Problem Chosen	2025 MCM/ICM Summary Sheet	Team Control Number
C		2501869

Medal Prediction Based on Tobit and Hurdle Models and Great Coach Effect Insights

Summary

In **Question 1**, this paper proposes a **Hurdle-Tobit fusion model** framework to address the characteristics of the **discrete nature** of the number of Olympic medals, the **zero inflation** characteristics, and the **heterogeneity between countries**. The Hurdle model solves the zero value problem through **two-stage modeling** (**logit classification** and **truncated count distribution**), and the Tobit model handles right censored data. Both models introduce the Revealed Comparative Advantage Index RCA and incorporate heterogeneity factors such as the discrete size of athletes and the host effect. Then, by comparison, we found that the hurdle and Tobit models were significantly **better than** the machine learning methods (random forest and XGboost) in terms of mean square error and prediction coverage. Among them, the **Hurdle** model has a better predictive effect on the number of medals of **sports powers**, while **Tobit** has a better predictive effect on countries that have won **a few medals**. The prediction results show that **the United States, China, and India** will see the largest increases in medals, while **France, Russia, and Germany** will see significant decreases. The probability of **Angola, Bangladesh, and Cambodia** winning their first medal is **49.2%, 46.7%, and 42.5%**, respectively. Finally, using model complementation and variable reconstruction, it is revealed how the host selection project affects the number of medals.

In **Question 2**, the **Bayesian Change-point Detection** method is used to identify significant changes in the number of medals won by countries in specific events (such as a sudden breakthrough after a long period of depression), and attribute them to the "great coaching effect". In order to solve the problem of frequent changes in the simple weighted medal data, an innovative **data preprocessing algorithm with memory enhancement** is used to preprocess the original medal data, which can effectively distinguish the significance of the medal data and the dynamic self-reinforcing mechanism from the real changes driven by the coaching effect. Then, a **two-stage least squares (2SLS)** method is used to construct a causal regression model to quantify the contribution of outstanding coaches to the number of medals, overcoming the endogeneity bias of the traditional OLS method. Model validation shows that 2SLS significantly outperforms OLS in explaining the coaching effect. Based on this, it is proposed that **the United States, the United Kingdom, and France** should prioritize the introduction of multinational coaches in wrestling, rowing, table tennis, and other events, which is expected to increase the number of medals by **1-2**.

In **Question 3**, we continue the analysis of Questions 1 and 2. Based on the **Bayesian change-point detection** method, we identify the potential "great coaching effect" time nodes in each sport in previous Olympic Games and incorporate them into the Hurdle and Tobit models as **key explanatory variables** for reanalysis. At the same time, in order to capture the dynamic self-reinforcing mechanism of medal counts, we further explain the **lagged terms** (such as the host country effect lagged by 4 years, the historical medal count lagged term, etc.) to analyze the **heterogeneity** of short-term and long-term effects. Ultimately, we obtain policy implications for promoting **post-competition resource redistribution** and for balancing the "**depth**" and "**breadth**" of National Olympic Committees.

Keywords: Tobit, Hurdle, 2SLS, Bayesian Change-Point Detection, Medal Prediction

Contents

1 Introduction

1.1 Background

The Olympic medal table reflects national sporting prowess and serves as a global competitive benchmark. At the 2024 Paris Olympics, the U.S. and China tied with 40 golds each, while host France ranked fifth in golds (16) but fourth overall. Notably, Albania and Cape Verde achieved their first-ever medals, yet over 60 countries remain without Olympic medals.

1.2 Restatement of the Problem

Task 1 Develop an Olympic medal prediction model using historical data to forecast: (1) 2028 Los Angeles medal counts per country with prediction intervals and accuracy metrics, identifying upward/downward mobility relative to 2024. (2) First-time medal-winning probability for nations, contextualized through historical breakthrough patterns. (3) Sport-event structure's impact on medal distribution, including identification of national dominance events. (4) Quantification of host advantage through program adjustment impacts.

Task 2 (1) Statistically validate coach replacement effects on medal changes and quantify "great coach" contributions. (2) For three nations, identify underperforming sports with medal potential and assess feasibility/benefits of elite coach recruitment.

Task 3 (1) Identify non-traditional factors influencing medal allocations. (2) Generate evidence-based policy recommendations for Olympic committees.

1.3 Literature Overview

Empirical analysis of the Olympic Games has emerged as a critical area of research, encompassing both forecasting (e.g., De Bosscher et al., 2006) and broader analytical perspectives (e.g., Streicher et al., 2020).

Since Ball (1972) introduced a correlation-based scoring model, forecasting methodologies have become increasingly sophisticated. Initially, many researchers relied on ordinary least squares (OLS) regressions due to their ease of interpretation (e.g., Baimbridge, 1998; Condon et al., 1999; Kuper and Sterken, 2001). However, a significant challenge in predicting Olympic medals is accurately reflecting the substantial number of nations that have yet to achieve any medals. The exponential function employed in these models tends to impose penalties on low predicted medal counts. Consequently, some scholars have transitioned to Poisson-based models (e.g., Lui and Suen, 2008; Leeds and Leeds, 2012; Blais-Morisset et al., 2017).

Afterwards, machine learning techniques (e.g. Random forest and XGboost) have gained traction in sports-related contexts. The random forest approach has demonstrated exceptional performance in tasks such as predicting the outcome of football matches (Groll et al., 2019) or horse racing outcomes (Lessmann et al., 2010). However, random forests and XGBoost, as ensemble learning methods based on decision trees, excel at handling non-linear relationships and feature interactions, but they lack the ability to explicitly model truncation and zero-inflated data. This means that these algorithms may not accurately capture the truncation points or zero-value generation mechanisms in the data, leading to prediction bias (Breiman, 2001; Chen & Guestrin, 2016).

Given that the dependent variable, typically the number of medals, has been set to zero, most authors have employed Tobit regression to predict Olympic success (e.g., Tcha and Pershin,

2003; Forrest et al., 2015; Rewilak, 2021). However, recently, a two-step approach, estimating the probability of winning any medal before determining the exact number of medals in case of success, has become more prevalent. In particular, both Scelles et al. (2020) and Rewilak (2021), employing a Mundlak transformation of the Tobit model, could, again, increase the prediction accuracy with their respective Hurdle models.

This finding aligns with the modeling approach adopted in this study, which utilizes Tobit and Herdle models to address the unique challenges associated with Olympic medal forecasting.

1.4 Our Work

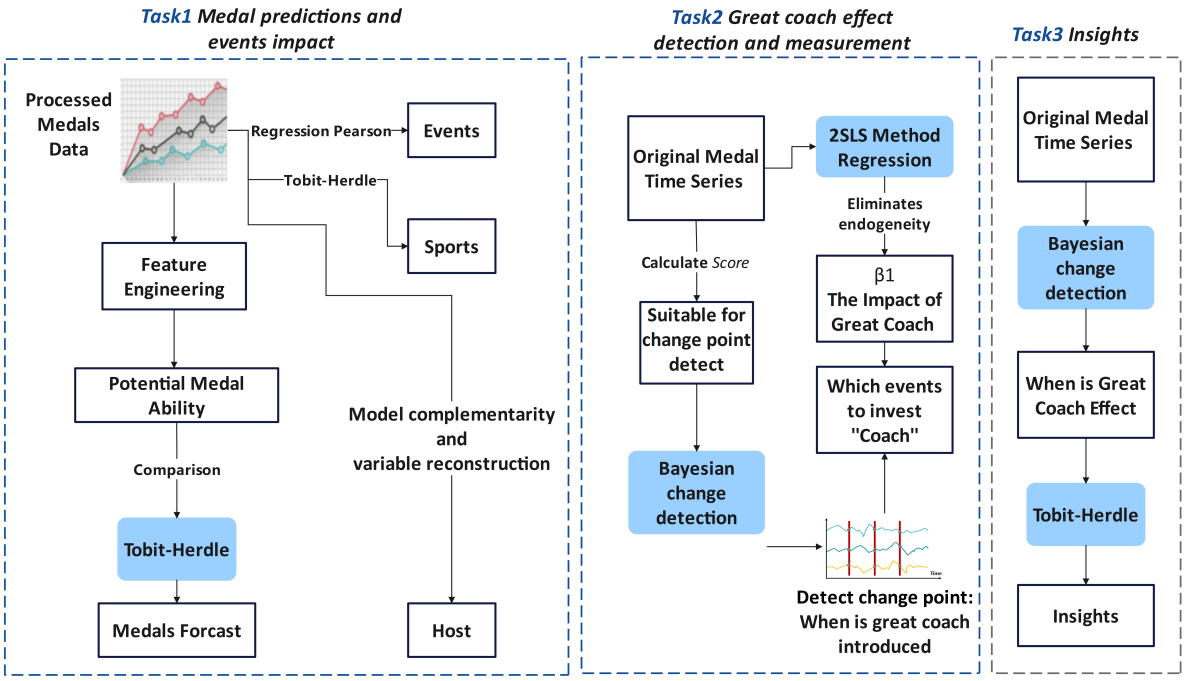

Figure 1 Our Work

2 Assumptions and Justification

- All models are built based on the normal holding of the Olympic Games, and situations such as the postponement of the Olympic Games will not occur.
- The models are built without the influence of external factors (such as political events and national bans), and are based only on existing data.
- All data used in this article is real data without statistical errors.

3 Notations

Notation	Significance
$Mapdisq^*_{i,t}$	Potential medal ability
RCA	National dominance sports index
$Na_{i,t}$	Categories to which the number of players belongs
NAE	The number of events that won medals in a sport for a country
$M_{i,t}$	The expected number of medals won by country
$Medal\_PCA$	The number of medals after dimensionality reduction

4 Data Processing

4.1 Data Cleaning

To facilitate forecasting, we exclude vacant items in the tabular data, harmonize country names (e.g., the Russian Olympic Committee and Unified Team in 2024 are grouped together as Russia), harmonize athlete names (e.g., due to differences in hosts and entry rules, Ma Long and Long Ma are actually the same), harmonization of event names (e.g., "Tabble Tennis Men's single" in 2024 is the same as "Men's single " of table tennis in 2020).

4.2 Data Overview

The following figure shows a box plot of the total number of medals for all participating countries in the Olympics from 1896-2024. At the same time, we extracted the total medal counts of USA, China, France, Great Britain and Australia to draw a line graph.

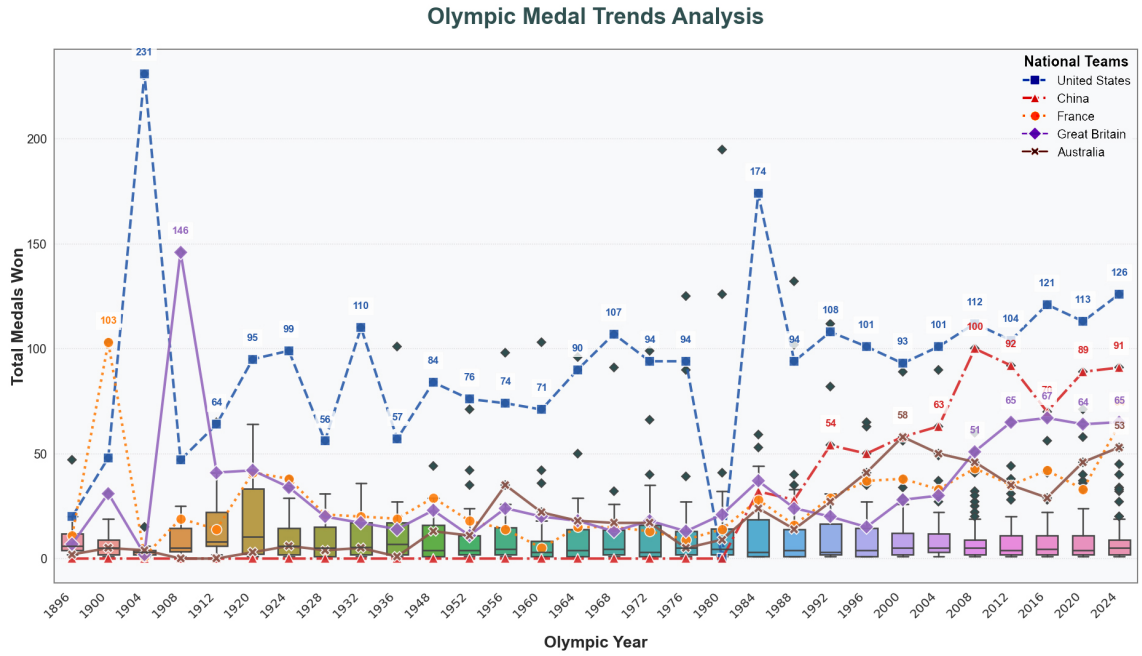

Figure 2 Box Plots of Total Medals from 1896 to 2024 and Ray Plots of Medals for 5 Countries

As shown in the figure, it can be seen that the box plot shows a right-skewed distribution, i.e., the maximum is much higher than the median and the tails extend to the right, reflecting that a few countries dominate the medal table.

We calculated the mean, standard deviation, maximum and minimum values for the number of athletes, number of medals, total number of medals, host, number of sport type and number of events in which the countries participates for the 2024 Olympic Games in Paris, France, as shown in the following table.

Table 1 Descriptive Statistics

Variable	Mean	SD	Min	Max
Number of Athletes	73.40	144.9809	2	856
Total Medals	11.42	19.7028	1	126
Gold Medals	1.59	4.97	0	40
Silver Medals	1.59	4.83	0	44
Bronze Medals	1.85	4.83	0	42
Host Country	0.01	0.07	0	1
Number of Events	31.27	46.84	1	234
Number of Sports	10.07	10.30	1	47

According to the above table, it can be seen that

- The large standard deviations of the number of athletes and medals indicate a large degree of dispersion.
- Only 1% of the observations are host countries, indicating that the sample of host countries is extremely small, which may affect its statistical significance, which may affect its statistical significance and should be interpreted with caution.
- The number of athletes is highly correlated with the number of medals, but it should be noted that there is covariance between the host country variable and the number of athletes in the model, resulting in the host country effect being explained by the number of athletes.

5 Comparative Modeling of Olympic Medal Projections: Hurdle-Tobit Framework with Host Country Strategy Analysis

5.1 Hurdle and Tobit Models

Since the number of medals is a discrete non-negative integer (e.g., 0, 1, 2, etc.), it becomes critical to deal with such a large amount of discrete data, and therefore the counting model is used for the analysis. So we first considered the suggestion of Blais-Morissetr, and Fortin (2017) that a Poisson model be used for handling such data in this case. However, due to significant heterogeneity across countries, the variance is larger than the mean (i.e., the overdispersion problem), which makes the assumptions of the Poisson model not valid.

To deal with the over-dispersion problem, we turn to consider the Zinb model (Zero-Inflated Negative Binomial Model), which is applicable when there are a large number of zeros in the data and assumes that the zeros come from two different sources, a “structural zero” (e.g., the country did not (e.g. the country did not participate in the Olympic Games) and “sampling zeros” (e.g. the country participated but did not win a medal). However, since the study

focuses on countries that have participated in the Olympics, structural zeros do not exist and the assumptions of the Zinb model no longer apply.

Then, we turn to use Hurdle model, which differs from the Zinb model in that it does not distinguish between the sources of the zero values, but rather divides the model into two parts: the first part is the decision of whether or not there is at least one medal (Binary Classification Problem), and the second part is the distribution of the number of medals in the event that there is at least one medal. This is similar to the Tobit model (Truncated Regression Model), but whereas the Tobit model assumes that observations below a certain Hurdle are truncated, the Hurdle model explicitly distinguishes between the two processes of whether or not that Hurdle is crossed.

Therefore, for the two different scenarios of countries that have never won a medal and sports powerhouses, we improve the use and compare the predictive effectiveness of the four methods of Tobit model, hurdle model, random forest and XGBoost, and select the optimal model for the different scenarios.

Step 1: Variable definition and treatment.

$NA_{i,t}$ denotes the number of athletes competing in the Olympics for country$i$ in year t, which was initially analyzed as a continuous variable. However, the marginal returns to the number of athletes may be nonlinear and the exact number of participants cannot be known in advance when predicting future medal totals. Therefore, we introduce $Na_{i,t}$ to discretize the number of participating athletes into four categories:

- $Na_{i,t} = 1$: 0 to 9 athletes;
- $Na_{i,t} = 2$: 10 to 49 athletes;
- $Na_{i,t} = 3$: 50 to 149 athletes;
- $Na_{i,t} = 4$: 150 athletes and over

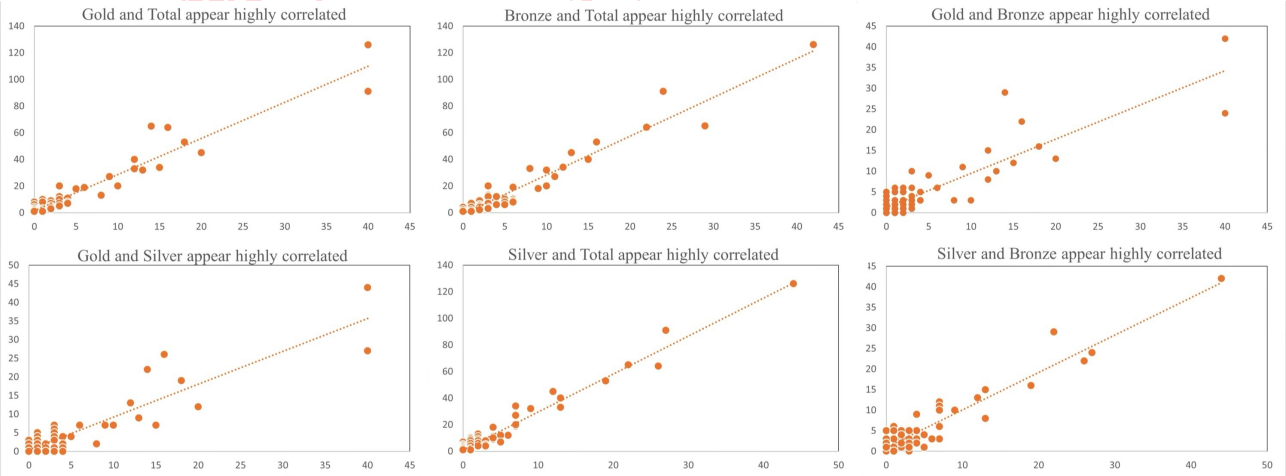

Figure 3 The correlation between medal counts

From the figures above, it is worth noting that there is a significant multicollinearity between the number of gold medals, silver medals, bronze medals, and medal table. To simplify the

model, we actually use Principal Component Analysis (PCA) to downscale to one dimension before using it, i.e., $Medal\_PCA$.

At the same time each country should have its dominant sport, in order to include it in the model we introduce RCA to refer to quantify the dominant sport of each country. RCA is given by the following formula:

$$RCA = \frac{M_{ij}/M_i}{T_j/T} \tag{1}$$

Here, M_{ij} represents the number of medals for country i in sport j;M_i represents the total number of medals for country i; T_j represents total global medals in sport j; T represents total medals in all sports.

$RCA > 1$: country has a comparative advantage in the given sport;

$RCA < 1$: country has a comparative disadvantage in the given sport.=

The figure below shows the RCA index heat maps of the countries that won medals at Paris 2024.

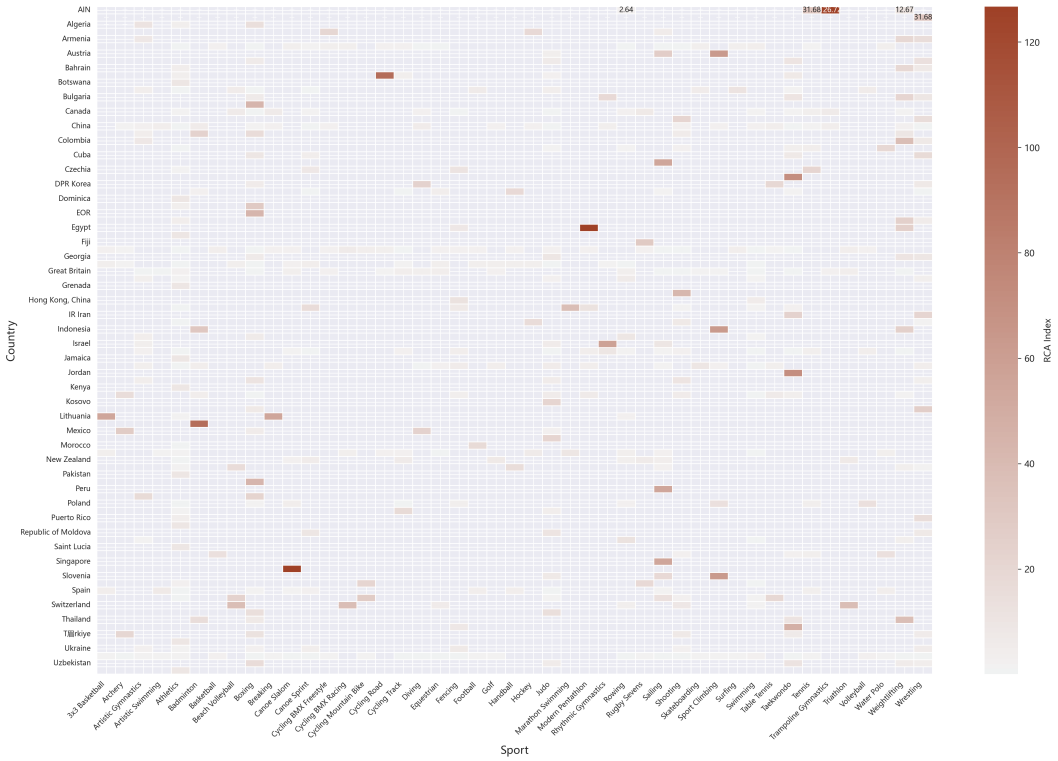

Figure 4 RCA Index Heatmap

Step 2: Define latent variable $X_{i,t}\Theta$ and potential medal ability$Mapdisq^*_{i,t}$.

Depending on the set of explanatory variables selected, we define latent variable $X_{i,t}\Theta$ by:

$$X_{i,t}\Theta = c+\lambda \ln N_{i,t-4}+\rho Host_{i,t}+\sum_p \gamma RCA_{p,i}+\delta Medal\_PCA_{i,t-4}+\sum_p \zeta Num\_Events_{p,i,t}+\kappa \sum_p NAE_p, \tag{2}$$

Here, $X_{i,t}\Theta = [X_{i,t}\Theta_{Gold}, X_{i,t}\Theta_{Silver}, X_{i,t}\Theta_{Bronze}]^T$. $Host_{i,t}$ is the host variable ($Host_{i,t}$=1, country i is the host of the t Olympic Games). $Num\_Events_{p,i,t}$ is the number of events in

which the i country participated in the p sport of the t Olympic Games. c, λ,ρ, γ, δ, ζ, $\kappa \in \mathbb{R}^3$are all coefficients, representing the value of $X_{i,t}\Theta$ for gold, silver and bronze respectively, which are calculated by maximum likelihood estimation (MLE) based on historical data.

We define potential medal ability $Mapdisq^*_{i,t}$ by:

$$Mapdisq^*_{i,t} = X_{i,t}\Theta + u_i + \epsilon_{i,t} \tag{3}$$

Here, u_i represents country-specific random effects (such as culture, training system and other unobserved factors), following a normal distribution $N\left(0, \sigma_u^2\right)$; $\epsilon_{i,t}$ is a random error term.

Step 3: Construct Tobit model.

For the Tobit Model, the general specification is:

$$Mapdisq_{i,t} = \begin{cases} Mapdisq^*_{i,t} & , \text{ if } Mapdisq^*_{i,t} > 0 \\ 0 & , \text{ if } Mapdisq^*_{i,t} \leqslant 0 \end{cases} \tag{4}$$

The predicted medal count $Mapdisq_{i,t}$ is a truncated version of the potential medal ability $Mapdisq^*_{i,t}$.

Step 4: Construct Hurdle model.

1) Calculate the probability of winning $P\left(Mapdisq^*_{i,t} > 0\right)$.

   We use the Logit model to calculate the probability that the i country will get at least one medal, as follows:

$$P\left(Mapdisq^*_{i,t} > 0\right) = \frac{1}{1 + e^{-X_{i,t}\Theta}} \tag{5}$$

2) Calculate the expected number of medals won by the winning country$M_{i,t}$.

   First, we define $\mu_{i,t}$ as the theoretical mean of the number of medals for country i under the condition of "winning at least one medal", the formula is:

$$\mu_{i,t} = e^{X_{i,t}\Theta} \tag{6}$$

   Then, to adjust the expected value since only the winning country is considered (i.e. zero values are excluded), we define the truncated correction $E(Mapdisq^*_{i,t})$ as:

$$E(Mapdisq^*_{i,t}) = \mu_{i,t} \cdot \frac{1}{1 - (1 + \mu_{i,t}/r)^{-r}} \tag{7}$$

   where r is a discrete parameter.

   Finally, the expected number of medals won by the winning country $M_{i,t}$ is given by:

$$M_{i,t} = P\left(Mapdisq^*_{i,t} > 0\right) \cdot E(Mapdisq^*_{i,t}) \tag{8}$$

5.2 Model Selection and Validation

5.2.1 Model Selection: Hurdle, Tobit, XGBoost and Ramdom Forest

In this study, we employ four models—the Tobit model, the Hurdle model, Random Forest, and XGBoost—to predict the total number of medals won by the United States in the 2024 Olympic

Games in Paris, France. This approach is taken to assess the accuracy of the four models' prediction results.The comparison between the prediction results based on the four methods and the real data is shown below.

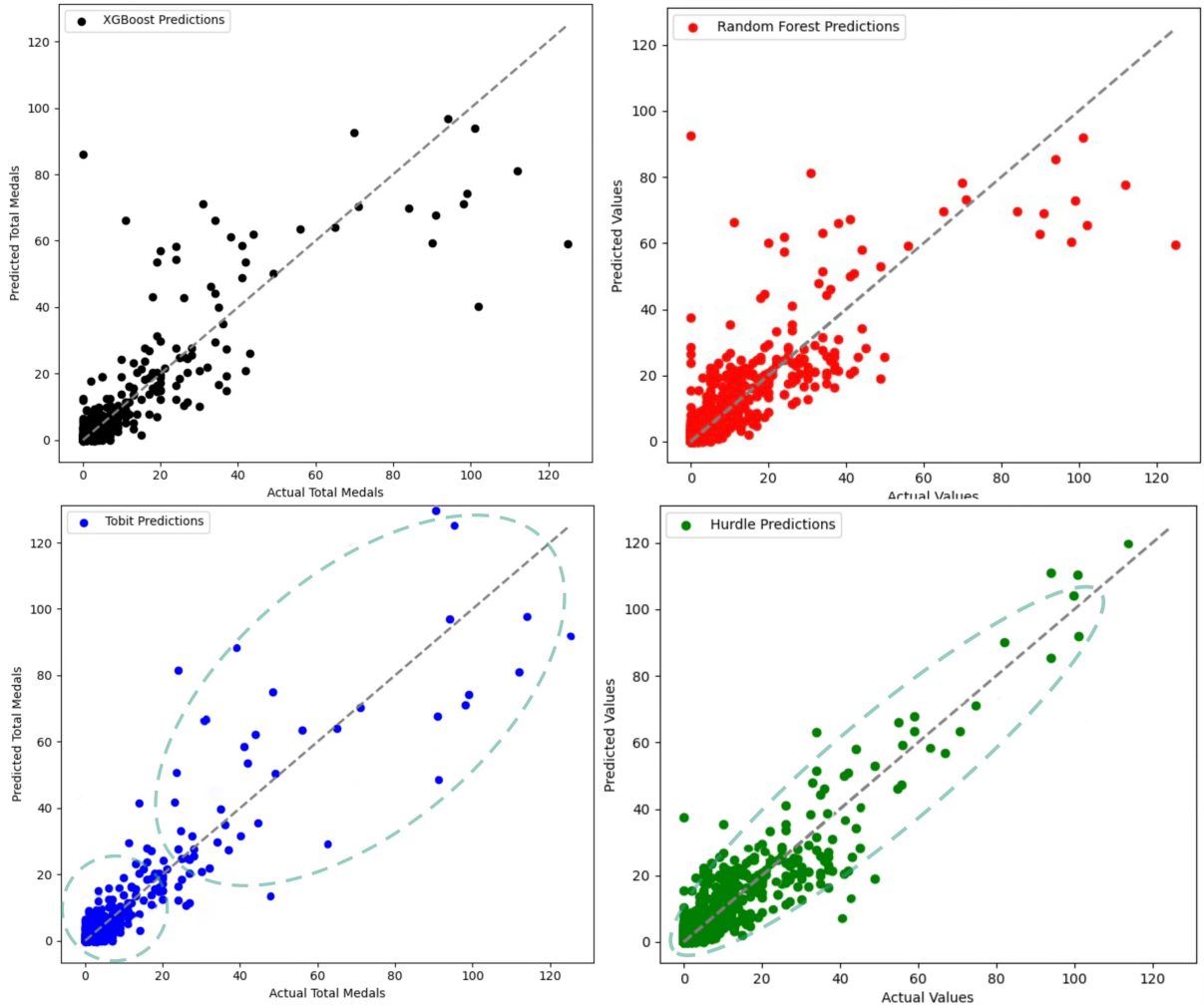

Figure 5 Prediction Results Based on the Four Methods Compared with the Real Data

Table 2 Model Performance Comparison (MSE and R^2)

	Hurdle	Tobit	XGBoost	Ramdom Forest
MSE	20.15	23.01	63.67	51.25
R^2	0.8545	0.8026	0.7121	0.758

As shown in the figure, Hurdle and Hurdle perform better, and the prediction results of Random Forest & XGBoost are poorly fitted to the real values. In the case of Random Forest with XGBoost, the predicted values are biased in the region of actual medal count >15, i.e., it is difficult to deal with the distributional characteristics of zero-inflated data effectively. Meanwhile, the prediction points in countries with large medal counts (actual values > 50) are scattered and irregular, reflecting the model's insufficient ability to capture extreme values.

Meanwhile, the MSE of Hurdle and Tobit is significantly lower than that of XGBoost and Random Forest, which indicates that the former two models are better in prediction accuracy.The

R^2 of Hurdle and Tobit is significantly higher than that of the former two. This indicates that Hurdle and Tobit models are more effective in explaining data variability.

5.2.2 Model Validation: 2024 Olympic Games Predictions Based on Hurdle and Tobit

We choose the Hurdle and Tobit models with better prediction performance to present the prediction results for the prediction interval and accuracy of the total medal count of 2024 Paris Games.

Table 3 Forecast of Olympic Medals for the 2024 Paris Games

Country	Actual Medals	Hurdle Model			Tobit Model		
		Forecast	Lower CI	Upper CI	Forecast	Lower CI	Upper CI
United States	126	**120**	103	134	**105**	95	113
China	91	**89**	86	93	**106**	95	117
Great Britain	65	**68**	63	53	**56**	51	61
France	64	**70**	63	77	**69**	66	71
Australia	53	**48**	44	52	**46**	44	58
Japan	45	**47**	42	51	**44**	42	46
Italy	40	**43**	38	48	**48**	43	53
Rate of right forecasts for 2024 Paris Olympic Games							
All countries (192)							
CI to 95%(+ or -2)		88.6% (93.2%)			83.8% (90.6%)		
Exact forecasts (+ or - 1)		21.9% (77.1%)			43.2% (74.5%)		
Exact forecasts (0 medal) (107 countries)		41.3%			69.1%		
Exact forecasts (non-0 medal) (85 countries)		22.4%			10.4%		
Countries with at least 3 medals(56)							
CI to 95%(+ or -2)		62.3% (76.8%)			50.2% (69.6%)		
Exact forecasts (+ or - 1)		8.9% (38.5%)			8.9% (31.4%)		

Note: CI = confidence interval; forecasts in bold.

First, in the 95% confidence interval prediction coverage, the Hurdle model covers 88.6% of all 192 countries, which is higher than the 83.8% of the Tobit model; among the 56 countries with at least 3 medals, its coverage (62.3%) is also much higher than that of the Tobit model (50.2%). In addition, the prediction accuracy of the Hurdle model for countries with non-zero medals (±1 error rate of 22.4%) is more than twice that of the Tobit model (10.4%), highlighting the advantage of its two-stage modeling in adapting to zero-inflated data.

Although the Tobit model has a higher rate of fully accurate predictions in zero-medal countries (69.1%) than the Hurdle model (41.3%), this advantage comes at the expense of the ability to predict non-zero medals.The Hurdle model is more balanced between the two types of scenarios by separating the mechanisms of zero value generation and positive prediction, and it is particularly robust in countries with high medal counts (±1 error bracket value 38.5% vs. 31.4%).

In summary, the Tobit model fits better around zero medals, thanks to its truncated regression

assumptions for the lower bound values. However, as the number of medals increases, the dispersion increases significantly, exposing the limitations of the single continuous distribution assumption. In contrast, Hurdle's model, through two-stage modeling, shows better prediction results in both the zero-value region and the high medal region, with a concentrated distribution of points and low dispersion, which effectively balances the zero-value generating mechanism with the flexibility of positive prediction.

5.3 2028 Los Angeles Olympics Medal Count Projections

The figure below shows the medal standings for all participating countries that are predicted to win at least one medal.

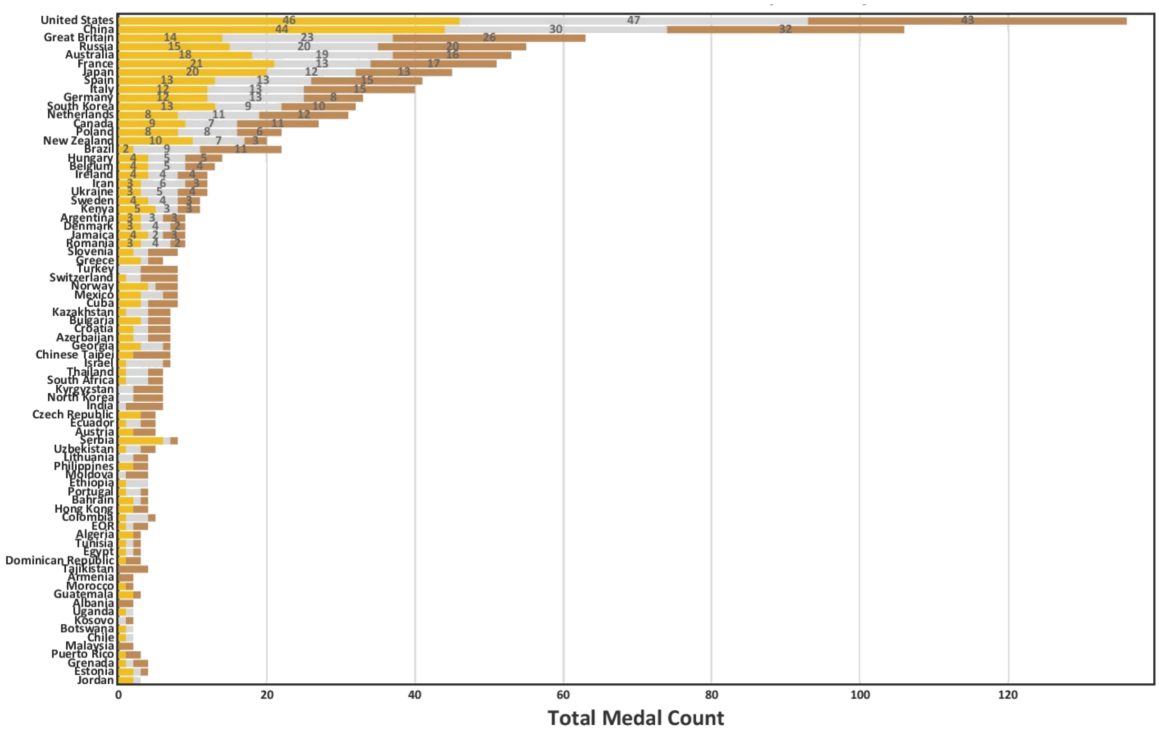

Figure 6 2028 Olympic Medal Predictions Gold/Silver/Bronze Distribution by Country

Based on the above prediction results, we have listed the top five countries with the highest increase and decrease in total medals, as follows.

Table 4 Top 5 Countries with the Highest Improve and Decline

Improve	USA	China	India	Brazil	Japan
Magnitude	15	10	7	5	5
Decline	France	Russia	Germany	Great Britain	Australia
Magnitude	-8	-4	-4	-3	-2

5.4 Projecting First-Time Medal Winners and Probability Estimates

The prediction results show that a total of three countries will win their first medals at the 2028 Los Angeles Olympics, namely: Angola, Bangladesh and Cambodia. The probability of

each country winning their first medal is shown in the table below.

Table 5 Projected First-Time Medal Winners and Probability

Country	Angola	Bangladesh	Cambodia
Possibility	49.23%	46.67%	42.49%

5.5 The Influence of Item Selection on the Number of Medals

The below table shows the mean values of the statistical results obtained by each country when Tobit and Hurdle were used.

Table 6 Results of Two Forecasting Models

Variables	Hurdle Model		Tobit Model	
	Coef	SD	Coef	SD
Constant				
Host country in 4 years	0.336*	0.1	8.864***	2.02
Host country t	1.566***	0.11	12.520***	2.03
Host country 4 years ago	0.85	0.11	-4.760**	2.08
Athletes [0,10]	0.016*	0	0.897*	0.02
Athletes [10,50]	0.510*	0.23	5.126***	0.67
Athletes [50,150]	0.989*	0.24	7.394***	0.87
150 athletes and more	1.559***	0.27	9.314***	1.15
Medals_PCA(t-4)	3.725***	0.12	3.42***	4.57
RCA	2.538***	0.22	1.828**	1.07
Medals_PCA(t-4)	3.725***	0.12	3.42***	4.57
Num_Events	0.538**	0.22	1.828**	1.07
Num_sports	0.618**	0.21	1.719**	0.97
NAE	2.742***	0.22	2.705***	0.26
$g_{i,t}$	-3.408***	0.25	—	—
σ_u^2	0.115***	0.03	26.68***	1.66
Observations total	529	—	1232	—
Observations noncensored	—	—	529	—

*Note: ***Significant at the 1 percent level; **5 percent level; *10 percent level.*

The results show that both the number of events and NAE have a positive and significant effect, and this significance is common to all countries. The coefficient of RCA is large and the variance is large, which may indicate significant heterogeneity in the importance of different sports for different countries. At the same time, the host country has a positive impact in both the Tobit and Hurdle models, and in the Hurdle model, the coefficient of the host country is smaller than that of the Tobit. The explanation is that there is obvious multiple collinearity between the host country and other variables such as NAE and RCA, and these variables will dilute the significance of the host country, and Hurdle is more likely to capture this effect.

5.5.1 Relationship Between Olympic Events and Medal Distribution

Initially, an effort was made to construct a graph representing the relationship between $MeanNAE$, Num_Events, and $Total$ in historical data for China, the United States, Japan, and France. Subsequently, the Pearson correlation coefficients of these countries regarding the two covariates and Total were calculated. The graph of the relationship between $Total$ and $meanNAE$ and correlation coefficient tables are presented below.

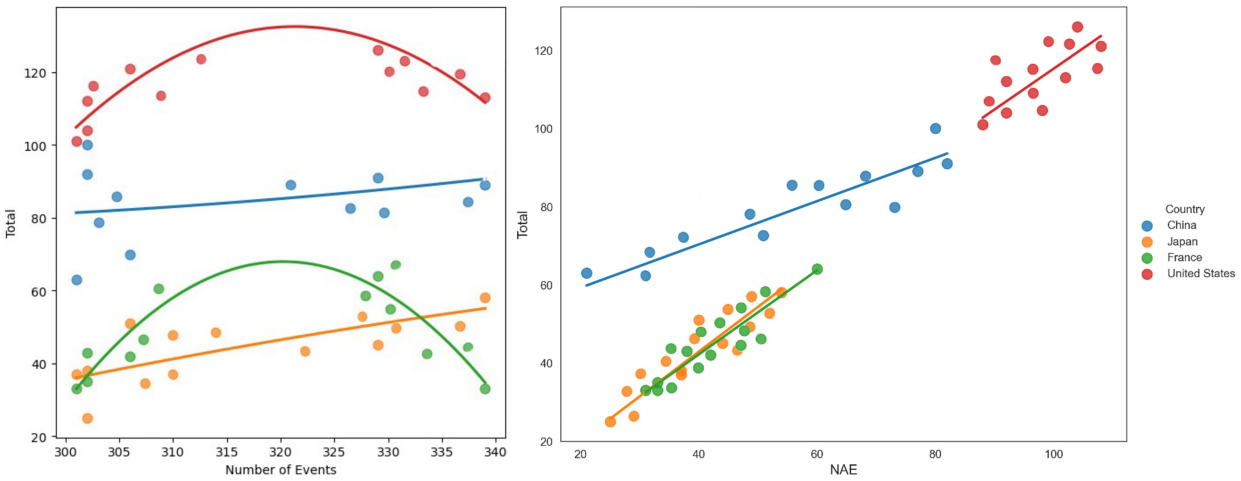

Figure 7 Comparison of RCA, Num_Events and $Total$ by Country

Table 7 Pearson Correlation Coefficient Tables

	Mean	SD
num_events	0.238	0.216
NAE	0.952	0.051

The findings revealed a robust linear relationship between $MeanNAE$ and $Total$, with a concomitant small standard deviation, suggesting its universality. However, the significance of Num_Event is negligible, and this phenomenon exhibits substantial heterogeneity across different countries.

5.5.2 Analysis of Key Sports and Their Importance for Different Countries

The RCA index is a quantitative metric that can be used to compare the relative popularity of different sports in different countries. As illustrated by the heatmap, Modern Pentathlon has emerged as Egypt's preeminent sport, a finding that aligns with its status as the nation's sole gold medalist and world record holder in this discipline. A similar observation can be made when examining the United States, another nation with a strong sporting reputation. The RCA index for the United States exceeds 1 in several sports, suggesting a comparable level of dominance in various sports. This finding is consistent with the notion that the United States possesses a diverse array of sports with similar levels of strength.

In order to ascertain the sports that are of the utmost importance to each nation, the present study utilizes Egypt and the United States as exemplars. The Tobit and Hurdle models are

employed to predict the total for the 2024 Summer Olympics, calculate the corresponding RCA and coefficient, use these as a benchmark to rank the sports, and extract the number of medals contributed by these sports. In essence, the sports displayed in the figure represent the most significant contributors to the overall medal count of the United States and Egypt, as evidenced by their substantial RCA and corresponding coefficient values.

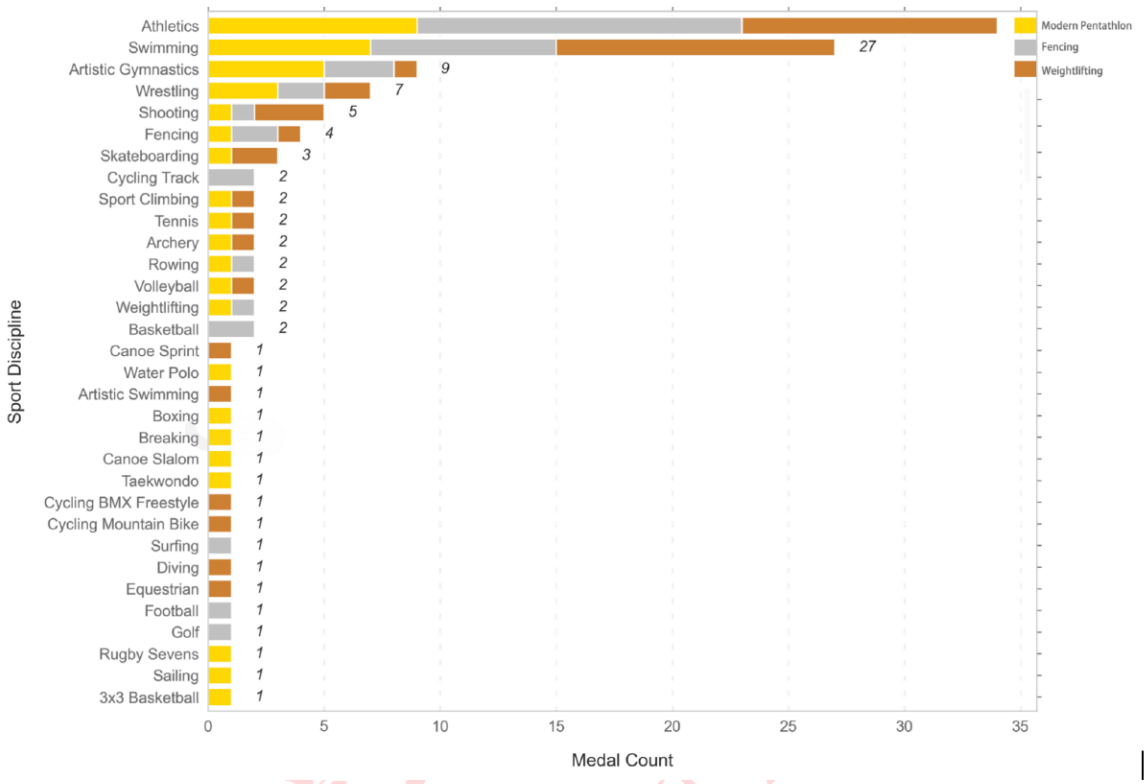

Figure 8 Top Contributing Sports to Medal Counts: the United States and Egypt"

5.5.3 Impact of Host Country Event Selection on Performance

Based on historical data observations, there is a significant synergistic growth between the host country's RCA and NAE (as shown in the figure below).

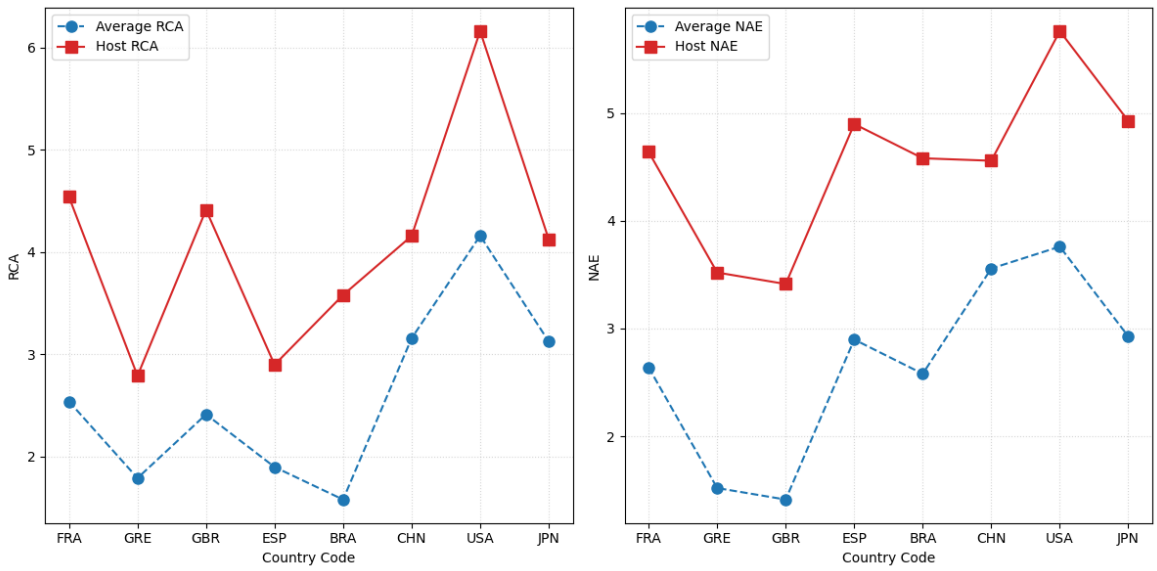

Figure 9 NAE and RCA Indices of the Host Country

This phenomenon may be attributed to two non-mutually exclusive mechanisms:

- Strategic project selection, wherein the host country utilizes the rules of the competition to incorporate temporary advantageous projects (such as niche unpopular projects) to artificially augment RCA through "institutional dividends".
- Systematic capacity enhancement. The long-term resource investment (e.g., athlete training) brought about by hosting the event promotes the overall upgrading of sports strength, which in turn enhances NAE and RCA.

To quantify the independent impact of these mechanisms, this study employs a dual strategy of model complementarity and variable reconstruction. While the hurdle model has been demonstrated to more effectively capture the host effect, the Tobit model has been shown to be more adept at addressing sports with weaker nations (i.e., medal count less than or equal to 0). This is due to the truncated regression characteristics of the Tobit model, which are more conducive to the "zero value accumulation + low value continuity" distribution. Consequently, the Tobit model can more sensitively detect the marginal effect of the host country surpassing the zero medal threshold through RCA_add and NAE. Consequently, it is imperative to construct two models based on Tobit and Hurdle through variable refinement.

- Convert continuous NAE to a three-category variable (low/medium/high) to identify nonlinear threshold effects accumulated empirically, and calculate the average RCA of traditional events that have existed stably in previous events to reflect the country's long-term competitiveness.
- Separate the RCA of permanent events and new events. For new event RCA_add: extract the historical maximum ($max(RCA)$) of the RCA of temporary events added by the host country to capture the peak effect of its strategic selection of events. maximum value ($max(RCA)$) to capture the peak effect of its strategic selection of events.

The ensuing results are presented in the subsequent table.

Table 8 Results of Four Explanatory Models

Variables	Model (1)–Hurdle		Model (1)–Tobit		Model (2)–Hurdle		Model (2)–Tobit	
	Coef	SD	Coef	SD	Coef	SD	Coef	SD
Host country in 4 years	0.136*	0.21	5.864***	0.12	0.236*	0.12	5.864***	0.22
Host country t	1.354***	0.41	5.520***	0.23	1.546***	0.21	5.520***	0.03
Host country 4 years ago	1.085***	0.21	4.145**	0.18	0.879**	0.11	4.547**	0.12
Athletes [0,10]	0.016*	0	0.745*	0.22	0.016*	0	0.789*	0.22
Athletes [10,50]	0.241*	0.34	0.826***	0.88	0.510*	0.12	0.726***	0.57
Athletes [50,150]	0.789**	0.54	0.924***	0.65	0.989**	0.23	0.954***	0.65
150 athletes and more	0.978***	0.25	9.314***	1.15	0.834***	0.12	0.614***	1.15
Medals_PCA(t-4)	3.725***	0.28	3.42***	4.57	3.702***	0.23	3.42***	4.47
RCA	1.548***	0.32	1.828**	1.42	2.538***	0.24	1.828**	1.17
RCA_usual					0.538*	0.22	0.778**	0.92
RCA_add					2.618***	0.41	1.994**	0.93
NAE					2.911***	0.122	2.414***	0.48
NAE [0,20]	1.745***	0.24	2.948**	1.15				
NAE [20,50]	1.857***	1.71	1.011**	0.75				
NAE 50 and more	1.912***	0.92	1.112***	0.36				
Num_Events	0.575**	0.24	1.148**	1.72	0.538*	0.22	1.178**	0.92
Num_sports	0.487**	1.69	1.572**	1.07	0.618**	0.41	0.994**	0.93
$g_{i,t}$	-3.408***	0.75	—	—	-3.408***	0.235	—	—
σ_u^2	0.350***	0.05	26.68***	23.67	0.334***	0.03	51.907***	8.36
Observations total	554	—	1289	—	554	—	1289	—
Observations noncensored	554				554			

Note: ***Significant at the 1 percent level; **5 percent level; *10 percent level..

According to the results in the table, strategically adding temporary dominant events (RCA_{add}) is the core driver of host countries' medal gains. In Model 2, the coefficient of RCA_{add} is significantly positive (2.618, 1.994), indicating that host countries can significantly increase their medal gains in the short term by adding new events with the best historical performance (such as karate at the Tokyo Olympics) and focusing their resources on these events. In contrast, the impact of long-term competitiveness (RCA_{usual}) is relatively limited, and the significance is not prominent in Model 1 and Model 2. Secondly, for Tobit, the number and type

of events (Num_Events, Num_Sports) are significantly positive (Num_Sports coefficient 1.572 in Model 1), indicating that adding sub-events or new sports types can create more medal opportunities for weak sports countries. In addition, Model 1 shows that experience accumulation (NAE categorical variable) and breaking through the zero medal threshold as a host country for weak sports countries are particularly critical through the addition of events (NAE low group coefficient 2.948*); while the host effect ($Host \quad country \quad t$ coefficient 5.520***) is stronger in the Tobit model, which confirms its marginal effect on the breakthrough of medals for weak countries. In summary, the host country achieves structural growth in the number of medals through a three-pronged approach to event selection: "short-term focus on strengths + expansion of event scale + preferential treatment of home resources". This mechanism is of even greater decisive significance for countries with low competitiveness.

6 The Great Coach Effect: A Strategy Optimization Study Based on Bayesian Change-Point Detection and Causal Inference Models

6.1 Discovery of Great Coach Effect Based on Bayesian Change-Point Detection

When there is a significant change in the number of medals won by a country in a certain event in a certain edition, e.g. winning at least one medal in a certain edition but failing to win a medal in several consecutive Olympic Games, or a significant increase in the number of medals won but also winning gold medals in previous editions. We believe that such significant changes are the result of the "great coach effect". The reason for this idea is that a sudden breakthrough after a long period of downturn, or a significant increase in strength in a certain edition of the Games, may be related to an outstanding coach, even if the country had previously had a certain level of sporting strength. Such obvious turning points can be identified using a Bayesian change-point detection.

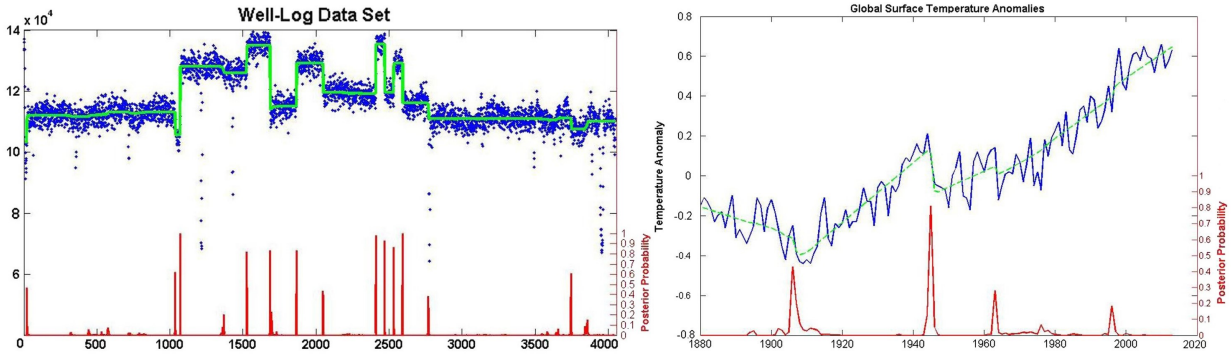

Figure 10 Two Types of Change-Point

6.1.1 Data Handling

Bayesian change-point detection has a wide range of applications in fields such as medicine, such as real-time detection of sudden changes in a patient's heart rate. Or the time point at which significant changes occur in historical data. For medal data, however, if bronze, gold, and silver medals are respectively set to 1, 2, and 3, the fluctuations in a unit of time are too

large, causing the model to identify a large number of change points. Therefore, we process the medal data to make it more stable within a unit of time. The calculation uses a sliding window method, and the formula is as follows:

$$S_t = \left(1 + \frac{1}{N_{Total}}\right) \cdot \begin{bmatrix} a \\ b \\ c \end{bmatrix} \cdot \begin{bmatrix} N_{Gold} & N_{Silver} & N_{Bronze} \end{bmatrix} \tag{9}$$

$$Score_t = \sum_{i=0}^{K} S_{t-4,i} e^{-i} \tag{10}$$

Before and after data processing:

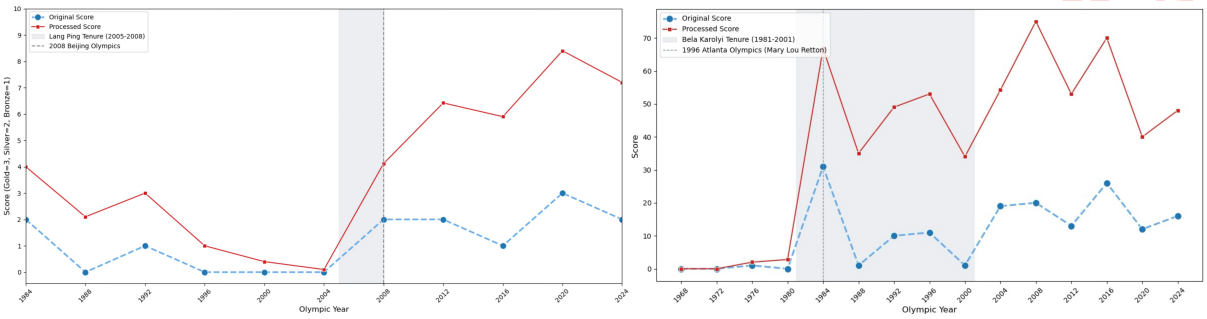

Figure 11 Two Types of Change-Point

6.1.2 Bayesian Change-point Detection

The fundamental principle of Bayesian change-point detection involves the segmentation of data into distinct components, under the assumption that each segment follows a specific probability distribution (e.g., the normal distribution). This approach entails the estimation of the location and number of change points through the application of Bayesian inference. This method integrates the observed data, the parameter prior, and the change point prior through a joint posterior distribution, leveraging marginal likelihood to streamline the calculation. This facilitates the inference of the most probable change point configuration.

6.1.3 Evidence of Coaching Effect in the United States

By bringing in the data of the US volleyball team and the gymnastics team and using Bayesian Change-point Detection, we found that it detected the change point, i.e., great coaching effect, as shown in the table below.

Table 9 Evidence of Great Coaching Effect for the United States

Sports	Change-point	Posterior Probability	Data Segment	Variance	Confidence Interval (95%)
Volleyball	2008	0.95	2004-2008	1.3	[4.8, 5.6]
Gymnastics	1984	0.87	1980-1984	2.1	[7.3, 8.3]

6.2 Quantitative Evaluation of Great Coaching Effect on Medal Counts

Regression models are an important tool in statistics for estimating the relationship between variables. They are significantly superior to machine learning models such as XGBoost and random forests in terms of interpretability (cited literature). However, when using regression models to analyze the contribution of great coaches to the number of medals, it is impossible to avoid the estimation bias caused by endogeneity (i.e., the influence of the mutual relationship between variables, missing variables, etc.), which makes it impossible to accurately identify the true influence of coaches. Therefore, here we must abandon general parameter estimation methods (such as OLS, DID, and IV) and develop a parameter estimation method that eliminates endogeneity and focuses on the strong causal effect of coaches and medals, namely two-stage least squares(2SLS).

6.2.1 Regression Model Using 2SLS

Step 1: Regress Endogenous Variable D Using Instrumental Variable $Score_t$

The endogenous variable D indicates whether or not the coach is great. If $D = 1$, it indicates that the coach has a great effect; otherwise, $D = 0$. The expression that defines D is:

$$D = \alpha_0 + \partial_1 Score + \alpha_2 RCA + \alpha_3 Num\_Events + \alpha_4 Num\_Athlete + \alpha_5 NAE + m \quad (11)$$

where $Score$ is the instrumental variable strongly correlated and satisfying exogeneity (not directly affecting the existence of a great coach). The remaining explanatory variables are exogenous variables.

Step 2: Regress the Number of Medals N Using the Predicted Value D

The number of medals N can be expressed as

$$N = \beta_0 + \beta_1 D + \beta_2 RCA + \beta_3 Num\_Events + \beta_4 Num\_Athlete + \beta_5 NAE + n \quad (12)$$

where $N = [N_{gold}, N_{silver}, N_{Bronze}]^T$, β_0, β_1, β_2, β_3, β_4, $\beta_5 \in \mathbb{R}^3$ are coefficient.

If β_1 is positive, the three components of β_1 represent the contributions of great coaching effect to the gold, silver and bronze medals, respectively.

6.2.2 Model Validation

We use the United States as an example and use historical data from 1960 to 2024. First, Bayesian change-point detection is used for each sport to statistically determine the possible event of the great coaching effect. Then, MLE regression is performed on each sport using 2SLS and OLS in turn to statistically determine the impact of 2SLS and OLS on the number of medals won by great coaches for each event and compare the results with the actual values, as shown in the figure.

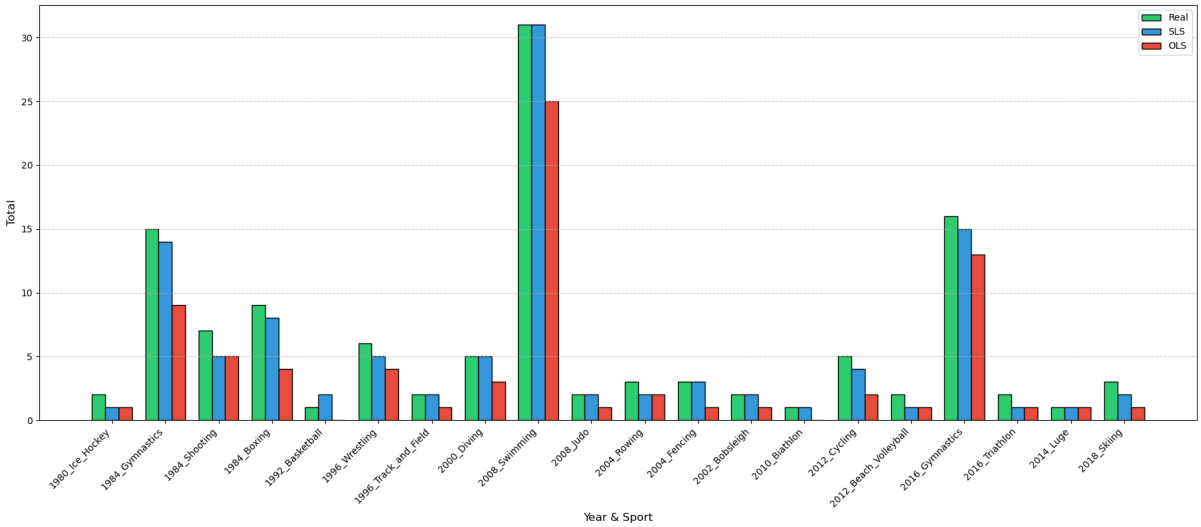

Figure 12 Comparison of SLS, OLS Models vs Real Data

Both 2SLS and OLS are explanatory of the impact of great coaches on the number of medals, but 2SLS is significantly better than OLS in terms of explanatory effect, and the problem of endogeneity is better solved.

6.3 Targeted Coaching Investments for Three Nations

The statistical results show that the United States, the United Kingdom, and France can obtain marginal benefits through strategic investment in coaches. The expected increase in the number of medals for each country after the introduction of coaches is shown in the figure below.

Table 10 The Targeted Coaching Investments and Predicted Improvement Magnitude in the United States, the Great Britain, and France

The United States	Magnitude	Great Britain	Magnitude	France	Magnitude
Wrestling	2	Wrestling	1	Table tennis	1
Rowing	1	Canoeing	2	Badminton	1
Judo	1	Judo	1	Athletics	1
Archery	1	Shooting	1	Gymnastics	2
Triathlon	2	Weightlifting	1	Rowing	1

The above strategy is based on two core logics: first, selecting projects with outstanding marginal benefits with the intervention of existing coaches to achieve medal growth at minimal cost; second, targeting weaknesses in the international competitive landscape (e.g., table tennis is dominated by Asia) and filling technical gaps through cross-border coach cooperation. In the long term, such investment can improve short-term results and build sustainable competitiveness, especially in Olympic events with a highly concentrated distribution of medals.

7 Insights: Strategic Resource Allocation and Path Dependency in Olympic Medal Performance

For question 3, we continue the analysis of questions 1 and 2, and use Bayesian variable point detection to label the possible great coaching effects of various sports in each country in previous Olympic Games, and add them as explanatory variables to the Hurdle and Tobit models. At the same time, due to the self-reinforcing mechanism of medal dynamics, we treat the explanatory variables in the same way (i.e., (further explaining the lagged terms), and then regress them. The results are shown in the table below.

Table 11 Results of Explanatory Models

Variables	Hurdle Model		Tobit Model	
	Coef	SD	Coef	SD
Host country t	1.354***	0.41	5.520***	0.23
Host country 4 years ago	1.085***	0.21	4.145***	0.18
Athletes [150+]	0.978***	0.25	2.314***	1.15
Medals_PCA	1.548***	0.32	1.828***	1.42
Num_Events vs. Num_Sports	0.575** vs. 0.487**	0.19	1.148*** vs. 1.572***	1.03

*Note: Only display the variables to be analyzed to interpret the coefficients; ***Significant at the 1 percent level; **5 percent level; *10 percent level.*

7.1 Dynamic Attenuation of the Host Country Effect and Resource Redistribution

From a temporal perspective, in the short term, the coefficient of host country t is 1.354* (Hurdle model), indicating that the advantage of the host country is most significant during the current Olympic Games. In the long term, the coefficient of host country 4 years ago is 1.085* (Hurdle model), indicating that the host country effect still exists 4 years later, but the strength has decreased by about 20%.

Policy implications:

- The host country needs to transform the infrastructure into a long-term training base within 2-4 years after the event to avoid idle resources.
- Non-host countries can target the "post-Olympic trough period" of previous host countries and obtain their surplus resources (such as venue leasing and coach exchanges) through cooperation.

7.2 The "Double-Eged Sword Effect" of the Number of Events and Sports

The coefficient of the Num_Events is 0.575* (Hurdle model)**, which shows that increasing the number of events can improve medal opportunities. However, the coefficient of the Num_Sports is only 0.487* (Hurdle model)**, indicating that excessive diversification may dilute resources.

Policy implications: National Olympic committees need to balance "breadth vs. depth".

- Small countries: choose 3-5 high-potential sports to avoid fighting on multiple fronts.
- Large countries: leverage the advantages of multiple sports, but need to establish a "core-peripheral" echelon (e.g., track and field as the core, modern pentathlon as a supplement).

8 Sensitivity Analysis

When using the hurdle model, a binary classifier is required in the first stage to classify which countries can win at least one medal. We use a logit classifier. Therefore, we will conduct a sensitivity analysis on the selection of the binary classifier to explore whether the selection of the binary classifier will affect the interpretability of the hurdle model for the number of medals and thus affect its predictive performance. We use the random forest classifier, decision tree, and logistic regression to replace the logit classifier in the hurdle, use the output probability as the result of the hurdle stage 1, regress it, and use the logit classifier as the benchmark to compare the changes in the coefficients of each covariate in the original model.

Variables	Random Forest Classifier	Decision Tree	Logistic Regression
Host country in 4 years	0.052	0.022	-0.057
Host country t	-0.011	0.025	-0.088
Host country 4 years ago	0.062	-0.013	0.565
Athletes [0,10]	-0.122	0.051	-0.041
Athletes [10,50]	0.011	0.052	0.251
Athletes [50,150]	0.062	-0.045	0.051
150 athletes and more	0.032	0.000	0.152
Medals_PCA(t-4)	-0.052	0.052	0.142
RCA	0.012	0.069	-0.098
Medals_PCA(t-4)	0.012	-0.052	-0.095
Num_Events	0.012	0.021	0.054
Num_sports	0.041	0.065	0.013
NAE	-0.032	0.054	-0.064
$g_{i,t}$	0.055	-0.048	0.052
σ_u^2	-0.011	0.054	-0.098
Observations total	0.072	-0.067	-0.052
Observations noncensored	-0.065	0.002	0.082
	0.025	0.051	-0.054

As can be seen from the table, the error between all the coefficients of the covariates and the explanatory variables is less than 10%, indicating that the classifier model we selected has no significant impact on the performance of the hurdle model, and our model is relatively robust.

9 Model Evaluation

9.1 Strength

- The interpretability and performance of the Hurdle and Topit models are both strong, rather than simple machine learning algorithms.
- Before using Bayesian variable point detection, the data is first preprocessed to reflect the significance of medals and the dynamic self-reinforcement mechanism.

- When explaining the great coaching effect, the use of 2SLS eliminates endogeneity between variables.

9.2 Weakness

- Due to the limitations of the subject, the model does not introduce the influence of factors such as economics and humanities, which may lead to errors.

References

[1] De Bosscher, V., de Knop, P., van Bottenburg, M., & Shibli, S. (2006). A conceptual framework for analysing sports policy factors leading to international sporting success. European Sport Management Quarterly, 6(2), 185-215.

[2] Streicher, T., Schmidt, S. L., Schreyer, D., & Torgler, B. (2020). Anticipated feelings and support for public mega projects: Hosting the Olympic Games. Technological Forecasting and Social Change, 158, 120158.

[3] Ball, D. W. (1972). Olympic games competition: Structural correlates of national success. International Journal of Comparative Sociology, 13(2), 186-200.

[4] Bainbridge, M. (1998). Outcome uncertainty in sporting competition: The Olympic Games 1896–1996. Applied Economics Letters, 5(3), 161-164.

[5] Condon, E. M., Golden, B. L., & Wasil, E. A. (1999). Predicting the success of nations at the summer Olympics using neural networks. Computers & Operations Research, 26(13), 1243-1265.

[6] Kuper, G. H., & Sterken, E. (2001). Olympic participation and performance since 1896. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.274295

[7] Lui, H.-K., & Suen, W. (2008). Men, money, and medals: An econometric analysis of the Olympic Games. Pacific Economic Review, 13(1), 1-16.

[8] Leeds, E. M., & Leeds, M. A. (2012). Gold, silver, and bronze: Determining national success in men's and women's Summer Olympic events. Jahrbücher für Nationalökonomie und Statistik, 232(3), 279-292.

[9] Blais-Morisset, P., Boucher, N., & Fortin, B. (2017). The impact of public investment in sports on the Olympic medals. Revue économique, 68(4), 623-642.

[10] Groll, A., Ley, C., Schauberger, G., & van Eetvelde, H. (2019). A hybrid random forest to predict soccer matches in international tournaments. Journal of Quantitative Analysis in Sports, 15(4), 271-287.

[11] Lessmann, S., Sung, M.-C., & Johnson, J. E. (2010). Alternative methods of predicting competitive events: An application in horserace betting markets. International Journal of Forecasting, 26(3), 518-536.

[12] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[13] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM. https://doi.org/10.1145/2939672.2939785

[14] Tcha, M., & Pershin, V. (2003). Reconsidering performance at the summer Olympics and revealed comparative advantages. Journal of Sports Economics, 4(3), 216-239.

[15] Forrest, D., McHale, I. G., Sanz, I., & Tena, J. D. (2015). Determinants of national medal totals at the summer Olympic games: An analysis disaggregated by sport. In The Economics of Competitive Sports. Edward Elgar Publishing.

[16] Rewilak, J. (2021). The (non) determinants of Olympic success. Journal of Sports Economics, 22(7), 152700252199283.

[17] Scelles, N., Andreff, W., Bonnal, L., Andreff, M., & Favard, P. (2020). Forecasting national medal totals at the summer Olympic games reconsidered. Social Science Quarterly, 101(2), 697-711.

[18] Streicher, T., Schmidt, S. L., Schreyer, D., & Torgler, B. (2020). Anticipated feelings and support for public mega projects: Hosting the Olympic Games. Technological Forecasting and Social Change, 158, 120158.

[19] Scelles, N. , Andreff, W. , Bonnal, L. , Andreff, M. ,& Favard, P. . (2020). Forecasting national medal totals at the summer olympic games reconsidered. Social Science Quarterly, 101(2).