test formula:

$$\Delta y_i=\gamma y_{t-1}+a+\delta t+\sum_{i=1}^{p}\beta_i\Delta y_{t-i}+u_t \tag{8}$$

where a is a constant term andδt is a linear time trend term.

5.3.4 Gray Time-series Combinatorial Modeling:

In this paper, ARIMA is combined with improved $GM(1,1)$ to construct a gray time series combination model [5]. Through the establishment of the combined prediction model, the prediction results obtained by various prediction methods are considered in a comprehensive manner, the index information obtained by multiple prediction methods is utilized to a greater extent, the consideration of the problem is more comprehensive and systematic, and the influence of individual prediction models by various random factors is reduced, so as to improve the accuracy of the model prediction.

Assuming that for a particular forecasting problem the vector of actual observations is $X = (x_1, x_2, x_3, \ldots\ldots, x_n)$,there are *m(m≥ 2)* forecasting methods, the elements of the vector $\omega = (\omega_1, \omega_2, \ldots\ldots, \omega_n)^T$ species are the weighting coefficients in the various combinations of forecasting models, respectively, and the predicted values of the j forecasting method are ,$\widehat{x_{1j}}\widehat{x_{2j}}$,......, ,$\widehat{x_{nj}}j = 1,2, \ldots, m$, and the predicted values ofx_t for the t year are as follows:

$$x_t = \sum_{j=1}^{m} \omega_j \widehat{x_{tj}} = \omega_1\widehat{x_{t1}} + \omega_2\widehat{x_{t2}} + \ldots\ldots + \omega_m\widehat{x_{tm}} \tag{9}$$

Where$t = 1,2, \ldots\ldots, n$, the actual observed value in year$t$ is x_t , and the predicted value of the method in$j$ in year$t$ is $\widehat{x_{tj}}$.

In this paper, we take the inverse of variance approach to solve for the magnitude of the power coefficients, which is expressed by the equation

$$\omega_j = \frac{e_j{}^{-1}}{\sum_{j=1}^{m} e_j{}^{-1}}, \sum_{j=1}^{m} \omega_j{}^{-1}, j = 1,2, \ldots\ldots, m \tag{10}$$

$e_j$is the sum of squared errors for the jth model, i.e.,$e_j = \sum_{t=1}^{n} \left(x_t - \widehat{x_{tj}}\right)^2$.

5.4 Vector Machine (SVR) Regression Based Medal Prediction Model

➢ **Modeling:**

Olympic winning medals are a kind of fluctuating, high noise, complex and unpredictable, nonlinear and uncertain time series, and the regression method in SVR model provides an effective solution for it. In this paper, SVR nonlinear extended samples are used to order the time series model by analyzing the change of the support vector set after new samples are added to the training set [6]. The basic principle is as follows: set the training set $S = \{(x_t, y_t) \in R^d \times R, t = 1,2, \ldots\ldots, n\}$,x_t is the$d$ dimensional input vector,y_t is the corresponding output vector,n is the total number of data, the parameter $\nu \in (0,1]$ is used to control the number of support vectors, the positive relaxation variable ζ_t, ζ_t^* is introduced, and the following planning problem is constructed:

$$minR(w, \zeta^{(*)}) = \frac{1}{2}||w||^2 + C(v^{\epsilon}) + \frac{1}{n}\sum_{i=1}^{n} (\zeta_t + \zeta_t^*)) \tag{11}$$

$$\begin{gathered} s \cdot t \cdot\ y_t - w\varphi(x_t) - b \leq \epsilon + \zeta_t \\ w\varphi(x_t) + b - y_t \leq \epsilon + \zeta_t^* \\ \zeta_t^{(*)} \geq 0, t = 1,2, \ldots\ldots, n, \epsilon \geq 0 \end{gathered} \tag{12}$$

where $\varphi(x)$ is the function in the higher dimensional space, w is the full vector, the coefficient b is the bias, and C is the error penalty parameter. Introducing the Lagrange multiplier methodλ_t, λ_t^* and the symmetric function$K(x, x_t) = exp(g||x_t - x_s||^2)$ which satisfies Mercer's condition transforms Eq. () to obtain the SVR model equation:

$$f(x, \lambda_t, \lambda_t^*) = \sum_{t=1}^{n} (\lambda_t - \lambda_t^*)K(x, x_t) + b \tag{13}$$

Combined with the set of influencing factors $Con = \{c_1, c_2, \ldots\ldots, c_6\}$,the multivariate nonlinear regression model of the number of medals (Me_t) was constructed:

$$Me_t = \lambda_0 + \lambda_1 c_1 + \lambda_2 c_2 + \ldots\ldots + \lambda_6 c_6 \tag{14}$$

➢ **Model solving:**

With the help of python operation, the number of medals of each country in 2028 was obtained. Then the Bootstrap method was used to resample multiple predictions, and the final prediction interval for each country's medal count was obtained at the 95% confidence interval. Where the prediction results for the United States are combined with the predicted 2028 host country effect values in 5.2, i.e.:

$$Me_{USA\_last} = Me_{USA}(1 + AE_i) \tag{15}$$

The median of the prediction intervals was selected to sort the results obtained to get the final 2028 predicted medal list. Due to page limitations, only the top 10 countries are shown in this article with their corresponding medal count prediction intervals:

Table 3 Forecast of the top 10 countries in the medal table in 2028 in terms of winning ranges

NOV	Gold	Silver	Bronze	Total
USA	47-52	35-43	35-38	117-133
CHN	44-49	29-32	26-29	99-110
GER	26-33	25-33	29-36	80-102
JPN	13-16	11~12	14~16	38-44
KOR	13-14	10~12	11~12	34-38
AUS	12~14	13-15	16-18	41-47
GBR	9~14	11~14	14~17	34-45
FRA	10~11	10~11	13~15	33-37
ITA	10~11	10~11	13~14	33-36
HUN	9~10	7~9	8~9	24-28

5.5 Predictive Model of Award Probability for Zero-award Countries Based on SVM And Logistic Coupling

In order to predict the probability of countries that have never won a medal to win in the next Summer Olympics, this paper builds a binary classification model [7] for countries that have never won a medal to predict and obtain the corresponding winning probability. Considering that the medal data of countries that have never won medals are all zero, this paper adopts a division by 7;3 through previous data, i.e., data before 1984 is used to train the model, and data from 1984-2024 is used to predict the probability of winning the next Olympic Games for countries with zero corresponding awards. With the help of excel screening and python calculations, the data of countries that have won medals and those that have never won medals are counted separately.

➢ **Constructing evaluation indicators:**

The dependent variable Y is assumed to be the country that has won a medal or not (0 means no, 1 means yes), and the set of impact indicators $X = \{x_1, x_2, x_3\}$ includes the number of total athletes who participated in SPORT in that country before 1984, the percentage of males among total participating athletes, and the number of times they have participated in the Olympic Games.

➢ **Prediction of whether to win the award based on SVM:**

The RBF kernel function $K(x, x_t) = exp(g||x_t - x_s||^2)$ is chosen, and the grid search method is used to find the penalty parameter C and the kernel function g to predict whether the country will win the award or not based on the set of evaluation indexes. Construct the following classification decision function:

$$f(x) = sign(\sum_{i=1}^{N} \alpha_i^* y_i K(x, x_i) + b^*) \tag{16}$$

where a component of $b^* = y_j - \sum_{i=1}^{N} \alpha_i^* y_i K(x_i, x_j)$, α^* α_j^* satisfies the condition $0 < \alpha_j^* < C$.

➢ **Logistic-based prediction of the probability of winning the award:**

A general linear model is built for the $0-1$ variable Y in conjunction with the evaluation metrics.

$$E(Y) = P(Y = 1) = P = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{17}$$

To simplify the model, the $logit$ transformation is applied to P :

$$logit(P) = ln(P/(1 - P)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{18}$$

$$P = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}} \tag{19}$$

Define $odds = P/(1 - P)$ as the ratio of the probability that the country will win a prize to the probability that it will not win a prize, then there is when x_i increases by one unit:

$$\frac{odds/x_{i+1}}{odds/x_i} = e^{\beta_i} \tag{20}$$

As a result, the predictive regression model of logistic regression on the value of P for non-awarded countries can be obtained. In this paper, we set a threshold value of 0.5, if $P > 0.5$, Y takes 1, otherwise it takes 0.

➢ **Maximum Likelihood Estimation:**

Let the data for the n th country be $(x_{i1}, x_{i2}, x_{i3}, y_i)(i = 1,2, \ldots\ldots, n)$, then the likelihood function for $y_1, y_2 \ldots, y_n$ is:

$$L = \prod_{i=1}^{n} p(x_i)^{y_i} [1 - p(x_i)]^{1-y_i} \tag{21}$$

Take the natural logarithm of both sides of the equation and then take the partial derivatives of the balls of $\beta_0, \ldots, \beta_3$ to maximize the likelihood function and obtain the corresponding estimates $\widehat{\beta_0}, \ldots, \widehat{\beta_3}$.

➢ **Model prediction results:**

with the help of python software, the logistic regression expression is obtained from the training set as:

$$logit(P) = -2.197 + 1.388x_1 - 0.379x_2 + 1.737x_3 \tag{22}$$

It can be seen that the effect of the percentage of male athletes on the prediction is negative, that is, the higher the percentage of male athletes, the lower the probability of winning the award for the first time; the number of athletes and the number of Olympic Games participated in the probability of winning the award for the first time are both positive. Substitute the data from the test set into SVM , to get the country that will win the prize for the first time in the next Olympic Games.

Table 4 Countries that will win a prize for the first time at the next Olympic Games and probability of doing so

Country	Bahamas	The British Virgin Islands	Samoa	Viet Nam

	(BAN)	(IVB)	(SAM)	(VIN)
P	0.679	0.656	0.705	0.701

5.6 Evaluation Model of Program-medal Relationship Based on Time Series Clustering

In order to simplify the model, this paper selects USA, CHA and GER, which are the top 3 countries in the 5.4 predicted medal list, and builds an evaluation model to explore the relationship between events/sport and the number of medals won by a country.

5.6.1 Constructing a Judgment Model for Advantageous Projects Based on Time Series Clustering

- **Time series clustering model construction:**

Because the data have obvious time-domain distribution characteristics, this paper converts the commonly used similarity calculation methods for static data into methods suitable for time series clustering, and then clusters the time series data with a set of clustering algorithms for static data [8].

- **Time series clustering similarity measure:**

In this paper, we choose the improved CORT method that combines the distance measure between time series and the correlation of the time series themselves [9].The similarity of the characteristics of the 2 time series' own attributes over time is measured by the first-order temporal correlation, and the first-order temporal correlation coefficients between the 2 intra-temporal sequencesX_T, Y_T are defined as follows:

$$CORT(X_T, Y_T) = \frac{\sum_{i=1}^{T-1}(X_{i+1} - X_i)(Y_{i+1} - Y_i)}{\sqrt{\sum_{t=1}^{T-1}(X_{t+1} - X_t)^2}\sqrt{\sum_{t=1}^{T-1}(Y_{t+1} - Y_t)^2}} \tag{23}$$

The value of$CORT(X_T, Y_T)$ is in the interval$[-1,1]$. The larger the value, the more similar the magnitude and direction of the rate of change of the two time series at each moment; the closer it is to -1, the rate of change of the two series is similar in magnitude but opposite in direction; and if it is 0, it indicates that the two series are linearly independent of each other. After combining the first-order temporal correlation coefficients of the 2 time series, the measure of this similarity is defined as follows:

$$d_{CORT}(X_T, Y_T) = \phi_k[CORT(X_T, Y_T)] \cdot d(X_T, Y_T) \tag{24}$$

where$\phi_k(\cdot)$ is an adjustment function that corrects the conventional distance$d(X_T, Y_T)$, defined in this paper, with a time-dependent coefficient:

$$\phi_k(u) = \frac{2}{1 + exp(ku)}, k \geq 0 \tag{25}$$

In this paper, k-means algorithm is selected to cluster the time series data. It is used with the help of python operation, after iterative convergence, to realize the evaluation of advantageous items according to the relationship between the number of medals won and the winning events.

To characterize the degree of adhesion of samples within a class and the degree of differentiation of samples between classes, a "profile value" is defined, and the profile value at point i :S_i

$$S_i = \frac{min(b_i) - a_i}{max[a_i, min(b_i)]} \tag{26}$$

wherea_i is the average distance from point i to the others within the class,b_i is the average distance from point i to all points in the other classes, andS_i varies from$[-1,1]$. Classification with a largeS_i mean value indicates that the classes are clearly differentiated, and negative values indicate points that may be misclassified. The optimal number of classifications requires that S_i has the largest average value and S_i has the smallest number of negative values.

Then for each cluster, its cluster center is recalculated. The new cluster center is the mean of all data points within that cluster and is calculated as follows:

$$c_i = \frac{1}{|S_i|}\sum_{x \in S_i} x \tag{27}$$

where S_i is the set of data points for the i cluster and $|S_i|$ is the number of data points in that set.

- **An Exploration of the Relationship Between Event and National Medal Winning:**

① **Spearman correlation analysis:**

Spearman analysis was used to analyze the relationship between the number of events and the number of medals won by a country in each session, and it was calculated that the value of ρ for USA was 0.72, the value of ρ for CHN was 0.95, and the value of ρ for CHN was 0.86. There is a strong positive correlation between the number of events and the total number of medals.

② **Time series clustering model advantageous item judgment:**

Selecting the Events that the country won each session with the number of medals won in that session, and explore which Events can help the country to win more medals. Determine the number of clusters k value, this paper uses the elbow method to estimate the optimal clustering k value of Events. The optimal clustering k-values of USA, CHA and GER are calculated to be 20, 7 and 11 in order as follows:

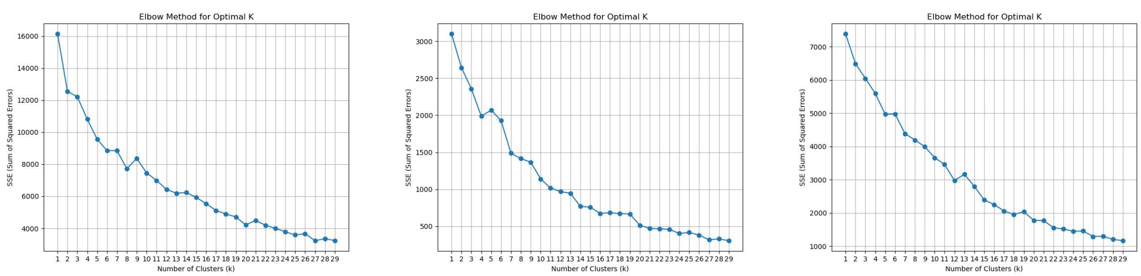

Figure 9 Elbow diagram (USA on left, CHA in meddle, GER on right)

The most dominant programs in USA, CHA, and CER, in order, are: "Basketball Men's Basketball", "Volleyball Women's Volleyball", and "Hockey Men's Hockey".

③ **Sports' exploration of the relationship between different countries:**

Selected countries win sports per session versus the number of medals won in that session to explore which sports are most important to different countries. Based on the solution principle of (ii), Swimming is calculated using python to obtain that Swimming is the most important to the three selected countries.

5.6.2 Gray Correlation-based Analysis of Host Country Awards and Events Correlations

In order to explore the impact of the host country's choice of projects on its medal results, this paper selects the relevant data of USA, FRA and GBR, which have been the host countries for many times, for analysis. With the help of python operation, the total medals corresponding to each EVENT of these three countries in each session are counted, and then their gray correlation coefficients are calculated with the following formula:

$$\varepsilon_i(k) = \frac{\min_i \min_k |x_0(k) - x_i(k)| + \rho \cdot \max_i \max_k |x_0(k) - x_i(k)|}{x_0(k) - x_i(k) + \max_i \max_k |x_0(k) - x_i(k)|} \tag{28}$$

Model solving: with the help of spsspro using gray correlation to calculate the host country EVENTS settings on the total number of medals, this paper shows the correlation of each country for the top 20 items:

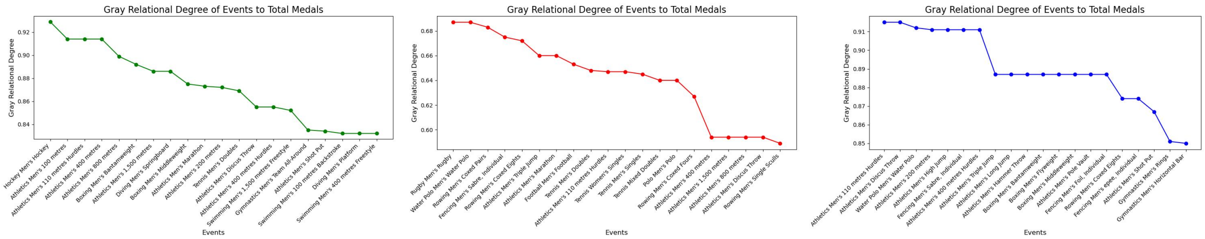

Figure 10 Correlation of GBR, FRA, and USA by Events on total number of medals

The analysis shows that Hockey men's Hockey, Athletics men's 100 metres, Boxing men's Bantamweight are good for increasing the number of GBR medals; Rugby men's Rugby, Water polo men's Coxed Pairs, and other disciplines are conducive to increasing the number of FRA medals; Athletics men's 110 metres hurdles, Athletics men 's Discus Throw, and Athletics men's high jump are good for increasing USA medals.

6 "Great coach" Effect

In order to investigate the impact of the "Great Coach" coaching on the number of medals, the specific flowchart for this section is shown below:

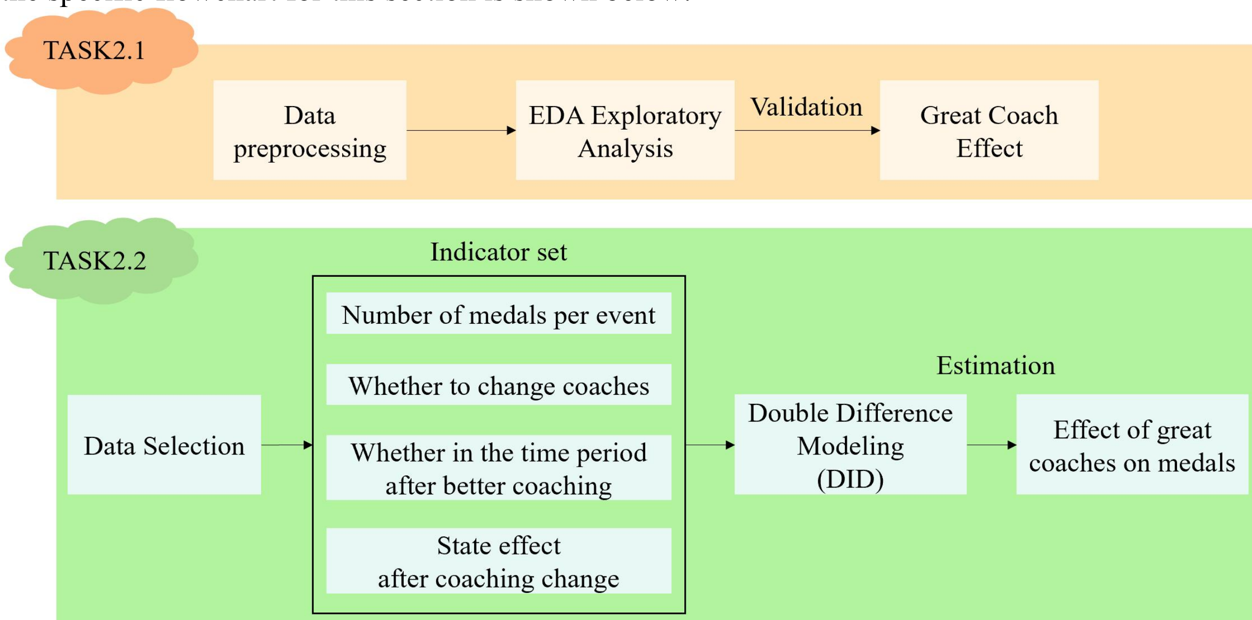

Figure 11 Flowchart of specific bodies in this section

6.1 Validation of the "Great coach" Effect Based on EDA Exploratory Analysis

The data provided were analyzed by EDA to identify the characteristics of the data and to verify the evidence of the "great coach" effect on the number of medals won. Use python to count the number of medals won by each country in each Olympic Games, and check whether there are outliers in the time series of the number of medals won by each country in each sport, which are considered to be the "Great Coach" effect, and explain the corresponding outliers by combining with the list of coaches of the corresponding sports in the past years published by the International Olympic Committee. The corresponding outliers are interpreted in the light of the list of coaches of the corresponding sports published by the IOC. Due to the limitation of the number of pages in this article, only 4 types of events with obvious "Great coach" effect are shown in this article.

Table 5 Events medal outliers

event	Swimming							Boxing					
country		Canada			China				Cuba			China	
time	gold	silver	copper	gold	silver	bronze	time	gold	silver	copper	gold	silver	bronze
2008	0	0	1	1	7	6	2008	0	4	4	2	1	1
2012	0	1	2	5	2	8	2012	2	0	2	1	1	1
2016	1	1	13	1	2	3	2016	3	0	3	0	1	3
2020	1	7	7	8	5	1	2020	4	0	1	0	2	0
2024	3	2	2	6	8	22	2024	1	0	1	3	2	0
event	Volleyball							Gymnastics					
country		USA			China				Romania			USA	
time	gold	silver	copper	gold	silver	bronze	time	gold	silver	copper	gold	silver	bronze
2000	0	0	0	0	0	0	1972	0	0	0	0	0	0
2004	0	0	0	12	0	0	1976	3	7	3	0	0	1
2008	0	12	0	0	0	12	1980	2	8	2	0	0	0
2012	0	12	0	0	0	0	1984	10	1	2	10	9	6
2016	0	0	12	12	0	0	1988	3	8	3	8	6	7

Through statistics, it can be found that in Swimming, Canada has a sudden increase in the number of medals in 2016-2020, but a decrease in the number of medals in 2024, but China has a sudden increase in the number of medals in 2024, and by checking the information, Rafadan Pierre moved from Canada to China as the head coach of the swimming team in 2022, and in 2024 to help China's swimming team's medal tally reach a new all-time high in the last five years; in Boxing, Cuba's gold medal tally declined in 2024, while China's gold medal tally showed a sudden increase in the same year, corresponding to the period when Fernandez Liranza Raul Ange transferred from Canada to China to coach in 2024; in Volleyball. USA's medal performance declined, and China regained the gold medal in the same year after two editions, corresponding to Lang Ping's coaching period in China; in Gymnastics, Romania achieved a breakthrough of zero medals in the event in 1976, and the number of medals declined in 1988, but USA won the first gold medal in the event in 1984, and has shown a sudden increase since then. In Gymnastics, Béla Károlyi coached Romania from 1976-1981, and then coached USA, with teams that helped their countries win multiple Gymnastics medals.

6.2 Estimation of the Impact of "Great coach" Based on the DID Model

➢ **Model building:**

Double difference (DID) model [10] is often used to assess the implementation effect of a policy or a program, the model is based on natural trial for research, the core idea is to analyze whether the policy or program has achieved the expected effect by comparing the difference between the experimental group and the control group before and after the implementation of the policy or program. The basic principle is shown in the table below:

Table 6 Principles of the double difference model

	Before investing in "great coach"	After investing in "great coach"	Difference
treat	A_1	A_2	$A_2 - A_1$
control	B_1	B_2	$B_2 - B_1$
DID	/	/	$\{(A_2 - A_1) - (B_2 - B_1)\}$

The DID model is applied with two differencing to eliminate the effect of heterogeneity. If the policy or program implementation has a significant effect, the coefficient of the DID estimator is statistically significantly different from 0. If the coefficient of the DID estimator is significantly greater than 0, it indicates that the policy intervention has had a positive impact; conversely, it indicates that the policy intervention has had a dampening effect.

In order to explore the impact of "Great coach" on the program, this paper uses the contribution value as the explanatory variable, which represents the number of medals that "Great coach" enhances for a particular sport in the country, and then calculates the growth rate of the corresponding program's medals to estimate the extent of its impact. then calculate the growth rate of medals of the corresponding program to estimate the degree of its impact.

Based on the DID model, the following base model is constructed:

$$C_{it} = \varphi_0 + \varphi_1 t\_g_i + \varphi_2 \eta_t + \varphi_3 \delta_{it} + \varepsilon_{it} \tag{29}$$

Where C_{it} denotes the country i in year t of $Change\_medal\ count$, t_g_i means whether it belongs to a country that has changed its coach, η_t denotes whether it is in the time period after better coaching, $\delta_{it} = t\_g_i \times \eta_t$ is the interaction term denoting the effect of the country i in year t after the coaching change, and the coefficient of the corresponding interaction term φ_3 denotes the change in the number of medals after the change of coaching.

➢ **Model solving:**

In this paper, the top 3 countries (USA, CHN and GER) and their corresponding dominant programs in the medal table are predicted in 5.4. The corresponding coefficient values are calculated with the help of python. R^2 All of them are greater than 0.8, the prediction results are better:

Table 7 Table of coefficients for DID prediction model

	φ_0	φ_1	φ_2	φ_3	R^2
USA	-1.5	-3.5	0.5	0.6	0.849
CHN	2.0	4.0	1.5	4.0	0.934
GER	-1	-2.0	0.5	4.5	0.975

➢ **Advice on investing in "great coach":**

the great coach effect contributes an average of 6.5 medals to USA sports and chooses to invest in basketball, with a medal growth rate of 23.21%; to GHA sports, with an average of 4 medals and chooses to invest in volleyball, with a medal growth rate of 400%; to GER sports, with an average of 4.5 medals and chooses to invest in hockey, with a medal growth rate of 207.31%; the value of contribution to the GER sport is an average of 4.5 medals and the choice to invest in field hockey with a growth rate of 128.46% in medals.

7 Unique Insights Into Olympic Medal Counts

In order to explore the characteristics related to the number of medals in the Olympic Games in more depth, based on the model established in the previous section, the relationship between medals and continents is explored from the perspective of each continent. The following aspects are mainly studied:

- **Exploring trends in medal counts over time by continent**
- **Explore trends in the number of athletes over time by continent**
- **Projected distribution of total medals and gold medals by continent in 2028**

With the help of python, we can get the consistent trend of the number of medals and athletes in each continent over time. Among them, Oceania, Africa and South America are relatively stable; Asia shows an increasing trend, Europe and North America show a fluctuating upward trend.

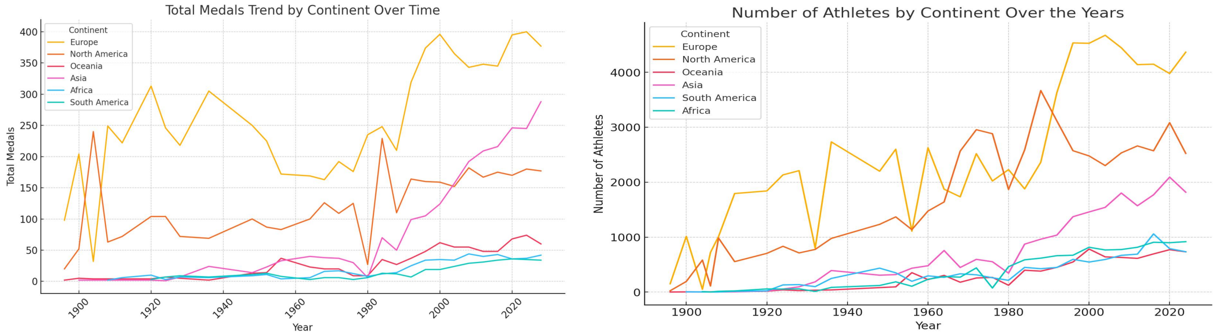

Figure 12 Trends in the number of data medals (left) and athletes (right) by continent over time

Based on the results of the previous Olympic medal prediction model, the total medals and gold medals by continent in 2028 are counted. Both percentages show Europe>Asia>North America>Oceania>Africa>South America.

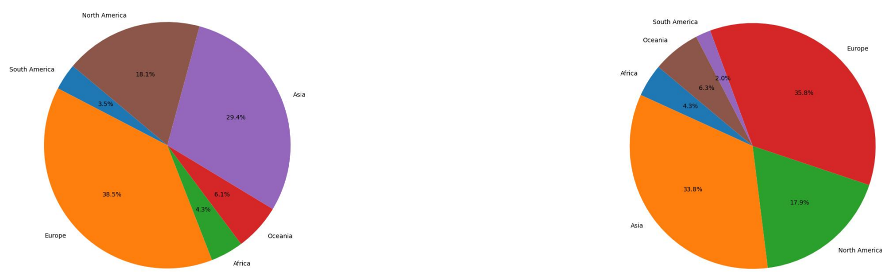

Figure 13 Projected distribution of total (left) and gold (right) medals by continent for 2028

Explain how these insight(s) can inform country Olympic committees:

- Increase investment in sports resources in Oceania, Africa and South America to provide more training facilities, event participation opportunities and coaching support to close the athletic gap between regions.
- Regular updating of the Olympic program to maintain its appeal to a global audience and to avoid over-concentration on the strengths of particular regions or countries.
- The Olympic Games are not only a stage for competition, but also a platform for showcasing culture and promoting regional development. Therefore, during the Olympic Games, countries should be organized to showcase their sports development plans and achievements and to promote global sports exchanges.

8 Sensitivity Analysis

8.1 Performance Evaluation of SVR-based Medal Prediction Models

In order to evaluate the performance of SVR, this paper utilizes python to train and predict the model of SVR and calculate the corresponding MSE, MAE and R^2 . The details are shown in the following table:

Table 8 SVR model predicted medal assessment indices for each type of medal

Prediction Performance (Optimized).	Gold	Silver	Bronze
MSE	0.3343	0.2588	0.5949
MAE	0.2414	0.2428	0.3346
R^2	0.9289	0.9322	0.9075

It can be obtained that the R^2 values of all types of medals are greater than 0.9, close to 1, and the model fits well. By MSE and MAE values close to 0.1, the model prediction error is small.

8.2 Performance Measures of Probabilistic Prediction Models for Zero Award Winning Countries

The performance of the model was evaluated using the logistic regression loss function, where the smaller the loss function, the better the model:

$$C = -\frac{1}{N}\sum_{i=1}^{N}[y_i ln(p_i) + (1 - y_i) ln(1 - p_i)] \tag{30}$$

where N is the sample size, y_i is the actual label $(1 or 0)$, and p_i is the probability predicted by the model.

With the help of python operations, the logistic regression loss for the training set is calculated to be 0.281<0.5 and the model reaches the learning result on the training set.

8.3 Evaluation of the Effectiveness of a Time Series Clustering-based Model for Evaluating the Relationship between Programs and Medals

According to the evaluation indicators obtained in each country, the profile coefficient tends to 1, which shows that the clustering is effective. The model classification is effective.

Table 9 Evaluation indicators of clustering effect of selected countries event

	Contour Coefficients	DBI	CH
USA	0.903	0.9527	68.7944
CHA	0.897	0.9051	52.6777
GER	0.912	0.8907	38.5480

8.4 Parallel Trend Test for "Great coach" Impact Estimation Based on DID Modeling

An important prerequisite for applying the double difference model is the existence of a common trend between the treatment and control groups before the intervention, i.e., the parallel trend (CT) assumption has to be satisfied. Therefore, in this paper, we qualitatively obtained by drawing graphs that the time trend graphs of the treatment and control groups before the implementation of the intervention followed essentially the same trend, indicating that the model meets the parallel trend assumption. After the implementation of the intervention, the trend changed, i.e., the time trend graph of the treatment group deviated after $t = 0$, indicating that the intervention had some effect.

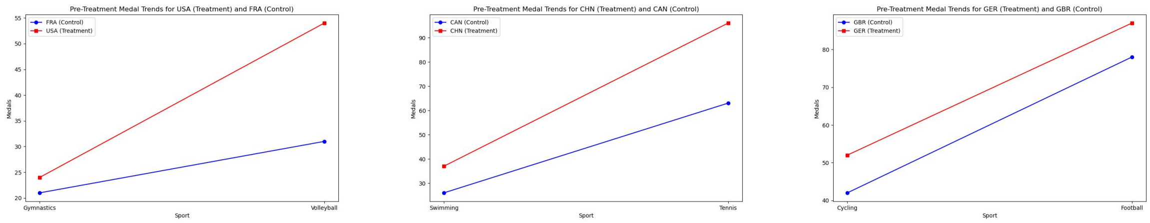

Figure 14 Double difference treat and control trend plots (USA on left, CHA in meddle, GER on right)

9 Model Evaluation and Further Discussion

- **Strength**:

(1) The SVR Olympic medal prediction model can fit the volatility of Olympic medal counts better, while using Bootstrap sampling method to provide confidence intervals for medal count prediction, giving the intervals of medal count prediction for different countries, and the model is reliable and interpretative.

(2) The coupled SVM and Logistic model can be used to accurately determine which non-winning countries are likely to win medals at future Olympics based on the existing training set, while the Logistic regression model provides good probabilistic predictions, making the results more flexible and easy to interpret.

- **Weakness:** the SVR Olympic medal prediction model has higher computational complexity and longer training time when the data volume is large.
- **Further discussion:** In the SVR Olympic medal prediction model, more factors that may affect the number of medals can be considered, such as the country's sports investment, the training level of athletes, and political factors.

References

[1] https://www.bugxianbao.com/n/11244.html

[2] Deng Yunlong. Analysis of the host effect of the Olympic Games and research on the participation strategy of Chinese sports delegations[J]. China Sports Science and Technology,2007,(01):3-6.DOI:10.16470/j.csst.2007.01.001.

[3] LUO Yubo, CHENG Yan fang , LI Mengyao ,et al. Prediction of the number of medals and overall strength of China in Beijing Winter Olympic Games - Based on host effect and gray prediction model[J]. Contemporary Sports Science and Technology,2022,12(21):183-186.DOI:10.16655/j.cnki.2095-2813.2112-1579-2956.

[4] CaiL ,LeiD ,WuF , et al.Combining Fourier Fractional GM(1,1|sin) Model with Rat Swarm Optimizer for Employment Rate Prediction[J].IEEJ Transactions on Electrical and Electronic Engineering,2024,19(6):1007-1013.

[5] SAN Rui,WANG Shuhua,GAO Donglian,et al. Research on combined forecasting model based on time series model and gray model[J]. Journal of Yanshan University,2012,36(01):79-83.

[6] ZHAO Wu,TANG Tianren,WANG Guofan. Research on the prediction of medals of London 2012 Olympic Games by optimal parameter v-SVR model[J]. Journal of Anhui Normal University(Natural Science Edition),2012,35(04):381-386+395.DOI:10.14182/j.cnki.1001-2443.2012.04.008.

[7] Hwa-Yeon Kim,Yong-Ze Liu. Research on financial fraud identification model based on fraud triangle theory--a coupled empirical analysis of support vector machine and logistic regression[J]. Journal of Dalian University of Technology (Social Science Edition),2014,35(01):92-97.DOI:10.19525/j.issn1008-407x.2014.01.017.

[8] Liao T. Clustering of time series data-a survey[J].Pattern Recognition,2005,38:1857-1874.

[9] Montero P,Vilar JA.TSclust:an R package for time series clustering[J].Journal of Statistical Software,2014,62(1):1-43.

[10] Zheng Chunli,Luo Chuanjian. The impact of issuing green bonds on the economic efficiency of listed companies--an analysis based on double difference model[J]. Wuhan Finance,2020,(10):38-44.