Problem Chosen	2024 MCM/ICM Summary Sheet	Team Control Number
C		2518812

In this report, we study the factors and nuances that affect a country's performance in the Olympics. We provide our solution to three questions: 1) How does one predict the gold medal list and total medals list in the 2028 Los Angeles Olympics; 2) How does coaching affect a country's performance in the Olympics; and 3) what are some insights the Olympics data could unravel, with an emphasis on the first question.

We employ a divide-and-conquer approach to forecasting the medal lists of the 2028 Olympics. Specifically, we categorized all Olympics Sports into three classes: "new" sports, which have participated in the Olympics for limited times; "stable" sports, whose winning countries do not fluctuate dramatically over time; and "unstable" sports, in which countries on the podium change swiftly. We then devised a Bayesian-Autoregressive model to describe patterns in the "stable" sports' medals, a non-parametric functional autoregressive model with multivariate exogenous variables to explain the "unstable" sports, and an autoregressive model with exogenous variables to model the "new" sports. The different classes of Sports are ultimately combined into a predicted gold medals list and a predicted total medals list, in which the 95% confidence interval of each country's medal counts is also provided. Additionally, using a Bayesian model, we also examine the possibility that countries who have never won an Olympic medal prior to the 2028 Olympics could make a breakthrough.

Contents

1 Introduction

1.1 Problem Background

The Olympic Games celebrate athletic excellence, with medal counts, especially Gold medals, symbolizing national pride. The 2024 Paris Olympics saw the United States leading in total medals (126) and tying with China for the most Golds (40). France performed well as the host, ranking 4th in total medals, while smaller nations like Dominica and Saint Lucia won their first Golds. However, over 60 countries still lack any Olympic medals, highlighting global disparities.

The host nation's role is crucial, as event selection and athlete investment often boost performance. The composition of sports also affects medal counts, with the U.S. excelling in swimming and track, while China dominates table tennis and weightlifting. These factors contribute to the unique nature of the Olympics, where success depends on various dynamics and the growing diversity of podium nations.

1.2 Problem Restatement

Q1: Prediction of medal counts for each country: Develop a model to predict the medal counts for each country, focusing on Gold and total medals, including estimates of uncertainty and performance metrics. Use this model to forecast the 2028 Los Angeles Olympics medal rankings, with prediction intervals for all countries. Identify which countries are likely to improve or perform worse compared to 2024. Include countries that have not yet earned medals, estimating how many will win their first medal in 2028. Analyze how the number and types of events impact medal performance, identifying key sports for each country and considering how the host country's event selection influences results.
Q2: "Great Coach" effect: Athletes face challenges in changing countries due to citizenship rules, but coaches can easily switch countries. This creates the potential for a "great coach" effect. Analyze the data for evidence of a "great coach" effect on medal counts. Estimate how much this factor contributes and suggest three countries and sports where investing in a "great" coach could have an impact.
Q3: Other insights about Olympic medal counts: Identify other original insights and explain how these insights can inform country Olympic committees.

2 A Divide-and-Conquer Approach

Traditionally, the task of forecasting Olympic medals is treated in a macroscopic sense. Many scholars solely use non-sport information, such as Gross National Product, to predict a nation's performance in the Olympics. (see Schlembach(2022) for example)

Alternatively, we adopt a divide-and-conquer path to the problem. Specifically, we divide all Olympics Sports into three classes ("stable" Sports, "unstable" Sports, and "new" Sports), devise independent models for each class, and combine our findings to answer the questions stated in Section 1.2. The classification criteria is defined as follows.

For a Sport z, denote by $p_{i,t}^{(z)}$ the proportion of gold medals of that Sport country i won at the t-th Olympics. We henceforth refer to this quantity as *gold proportions*. As an illustration, $p_{1,5}^{(10)}$ measures

$$\frac{\text{Number of Sport 10 golds won by country 1 at the fifth Olympics}}{\text{Total number of discipline number 10 golds at the fifth Olympics}}. \tag{1}$$

A Sport z is determined to be "new" if one of the following conditions is met.

1. If Sport z has been in the Olympics for fewer than 5 times, which we call a "existing new" Sport.

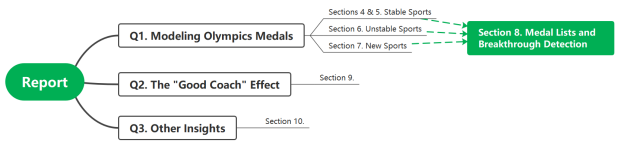

Figure 1: The workflow chart of our report.

2. If Sport z is to debut or return at the next Olympics, which we call a ”premiering new” Sport.

For non-”new” Sports, it is classified as ”stable” if one of the two conditions is met:

1. Recent dominance. If $p_{i,t-1}^{(z)}, p_{i,t-2}^{(z)}, p_{i,t-3}^{(z)} \geq 60\%$ for some country i.

2. Low gold turnover rate. We define the turnover rate of Sport z at the t-th Olympics by

$$TO_t^{(z)} = \sum_{i=1}^{n} |p_{i,t}^{(z)} - p_{i,t-1}^{(z)}|.$$

We calculate the weighted turnover rate for all disciplines over the past 5 Olympics, using exponentially-decaying weights with a 6-year half-life:

$$TO^{(z)} = 0.411TO_t^{(z)} + 0.259TO_{t-1}^{(z)} + 0.163TO_{t-2}^{(z)} + 0.103TO_{t-3}^{(z)} + 0.065TO_{t-4}^{(z)}$$

The median turnover rate is used as the cutoff value: Sports with a turnover rate lower than the median value are considered ”stable”.

The workflow of our report is shown in Figure 1. Ultimately, our findings are summarized in the summary sheet.

3 Data Analysis and Pre-processing

The link to the code: **Github Repository**. **Python** was the main language used in this process.

3.1 Overall cleaning process

3.1.1 data_cleaning.py

This file is the most crucial part of the data cleaning process, as it will standardize the **country names** and **discipline names** to make the dataset more consistent. As there were too many political

events in the past which might change a country's name appearing in the Olympic games, we chose to focus on some of the most famous events and work on renaming of the countries. The inheritance relationship was predominantly determined by population demographics. These are:

- **Tag der Deutschen Einheit (German Unity Day)**: We decided to rename all data before 1992 associated with West and East Germany to Germany.
- **The dissolution of the Soviet Union**: Russia will directly inherit all medals achieved by Soviet Union, while other successor states are treated as new participants.
- **The breakup of Yugoslavia**: Serbia will directly inherit all Yugoslavia's medals, while other successor states are treated as new participants.
- **Velvet Divorce**: Czech Republic directly inherits all Czechoslovakia's medals, while Slovakia is treated as a new participant.
- **Another special case**: some countries in the raw dataset have a hyphen in their names. This indicates that the event is a team match that a country might send out more than one teams to. All teams with this kind of format would be merged to the country itself (the data of **Germany-1** and **Germany-2** was merged to the data of **Germany**).

Another important part is the names of the disciplines. Note that we dealt with disciplines instead of specific events, as the name of events varied too much from year to year. The provided **Athletes** dataset contains multiple disciplines with different names, but they are actually the same ones. For example, **Marathon Swimming** and **Marathon Swimming, Swimming** should be the same discipline, but they appear to have different names. After examining the data, we decided to focus on some of the disciplines that may greatly affect our modeling:

- **Canoeing, Cycling, Swimming**: all Canoeing, Cycling, Swimming-related disciplines were merged together into **Canoeing, Cycling, and Swimming**, respectively.
- **Equestrianism**: Equestrianism and Equestrian were merged to be **Equestrianism**.
- **Gymnastics**: all Gymnastics-related disciplines were merged into **Gymnastics**, along with the special case Trampolining.

3.1.2 less5/great5_classification.py

The effect of the model depends on the quality of the data. The data should be classified into one of the 3 categories defined in **2. Classification Criterion**. Our first criteria is that a sport that has been in the Olympics for fewer than or exactly 5 times (denoted as **lt5**) is classified as unstable, and vice versa. For a more convenient modeling environment, we decided to make **all our dataframes follow the format** with following information:

- **Sport**: the name of the discipline, stored as a string.
- **Year**: the year of the Olympic game, store as an integer.
- **Team**: the country/team (here they are the same in the meaning), stored as a string.
- **gold/silver/bronze_medals**: integers representing the number of gold/silver/bronze medals a country get in this discipline and year.
- **total_medals**: this column contains an integer which is simply the sum of the previous 3 columns.
- **participants**: the number of participants from this country, stored as an integer.
- **sport_total_medals**: the total number of medals of this discipline in this year, stored as an integer.
- **sport_total_participants**: the total number of participants in this discipline and year, stored as an integer.
- **isHost**: a boolean value representing whether the country is the host of the Olympic or not.

There were many interesting edge cases that were ignored at first. For example, in a team event (like **3×3 Basketball**), we should only count gold medal 1 time for the champion instead of 3.

3.1.3 stability_classification.py

Disciplines that have been held for more than 5 times (denoted as **gt5**) in the Olympics history were stored in ***./processed/gt5_sports.csv***. If a discipline is **dominated by some countries**, or has a **low gold turnover rate**, it is considered to be a **stable discipline**. All **gt5** were classified to either "dominant" or "non-dominant", indicating whether there is a country that achieved more than 60% of gold medals

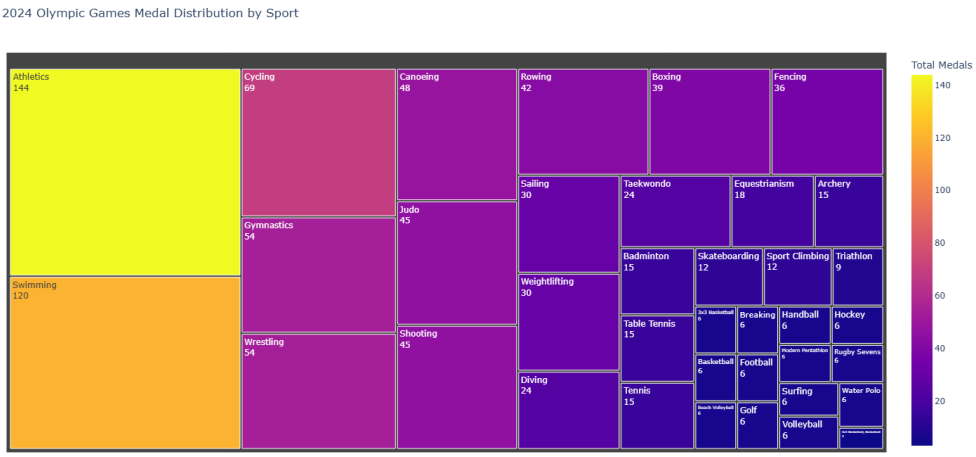

Figure 2: Total medal distribution among disciplines in 2024

in the most recent 3 Olympic games (2016, 2020, and 2024).

For those ”non-dominant” disciplines, we further computed their gold turnover rates and split them in half, where the half with **higher turnover rates were considered unstable**, and the other ones were considered **stable and merged with ”dominant” data**. Thus we have our data divided into 3 parts: **lt5**, **stable**, and **unstable**.

3.1.4 visualization.py

As the name indicates, this program provides an auxiliary tool for visualizing some of the data which makes our analysis more straightforward. There are 2 diagrams produced:

- **Treemap diagram of medal distribution**: we computed the overall distribution of total number of medals among disciplines in 2024 (**Figure 1**).
- **Line plot of gold/total medals of top 15 countries in 2024**: this is a line plot depicting the trend of the number of medals (gold and total) throughout the history of Olympic games achieved by the 15 countries that ranked top in the 2024 Olympic (**Figure 2 and 3**).
- **Turnover vs Gold medals of disciplines in 2024**: we wanted to figure out whether there's some relationships between the gold turnover with actual number of gold medals that one discipline has in 2024. **Min-Max normalization** was used to make sure every turnover value lies in the

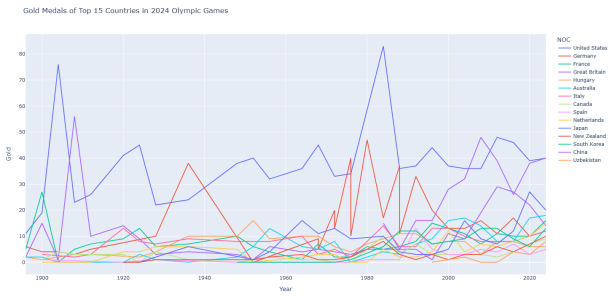

Figure 3: Gold medal trend of top 15 countries

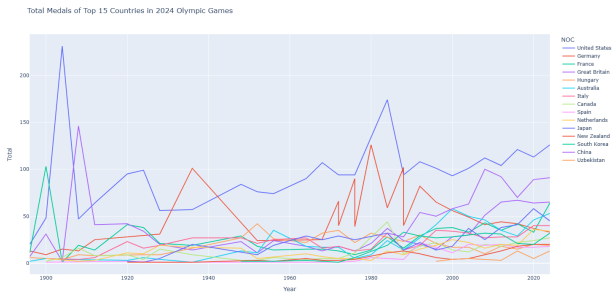

Figure 4: Total medal trend of top 15 countries

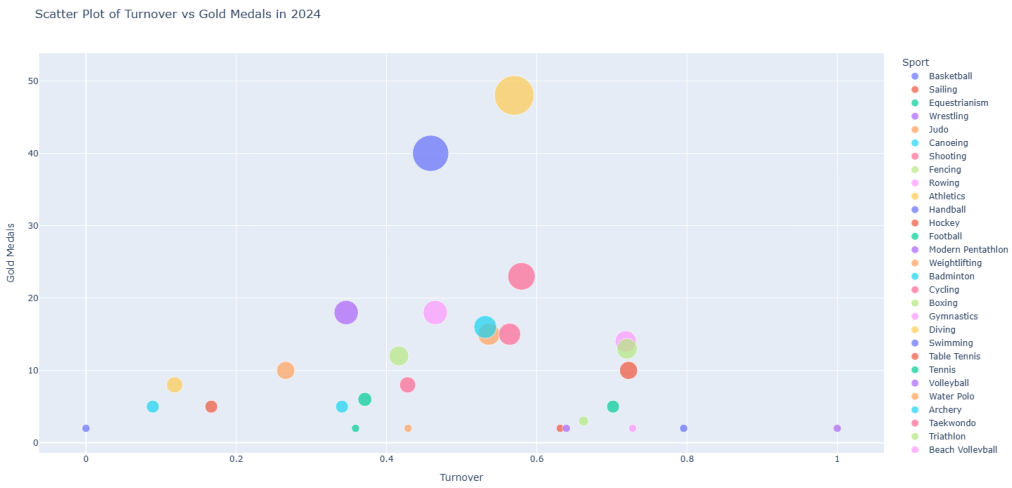

Figure 5: Turnover vs Gold medals in 2024

range of $[0, 1]$, and the sizes of dots were set to be proportional to the number of gold medals (**Figure 4**).

Properties of Historical Medal Data

Historical Olympic medal data (from 1896 to 2024) is characterized by:

- **Fragmented participation**: Many teams do not compete every cycle or in every sport, leading to sporadic appearances in the dataset.
- **Many zeros**: Even when participating, a large number of team–year combinations yield zero

medals (especially smaller or newer countries).

- **Momentum effects**: Performances in past Olympics often correlate with future success (teams that invest more, or due to latent socioeconomic factor not available in problem setting).

These data features motivate two different approaches when predicting overall medal tables versus gold-medal tables:

1. A **VAR** model that captures cross-lag dependencies among gold, silver, and bronze medal counts, along with partial pooling across teams (but *no explicit presence factor* in this example).

2. A **Bayesian AR(3) on gold ratio** for each sport, focusing on how the fraction of gold medals a team captures evolves over time (here, `present` *is* feasible to include, since per-sport participation is straightforward to track).

Table 1: Symbol List

Symbol	Meaning
i	Index for team (country).
t	Index for time (Olympic cycle, e.g. 1896, 1900, . . . , 2024).
$\text{gold}_{i,t}$	Number of gold medals won by team i in year t.
$\text{silver}_{i,t}$	Number of silver medals won by team i in year t.
$\text{bronze}_{i,t}$	Number of bronze medals won by team i in year t.
$\texttt{present}_{i,t}$	Binary indicator: 1 if team i participated in that sport/year.
$\mathbf{y}_{i,t}$	Vector $\left(\text{gold}_{i,t}, \text{silver}_{i,t}, \text{bronze}_{i,t}\right)^\top$.
$\text{Host}_{i,t}$	Binary indicator: 1 if team i was the host country in year t.
Σ	Error covariance matrix in the VAR.
$r_{i,t}$	Fraction (ratio) of gold medals for team i in a specific sport/year.
$\text{logit}(x)$	The logistic transform, $\ln\big(x/(1-x)\big)$.

4 Stable Program medal list Prediction

4.1 Bayesian AR(3) Model with Presence Factor (Negative Binomial)

4.1.1 Definition of the Presence Factor

We define a binary variable $\texttt{present}_{i,t}$ to capture whether team i competed in year t:

$$\texttt{present}_{i,t} = \begin{cases} 1, & \text{if the team } i \text{ has nonzero participants in year } t, \\ 0, & \text{otherwise.} \end{cases}$$

The historical presence (or absence) of a team can reflect investment in sports infrastructure or consistent engagement. In practice, one may also define lagged versions such as $\texttt{present}_{i,t-4}$ to capture the effect of presence four years prior (common in Olympic data).

4.2 AR(3) Negative Binomial Formulation

Let $Y_{i,t}$ denote the total medals for team i at time t. We model $Y_{i,t}$ using a Negative Binomial distribution with mean $\mu_{i,t}$ and an overdispersion parameter ϕ:

$$Y_{i,t} \sim \text{NegBinomial}\big(\mu_{i,t},\ \phi\big),$$

where $\mu_{i,t}$ depends on past medal counts (AR(3) structure), past or current presence, and potential hosting status. Concretely, one can write:

$$\begin{aligned} \log\big(\mu_{i,t}\big) = \beta_0 & \\ & + \beta_1 \ln\big(1 + Y_{i,t-1}\big) \\ & + \beta_2 \ln\big(1 + Y_{i,t-2}\big) \\ & + \beta_3 \ln\big(1 + Y_{i,t-3}\big) \\ & + \gamma \ln\big(1 + \texttt{present}_{i,t-4}\big) \\ & + \delta \,\text{Host}_{i,t} \\ & + u_i. \end{aligned}$$

Here:

- β_0 is an overall intercept capturing the baseline log-mean medal count.

- $\beta_1,\ \beta_2,\ \beta_3$ reflect the autoregressive momentum from medals earned in the three prior cycles.
- γ measures how being present (e.g. four years ago) contributes to current performance.
- δ captures any "hosting effect" ($\text{Host}_{i,t} = 1$ if team i hosts in year t).
- u_i is a random intercept for team i, $u_i \sim \mathcal{N}(0,\ \sigma_u^2)$, allowing partial pooling across teams.
- ϕ is the dispersion parameter of the Negative Binomial distribution.

4.2.1 Prior Specification and Posterior Inference

We assign priors to all unknown parameters:

$$\begin{aligned} \beta_0,\ \beta_1,\ \beta_2,\ \beta_3,\ \gamma,\ \delta\ &\sim\ \mathcal{N}\big(0,\ \sigma_\beta^2\big), \\ u_i\ &\sim\ \mathcal{N}\big(0,\ \sigma_u^2\big), \\ \phi\ &\sim\ \text{Gamma}\big(\alpha,\ \kappa\big). \end{aligned}$$

Priors

- $\beta_0,\ \beta_1,\ \beta_2,\ \beta_3,\ \gamma,$ and δ are global regression coefficients that control the effects of past medal counts, presence, and hosting on $\log(\mu_{i,t})$. Assigning $\beta_j \sim \mathcal{N}(0,\ \sigma_\beta^2)$ implies a soft prior centered at zero with variance σ_β^2, indicating no strong preconception about effect direction or magnitude.
- u_i is a team-specific random intercept capturing unobserved heterogeneity across teams. The prior $u_i \sim \mathcal{N}(0,\ \sigma_u^2)$ partially pools each team's baseline level of performance toward a global mean.
- ϕ is the overdispersion parameter of the Negative Binomial distribution. The prior $\text{Gamma}(\alpha,\ \kappa)$ encodes moderate beliefs about its likely range, balancing zero-inflated data and how dispersed the counts can be.

Given observed data $\{Y_{i,t},\ \texttt{present}_{i,t},\ \text{Host}_{i,t}\}$ for each team-year, the posterior distribution of $\Theta = \{\beta_0, \beta_1, \ldots, u_i, \phi\}$ is

$$p(\Theta \mid \{Y_{i,t}\},\ \{\texttt{present}_{i,t}\},\ \{\text{Host}_{i,t}\})\ \propto\ \prod_{i,t} \underbrace{\text{NegBinomial}\big(Y_{i,t}\ \mid\ \mu_{i,t}(\Theta),\ \phi\big)}_{\text{likelihood}} \times p(\Theta),$$

where $\mu_{i,t}(\Theta)$ follows the log link described above. We estimate Θ via MCMC, producing posterior samples that capture both parameter uncertainty and overdispersion.

Table 2: Posterior Summaries for the Negative Binomial Model on `total_medals`

Parameter	**Mean**	**SD**	**2.5%**	**97.5%**	$\hat{R}$
Random Intercept (Team)					
σ_{u_i} (sd(Intercept))	2.34	0.17	2.04	2.70	1.01
Regression Coefficients					
β_0 (Intercept)	-4.02	0.18	-4.38	-3.68	1.01
β_1: $\log(\text{lag\_total\_1} + 1)$	0.59	0.04	0.51	0.67	1.00
β_2: $\log(\text{lag\_total\_2} + 1)$	0.28	0.04	0.19	0.36	1.00
β_3: $\log(\text{lag\_total\_3} + 1)$	0.04	0.05	-0.05	0.13	1.00
γ: $\log(\text{lag\_present\_1} + 1)$	1.02	0.12	0.79	1.25	1.00
δ (isHost)	1.69	0.28	1.16	2.27	1.00
Additional Distribution Parameter					
ϕ (shape)	0.68	0.04	0.61	0.76	1.00

Interpretation

- The random-effect standard deviation $\sigma_{u_i} \approx 2.34$ indicates noteworthy variation in baseline medal-winning ability across the all teams.
- The negative intercept ($\beta_0 \approx -4.02$) reflects a low base count for teams with minimal prior medals and not hosting.
- Each one-unit increase in $\log(1 + \text{lag\_total\_1})$ raises the expected $\log(\mu)$ of total medals by about 0.59, indicating strong momentum from past one Game. The second (β_2, β_3) past game also contribute but with smaller coefficients. While the third lag has very insignificant effect. Considering the career cycle of professionals, it is not surprising.

- $\gamma \approx 1.02$ shows that being present four years prior (`lag_present_1`) can boost medals. Longer-term participation effects exist.
- Hosting country($\delta \approx 1.69$) remains a sizable advantage.
- The shape parameter $\phi \approx 0.68$ suggests moderate overdispersion typical of real-world medal data beyond Poisson assumptions.

5 Stable Program Gold List Prediction

5.1 Bayesian AR(3) for Gold Ratio in Each Sport

Motivation: Each sport has a fixed number of gold medals, making the problem *zero-sum* across teams. Modeling the *ratio* of gold medals each team wins in a given sport allows direct interpretation of how the share of medals evolves over time, and `present` tracks whether a team participated in that specific sport helps which will help us answer later problem 3.

5.1.1 AR(3) on the Logit Scale

For a given sport, let

$$r_{i,t} = \frac{\text{gold}_{i,t}}{\text{TotalGoldInSport}_t},$$

$$\begin{aligned}\text{logit}(r_{i,t}) = \alpha_i &+ \beta_1 \,\text{logit}(r_{i,t-1}) \\ &+ \beta_2 \,\text{logit}(r_{i,t-2}) \\ &+ \beta_3 \,\text{logit}(r_{i,t-3}) \\ &+ \gamma \,(\text{Host}_{i,t}) \\ &+ \delta \,(\texttt{present}_{i,t-1}) \\ &+ \varepsilon_{i,t}.\end{aligned}$$

Here:

- α_i is a random intercept for team i.

- β_k are global AR(3) coefficients.
- γ and δ capture host and presence effects.
- $\varepsilon_{i,t} \sim \mathcal{N}(0, \sigma^2)$.

Prior and Posterior Sampling

We specify priors,

$$\alpha_i \sim \mathcal{N}\big(\alpha_0,\ \sigma_\alpha^2\big),$$

$$\beta_k,\ \gamma,\ \delta \sim \mathcal{N}\big(0,\ 10^2\big),$$

$$\sigma,\ \sigma_\alpha \sim \text{Half-Cauchy}(1).$$

The posterior distribution for all parameters (and latent intercepts α_i) given $\{r_{i,t}\}$ is then:

$$p\Big(\{\alpha_i\},\ \alpha_0,\ \sigma_\alpha, \{\beta_k\}, \gamma, \delta, \sigma \ \Big|\ \{r_{i,t}\}\Big)$$

$$\propto \prod_{i,t} \underbrace{p\Big(r_{i,t} \ \Big|\ \text{logit}(r_{i,t}) \ = \ \alpha_i \ + \ \sum_{k=1}^{3} \beta_k \,\text{logit}(r_{i,t-k}) \ + \ \gamma\,\text{Host}_{i,t} \ + \ \delta\,\texttt{present}_{i,t},\ \sigma\Big)}_{\text{likelihood}}$$

$$\times\ p(\{\alpha_i\},\ \alpha_0,\ \sigma_\alpha)\ p(\beta_k)\ p(\gamma)\ p(\delta)\ p(\sigma).$$

Sampling is done via MCMC. After convergence, each draw yields a different set of parameter values, allowing us to forecast the distribution of future gold ratios.

Normalization for Predicted Gold Counts

To forecast for 2028, we compute

$$\widetilde{r}_{i,2028}^{(s)} \ = \ \frac{1}{1 + \exp\Big(-\,\text{logit}(r_{i,2028}^{(s)})\Big)},$$

then multiply by the total gold medals in that sport:

$$\widetilde{G}_{i,2028}^{(s)} \ = \ \widetilde{r}_{i,2028}^{(s)} \times \text{NumGoldInSport}_{2028}.$$

6 Modeling Unstable Sports

In this section, we devise a model to forecast the number of golds and total medals each country wins in Sports classified as unstable. Our classification criteria imply a country's gold proportion in an unstable Sport s, $p_{i,t}^{(s)}$, should be inconsistent in recent Olympics. Many factors contribute to the rapid movements of gold proportions, including the Sport's close competition, rapid development, and frequent coaching changes. The complexity and difficulty of quantifying such factors render the preceding parametric Bayesian model inappropriate, as it is nearly impossible to assume rigorous functional forms and factors. Alternatively, we fabricate a non-parametric time series model to forecast a country's gold proportion for such unstable Sports.

6.1 Model Formulation

We first devise the model for gold medals. The autoregressive (AR) model, a fundamental time series template, assumes the following:

$$p_{i,t}^{(s)} = \alpha_0 + \alpha_1 p_{i,t-1}^{(s)} + \cdots + \alpha_q p_{i,t-q}^{(s)} + \epsilon_t.$$

for some lag $q \in \mathbb{Z}_+$, autoregressive coefficients $a_0, \cdots, a_q \in \mathbb{R}$, and error term ϵ_t independent of previous observations $p_{i,t-j}^{(s)}, j \geq 1$ and having a finite second moment. By defining the lag operator $L : L^k p_{i,t}^{(s)} = p_{i,t-k}^{(s)}$, the AR model is compactly $\alpha(L)p_{i,t} = \epsilon_t$ where $\alpha(\cdot)$ is a polynomial.

We propose four improvements to the model:

1. Modeling the difference of gold proportions, $\Delta p_{i,t}^{(s)} := p_{i,t}^{(s)} - p_{i,t-1}^{(s)}$, is preferred as it contains information about the recent trend of country i's gold proportions in Sport s.

2. $q \leq 3$ is bounded because the information from 4 or more Olympics ago is intuitively far less significant than that from the recent three.

3. Autoregressive coefficients should be sport-specific to include the effects of disparate career lengths, competitiveness, and other factors. We hence replace a_i with $a_i^{(s)}, i = 0, \cdots, q$.

4. Exogenous variables could augment the explanatory power of the model. Particularly, we add a linear component to the model including three variables

- The host indicator, defined as

$$h_{i,t} = \begin{cases} 1 & \text{if country } i \text{ is the Olympics host at year } t; \\ -1 & \text{if country } i \text{ is the host at the previous Olympics;} \\ 0 & \text{otherwise.} \end{cases}$$

- The *participant proportion* at the previous Olympics, defined similarly as (1) by

$$n_{i,t}^{(s)} = \frac{\text{Number of Sport } s \text{ participants from country } i \text{ at the } t\text{-th Olympics}}{\text{Total number of Sport } s \text{ participants at the } t\text{-th Olympics}}.$$

- The *total medal proportion* at the previous Olympics, similarly

$$m_{i,t}^{(s)} = \frac{\text{Total number of Sport } s \text{ medals country } i \text{ at the } t\text{-th Olympics}}{\text{Total number of Sport } s \text{ medals at the } t\text{-th Olympics}}.$$

We combine the exogenous variables in $X_{i,t}^{(s)} = [h_{i,t} \quad n_{i,t-1}^{(s)} \quad m_{i,t-1}^{(s)}]^T$.

The model is thus an autoregressive model with exogenous variables (ARX):

$$\alpha^{(s)}(L)(\Delta p_{i,t}^{(s)} - \beta^T X_{i,t}^{(s)}) = \epsilon_t. \tag{2}$$

for some linear trend coefficient $\beta \in \mathbb{R}^3$ and some cubic polynomial $\alpha^{(s)}(\cdot)$. The model allows for a one-step-ahead prediction

$$\hat{\Delta p}_{i,t+1}^{(s)} = \hat{\beta}^T X_{i,t+1}^{(s)} + \hat{\alpha_0}^{(s)} + \hat{\alpha_1}^{(s)}(\Delta p_{i,t}^{(s)} - \hat{\beta}^T X_{i,t}^{(s)}) + \hat{\alpha_2}^{(s)}(\Delta p_{i,t-1}^{(s)} - \hat{\beta}^T X_{i,t-1}^{(s)}) + \hat{\alpha_3}^{(s)}(\Delta p_{i,t-2}^{(s)} - \hat{\beta}^T X_{i,t-2}^{(s)})$$

The model provides information about the correlation between the $p_{i,t}^{(s)}$ and selected factors. However, we question the suitability of constant autoregressive coefficients in the AR model. A crucial weakness of constant coefficients is that, for sports with few events, the random variable $p_{i,t}^{(s)}$ is drawn from a limited sample space. Hence, the number of possible values of the forecasted gold proportion $\hat{p}_{i,t+1}^{(s)} = p_{i,t}^{(s)} + \hat{\Delta p}_{i,t+1}^{(s)}$ is sparse, causing less accurate forecasts.

The functional autoregressive (FAR) model (Chen and Tsay, 1993) augments the expression in (2) by giving the autoregressive coefficients a functional form, allowing them to change continuously. Generally, the model postulates

$$\Delta p_{i,t}^{(s)} = \alpha_1^{(s)}(p_{i,t-d}^{(s)})\Delta p_{i,t-1}^{(s)} + \alpha_2^{(s)}(p_{i,t-d}^{(s)})\Delta p_{i,t-2}^{(s)} + \alpha_3^{(s)}(p_{i,t-d}^{(s)})\Delta p_{i,t-3}^{(s)} + \epsilon_t.$$

where $\alpha_k(\cdot)$ are locally continuous functions, ϵ_t are IID errors independent to $p^{(s)}_{i,t-j}$, $j \geq 1$, and $d \geq 1$ is a pre-determined lag. To incorporate information from the AR model, we replace the coefficient function's argument $p_{i,t-d}$ with some deterministic function $f(X^{(s)}_{i,t-k}, \Delta p^{(s)}_{i,t-k}, k = 1, 2, 3 | \alpha^{(s)}, \beta)$. For simplicity and to circumvent overfitting, we assume f is the one-step-ahead prediction:

$$f^{(s)}_{i,t} := f(X^{(s)}_{i,t-k}, \Delta p^{(s)}_{i,t-k}, k = 1, 2, 3 | \alpha^{(s)}, \beta) = \hat{\Delta p}^{(s)}_{i,t}.$$

Effectively, this formulation establishes a functional autoregressive model with exogenous variables (FARX), where the exogenous variables are implicitly given in the arguments of the autoregressive functions:

$$\Delta p^{(s)}_{i,t} = \alpha^{(s)}_1(f^{(s)}_{i,t})\Delta p^{(s)}_{i,t-1} + \alpha^{(s)}_2(f^{(s)}_{i,t})\Delta p^{(s)}_{i,t-2} + \alpha^{(s)}_3(f^{(s)}_{i,t})\Delta p^{(s)}_{i,t-3} + \epsilon_t. \tag{3}$$

For computational simplicity, we assume $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$. Fan and Yao (2005) proposed using Taylor Series and locally weighted least-squares to estimate the coefficient functions. The specific formulae could be found in Fan and Gijbels (1996). To obtain the one-step-ahead forecast using the FARX model, we calculate $f^{(s)}_{i,t+1}$, which is known from existing data, and use local methods such as splines or kernels to estimate each autoregressive coefficient $\hat{\alpha_j}(f^{(s)}_{i,t+1})$. The final estimation is

$$\hat{\Delta p}^{(s)}_{i,t+1} = \hat{\alpha_1}^{(s)}(f^{(s)}_{i,t+1})\Delta p^{(s)}_{i,t} + \hat{\alpha_2}^{(s)}(f^{(s)}_{i,t+1})\Delta p^{(s)}_{i,t-1} + \hat{\alpha_3}^{(s)}(f^{(s)}_{i,t+1})\Delta p^{(s)}_{i,t-2}. \tag{4}$$

We use the same model to explain total medals by reversing the roles of $p^{(s)}_{i,t}$ and $m^{(s)}_{i,t}$. Namely, we model $\Delta m^{(s)}_{i,t} := m^{(s)}_{i,t+1} - m^{(s)}_{i,t}$ and use $p^{(s)}_{i,t}$ as an exogenous variable in the FARX formulation. We proceed to empirically estimate the parameters of models (2) and (3).

6.2 Parameter Estimation

We first estimate the parameters for the gold medal model. The parameters in (2) are calculated using Maximum Likelihood Estimation. The results are shown in Table 3. Unsurprisingly, we observe a positive impact of hosting the Olympics, indicated by the positive β_1 estimates for all Sports. The performance of the ARX model is relatively poor for Handball, Hockey, Modern Pentathlon, and Volleyball, producing an in-sample standard deviation of over 15%.

Having obtained the ARX model coefficients, we turn to estimating the coefficient functions in the FARX model (3). To estimate $\alpha(f_0)$, Fan and Yao (2005) proposed using the first-order Taylor

Sport	α_1	α_2	α_3	β_1	β_2	β_3	Estimated σ (%)
Athletics	-0.444	-0.200	-0.136	0.051	0.145	-0.124	2.276
Beach Volleyball	-0.651	-0.348	0	0.184	0.279	-0.221	14.209
Boxing	-0.387	-0.066	0	0.087	-0.244	-0.159	5.744
Cycling	-0.560	-0.392	-0.094	0.144	0.136	-0.210	8.062
Handball	-0.714	-0.598	0	0	0.240	-0.292	18.475
Hockey	-0.786	-0.459	-0.189	0.175	0.168	-0.189	21.080
Modern Pentathlon	-0.812	-0.486	-0.255	0.075	0.191	-0.139	17.591
Rowing	-0.670	-0.268	0	0.026	0.256	-0.244	6.577
Sailing	-0.617	-0.370	-0.075	0.035	0.293	-0.228	5.964
Tennis	-0.749	-0.359	-0.154	0.046	0.444	-0.391	10.218
Triathlon	-0.605	-0.350	0	0	0.342	-0.315	12.544
Volleyball	-0.529	-0.474	0	0.118	0.360	-0.268	18.076

Table 3: Estimated parameters of the model in (2). Insignificant coefficients are replaced by 0.

expansion of the coefficient function at some f in the neighborhood of f_0:

$$\alpha_j(f) \approx \alpha_j(f_0) + \alpha_j'(f_0)(f - f_0) \equiv a_j + c_j(f - f_0).$$

where a_j and c_j are local intercept and slope estimators of $\alpha_j(\cdot)$. We choose n training data points with $f_{i,t}^{(s)}$ local to f_0 and perform a local least-squares estimation by minimizing the following expression

$$\sum_{i,t}[\Delta p_{i,t}^{(s)} - \sum_{j=1}^{3}\{a_j + c_j(f_{i,t}^{(s)} - f_0)\}\Delta p_{i,t-j}^{(s)}]^2 K_h(f_{i,t}^{(s)} - f_0)$$

where $K_h(x) = \frac{3}{4h} max\{0, 1 - (\frac{x}{h})^2\}$ is the Epanechnikov kernel with bandwidth h. Weighted Least-Squares theory could be used to find a local minimizer $(a_1, a_2, a_3, c_1, c_2, c_3)$. The bandwidth h is found using an estimation method by Cai, Fan, and Yao(2000), which performs a grid search and selects the bandwidth that minimizes the in-sample average prediction error.

We estimate the values of $\alpha_j(\cdot), j = 1, 2, 3$, at 100 equally spaced lattice points between the minimum and maximum values of $f_{i,t}^{(s)}$. However, for sports that have only participated in a few Olympics or contain few events, severe overfitting is observed. We remedy such occurrences by using the ARX coefficients when an extreme $\alpha_j(\cdot)$ value is estimated. A cubic spline is used to interpolate

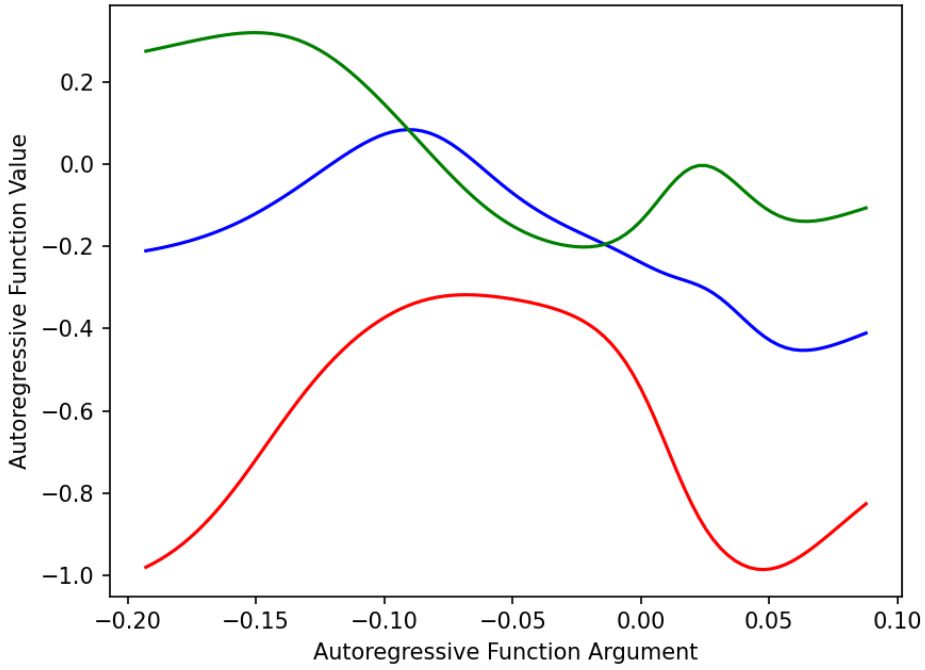

Figure 6: Estimated autoregressive functions of the Athletics sport. Red: α_1. Blue: α_2. Green: α_3.

Sport	Athletics (48)	Cycling (22)	Rowing (15)	Sailing (10)
ARX	2.276	8.062	6.577	5.964
FARX	1.833	5.292	5.568	3.980

Table 4: In-sample standard deviation (%) of the ARX and FARX models. The number in parentheses next to the sport names is the number of events belonging to the sport at the 2028 Olympics.

values between the lattice points. The in-sample variance of the FARX model is lower than that of the ARX model in sports with many events, as shown in Table 4.

The three estimated functions for Athletics are shown in Figure 6. Interestingly, the first- and second-order autoregressive coefficients are negative for almost all model inputs, indicating teams' performance in "unstable" sports display a significant reversal phenomenon – teams whose performances have declined over recent Olympics are likely to bounce back, and vice versa. This phenomenon is present in all "unstable" sports and is consistent with our parameter estimations for the ARX model, where $\alpha_1, \alpha_2 < 0$ for all sports.

We repeat the same process to model total medals. The in-sample standard deviations for both the ARX and the FARX models are significantly less than that for the gold medal model, and the FARX

Country	Expected Golds	95%CI Lower Bound	95%CI Upper Bound
United States	15.206	13.298	17.114
Great Britain	9.623	7.678	11.569
Netherlands	9.348	7.454	11.242
Italy	5.772	4.456	7.088
Australia	5.656	4.211	7.101

Table 5: Projected top gold medal winners of unstable sports at the 2028 Olympics.

model maintained a lower standard deviation for sports with many events, although the difference is less significant.

6.3 Model Forecasting

We use equation (4) to generate forecasts for each country's $\hat{\Delta p}^{(s)}_{i,2028}$ and $\hat{\Delta m}^{(s)}_{i,2028}$ in the "unstable" sports at the 2028 Olympics. The estimated gold proportions is found by $\hat{p}^{(s)}_{i,2028} = \hat{\Delta p}^{(s)}_{i,2028} + p^{(s)}_{i,2024}$ and similarly for medal proportions. Since we do not restrict $\hat{p}^{(s)}_{i,2028}$ to sum to 1 over all countries, we normalize it and calculate the expected gold medal count by

$$\hat{p}^{(s)}_{i,2028} \leftarrow \frac{\hat{p}^{(s)}_{i,2028}}{\sum_i \hat{p}^{(s)}_{i,2028}}; \quad E[g_{i,2028}] = \sum_s \hat{p}^{(s)}_{i,2028} m^{(s)}_{2028}$$

where $E[g_{i,2028}]$ is the expected number of "unstable" gold medals of country i in the 2028 Olympics, and $m^{(s)}_{2028}$ is the number of events belonging to Sport s. The same procedure is implemented to calculate the expectation of total medals.

The top winners are shown in Tables 5 and 6. We also present the 95% confidence interval of the estimated medal counts. The winner tables are coherent with our intuitions, with the United States remaining dominant and followed by Great Britain and the Netherlands, two countries that are historically top performers in Cycling, Rowing, and Sailing.

7 Modeling New Sports

In this section, we postulate models to predict the number of golds and total medals a country wins in "new" Sports, which either have had less than 5 occurrences at the Olympics (the existing "new"

Country	Expected Total Medals	95%CI Lower Bound	95%CI Upper Bound
United States	39.212	35.929	42.496
Great Britain	36.607	32.282	40.931
Netherlands	23.849	20.487	27.212
France	15.028	11.872	18.184
Australia	14.665	12.204	17.127

Table 6: Projected top total medal winners of unstable sports at the 2028 Olympics.

Coefficient	α_1	α_2	β_1	β_2	β_3
Estimated Value	0.256	0.085	0.132	-0.333	0.786

Table 7: Estimated parameters of the model in (5), q = 2.

Sports) or would premiere at the 2028 Olympics (the premiering "new" Sports).

7.1 Existing "New" Sports

We first model Sports that have recently joined the Olympics core program. According to our criterion, these existing "new" sports are hosted at the Olympics less than 5 times but have participated in the 2020 and 2024 Olympics. Assuming that the mechanism dictating how gold medals are distributed in the initial years of a Sport remains the same across all sports, we propose the following model

$$p_{i,t}^{(s)} = \beta^T X_{i,t}^{(s)} + \alpha_1 (p_{i,t-1}^{(s)} - \beta^T X_{i,t-1}^{(s)}) + \cdots + \alpha_q (p_{i,t-q}^{(s)} - \beta^T X_{i,t-q}^{(s)}) + \epsilon_t \tag{5}$$

which is an ARX model similar to (2), but $p_{i,t}^{(s)}$ is modeled instead of $\Delta p_{i,t}^{(s)}$ and the autoregressive coefficients are independent of the Sport. X_t is defined identically as in (2), except $h_i, t = 0$ if country i hosted the previous Olympics. The ARX order q should be (the Sport's number of occurrences in the Olympics - 1), which is the maximum possible to produce a forecast. Fitting (5) using Sports that have participated in more than 5 Olympics gives the parameter estimates. We provide an example of q = 2 in Table 7. The procedure for modeling total medals is similar to Section 5.2, where we reverse the roles of the explanatory variable $p_{i,t}^{(s)}$ and exogenous variable ($m_{i,t}^{(s)}$). A one-step-ahead forecast of $\hat{p}_{i,t+1}^{(s)}$ and $\hat{m}_{i,t+1}^{(s)}$ is calculated, and subsequent manipulations lead to an estimate of the number of gold medals and total medals a country would win in the existing "new" sports at the 2028 Olympics.

Country	Expected Gold	Expected Medals	Country	Expected Gold	Expected Medals
United States	7.070	16.306	Australia	2.195	6.552
Great Britain	3.632	8.598	Japan	1.867	5.703
Canada	2.249	5.457	France	1.215	7.115

Table 8: Top winners in the ”new” Sports.

7.2 Premiering ”New” Sports

At the time of writing, the International Olympic Committee has announced the addition of five Sports to the 2028 Olympics that are absent at the 2024 Olympics, two of which are debuting (Flag Football, Squash) and three of which are returning (Baseball/Softball, Cricket, and Lacrosse), totaling 10 events. Without holistic knowledge about the development of the five Sports, we use a simple model to estimate the gold and total medals a country wins.

1. For the returning Sports, we estimate $p_{i,2028}^{(\hat{s})}$ and $m_{i,2028}^{(\hat{s})}$ by calculating the proportion of Sport s gold(total) medals historically won by country i over all historic Sport s gold(total) medals.

2. For the new Sports, we set $p_{i,2028}^{(\hat{s})}$ to be the proportion of all gold medals country i won in the 2024 Olympics. Similarly, $m_{i,2028}^{(\hat{s})}$ is the proportion of all medals country i won in the 2024 Olympics.

While this straightforward model introduces room for bias, we do not see an effective measure to accurately predict the results for the premiering ”new” sports without running into risks of overfitting. In addition, the premiering ”new” Sports account for roughly 3% of total medals, hence allowing some bias would not significantly impact the accuracy of our estimated medal board.

7.3 Top Winners in the ”New” Sports

Table 8 presents six well-performing countries in the ”new” Sports. The United States once again dominates the list.

8 Medal List Prediction and Breakthrough Detection

8.1 Estimtaed Gold and Total Medal Lists

In the preceding sections, we developed and fitted distinct models for each group of Sports classified in Section 2. Each model produced estimates of the number of gold and total medals each country would win in each sport. In this section, we combine our previous forecasts into a medal list.

The expected number of gold and total medals country i wins are calculated by summing the forecasted medal counts for each group of Sports, thanks to the linearity of expectations. To generate a 95% confidence interval of the number of expected gold and total medals, we perform a Monte Carlo simulation to generate the error terms' empirical multivariate distribution. We display in Tables 9 and 10 the top 10 winners by gold medals and total medals, as well as their associated confidence intervals, rounded to the nearest whole number. The complete medal lists are uploaded to this Google Drive.

Country	**Estimated Golds**	**95%CI Lower Bound**	**95%CI Upper Bound**
United States	49	43	59
China	48	38	66
Japan	21	17	28
Great Britain	20	17	24
Australia	16	13	21
Germany	15	12	20
France	14	11	18
Italy	11	9	14
Netherlands	11	9	13
Canada	8	7	10

Table 9: Estimated Gold Medal List and Confidence Intervals.

The estimated tables are consistent with the recent trend that the United States and China battle

Country	Estimated Total Medals	95%CI Lower Bound	95%CI Upper Bound
United States	137	124	187
China	111	96	173
Great Britain	76	68	97
France	58	51	81
Japan	46	41	69
Germany	46	41	65
Australia	44	39	59
Italy	34	30	48
Netherlands	30	26	37
Canada	27	24	37

Table 10: Estimated Total Medals List and Confidence Intervals .

for the top spot. The model accurately captured the phenomenon that China's total medal count is significantly less than that of the US, despite the two having similar gold medal counts. Other historically top-performing countries, including Japan, Great Britain, Australia, France, and the Netherlands, are ranked in the top 10 in both tables as well. The top-heavy prediction falls in line with real-world observations. We're confident that our estimated medal lists resemble realistic ones and match medal list trends in recent Olympics.

8.2 2028 First Medal Breakthrough Detection

Since only historical medal counts and binary participation indicators are available, and purely data-driven approaches are required, predictions for zero-medal countries have to rely solely on these features. The entropy of forecasting first-time winners is contained in historical participation factors with the assumption that near-term participation patterns may heavily influence the likelihood of a breakthrough.

We extend our Negative Binomial framework to incorporate 2 lagged participant counts (lag 1 and lag 2) and 6 lagged "presence" indicators. This structure considers recent participant levels and how historically erratic attendance still conveys a small but persistent boost if a nation maintained an Olympic program.

Let $\{Y_{i,t}\}$ be the total medals for team i in year t. We fit a Negative Binomial regression:

$$Y_{i,t} \sim \text{NegBinomial}\big(\mu_{i,t},\ \phi\big), \tag{6}$$

$$\log\big(\mu_{i,t}\big) = \beta_0 + \sum_{k=1}^{2} \beta_{p_k} \log\big(\text{lag\_participants}_{k,i,t} + 1\big) + \sum_{\ell=1}^{6} \beta_{\text{pr}_\ell} \log\big(\text{lag\_present}_{\ell,i,t} + 1\big) + u_i. \tag{7}$$

Here:

- $\text{lag\_participants}_{k,i,t}$ is the participant count lagged by k cycles, with $\log(\cdot + 1)$ to handle zeros.
- $\text{lag\_present}_{\ell,i,t}$ is the binary indicator (1 if present, 0 otherwise), lagged by ℓ cycles.
- u_i is a random intercept by team, $u_i \sim \mathcal{N}(0,\ \sigma_u^2)$.
- ϕ is the Negative Binomial dispersion parameter.

8.2.1 Posterior Estimation

We place weakly informative priors on the regression coefficients $\beta_0,\ \beta_{p_k},\ \beta_{\text{pr}_\ell}$ and on $\sigma_u,\ \phi$. The posterior distribution is given by

$$p\big(\Theta \mid \{Y_{i,t}\}\big) \propto \prod_{i,t} \text{NegBinomial}\Big(Y_{i,t} \mid \mu_{i,t}(\Theta),\ \phi\Big) \times p(\Theta), \tag{8}$$

where $\Theta = \{\beta_0, \beta_{p_k}, \beta_{\text{pr}_\ell}, u_i, \sigma_u^2, \phi\}$.

8.2.2 Measure Metrics and Results

We estimate the probability that team i wins at least one medal by computing:

$$\widehat{P}\big(Y_{i,2028} \geq 1\big) = \frac{1}{S} \sum_{s=1}^{S} \mathbf{1}\big(Y_{i,2028}^{(s)} \geq 1\big), \tag{9}$$

where $Y_{i,2028}^{(s)}$ is the predicted medal count under the s-th draw of the model parameters, and $\mathbf{1}(\cdot)$ is an indicator function.

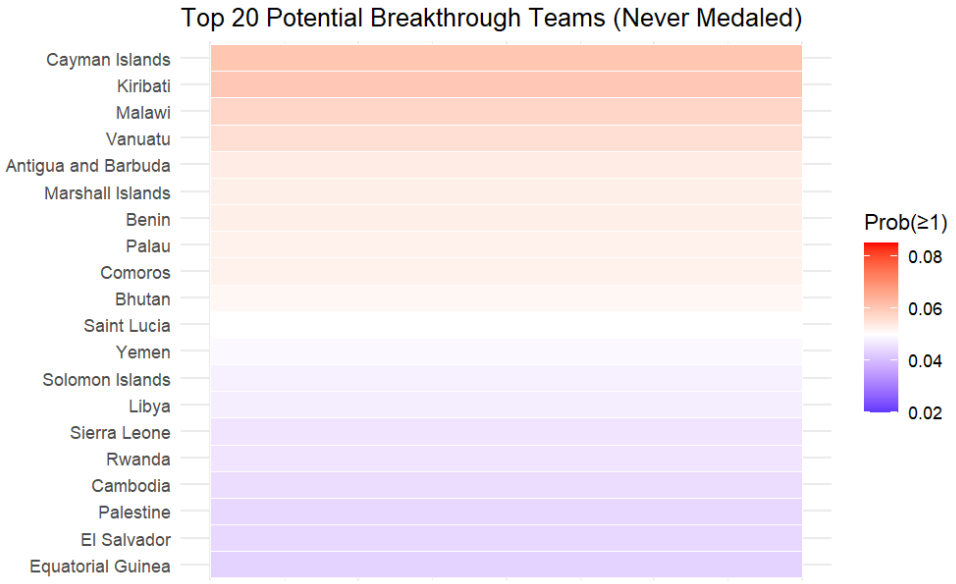

Figure 7: Countries with top possibilities of breakthrough.

Cayman Islands, Kiribati, Malawi, Vanuatu, Antigua and Barbuda, Marshall Islands, Benin, Palau, Comoros Bhutan have Probablity larger than 0.05 to win first medal in history.

9 Great Coach Effect

We analyzed the impact of Béla Károlyi on both the Romanian and U.S. gymnastics teams, demonstrating the existence of **Great Coach** effect in gymnastics and discussing its statistical significance. Following this, we examined Lang Ping's influence on the Chinese and U.S. volleyball teams, showing that while there is an effect, it is less pronounced compared to gymnastics. Finally, we explored potential benefits for other countries and sports that may consider investing in a "great coach," summarizing the insights and offering perspectives on future developments for some countries.

9.1 Béla Károlyi and Gymnastics

Béla Károlyi became the head coach of the Romanian women's gymnastics team in the 1970s. In 1981, Károlyi defected to the United States and began coaching the U.S. women's gymnastics team.

He coached the U.S. team until 1992, then took a break before returning as a national team coordinator in 1996. Károlyi remained involved with the U.S. team until he retired in 2001. Note that in **Figure**

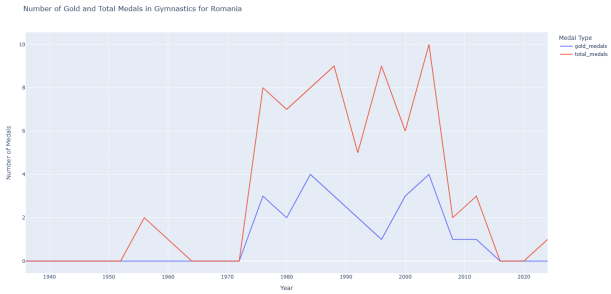

Figure 8: Medal trend of Romania Gymnastics

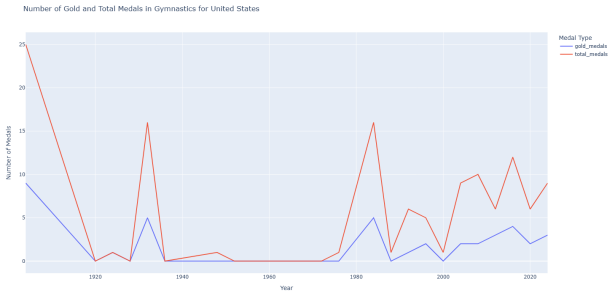

Figure 9: Medal trend of USA Gymnastics

8, there is a great spike starting during 1970s in the number of both gold and total medals earned by Romanian Gymnastics team. Similarly, we can also observe the same trend of spike in **Figure 9** for the USA Gymnastics team during 1980s. Károlyi's coaching brought a transformative impact to both teams, leading to significant improvements in their performances on the international stage. The sharp increase in medals during these periods highlights the importance of a skilled and influential coach, underscoring the existence and significance of the "Great Coach" effect in gymnastics.

9.2 Lang Ping and Volleyball

Lang Ping is a renowned volleyball coach who has made significant contributions to both the Chinese and U.S. women's volleyball teams. We all know that Lang Ping is definitely a great coach, and that throughout her career, Lang Ping's coaching has been instrumental in the success of both teams. However, we notice that the trend of Chinese Volleyball team (**Figure 10**) is not so significant, while USA team (**Figure 11**) does display a trend of increase in the number of medals they get during Lang Ping's coaching period. Note that Volleyball was first introduced to Olympics in 1964, and there are not enough data points for us to generalize and conclude the significance of **Great Coach** effect in the overall success of the team only with the data provided.

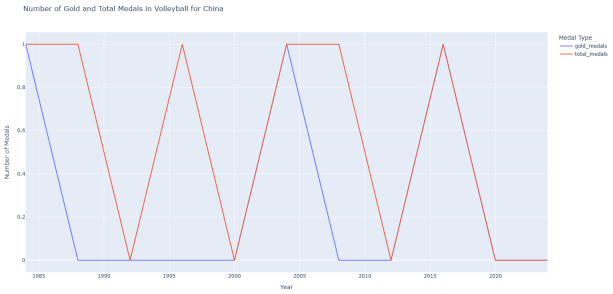

Figure 10: Medal trend of China Volleyball

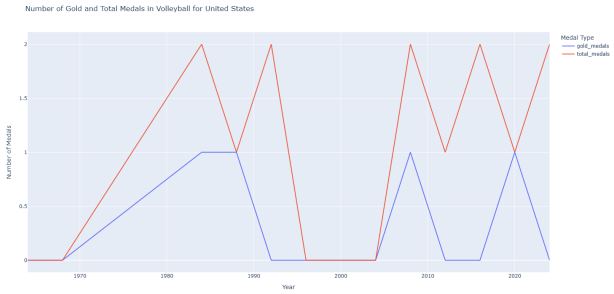

Figure 11: Medal trend of USA Volleyball

9.3 Investment in a great coach

From our analysis in **9.1** and **9.2**, we get to a conclusion that the **Great Coach** effect exists, but not as significant as thought. More specifically, much more data will be needed, not just medal data, in order for us to analyze the importance of a **Great Coach** comprehensively. Here are 3 examples of how a country can invest in a great coach:

- **India – Cricket**: India could benefit from investing in a ”great” coach for its cricket team to improve consistency, particularly in Test cricket. A top coach could refine techniques and strategies, helping India secure more victories and potentially increase their ICC Test Championship points.
- **Australia – Soccer (Football)**: Australia’s soccer team could improve with a ”great” coach to enhance technical skills, tactics, and overall performance. This could help them perform better in the World Cup and AFC Asian Cup.
- **Nigeria – Basketball**: Nigeria’s basketball team could rise to greater heights with a ”great” coach to develop players’ skills and team cohesion. This could help them perform better in the Olympics and FIBA World Cup.

10 Other Insights

We highlight two intriguing findings: evidence proving excessive doping in East Europe during the Cold War, and the significant disparities in Olympic performance between the United States and China

due to different Sports bureaucratic systems.

10.1 Cold War Doping Evidence

The fumes of the Cold War are everywhere found, including in the Olympic games. Opposing nations sought to prove their strengths by outperforming in the Olympics, and the use of banned substances emerged. Sports doping quickly became a trend in Eastern European countries and is most well-documented in East Germany (see Franke(1997)). While the athletes and countries involved with excessive doping won the moment, the physical damage and political backlash were equally as significant. Through our data, we spotted evidence suggesting the use of illegal substances by Eastern European countries.

Figure 12 shows the Bayesian model intercepts for different countries in descending order. It is worth noting that the intercepts could be considered an elusive part of the model, as they are not explained by empirical model inputs. Hence, large intercepts are equivalent to considerable unexpected deviations from the model, implying there is some omitted external influence.

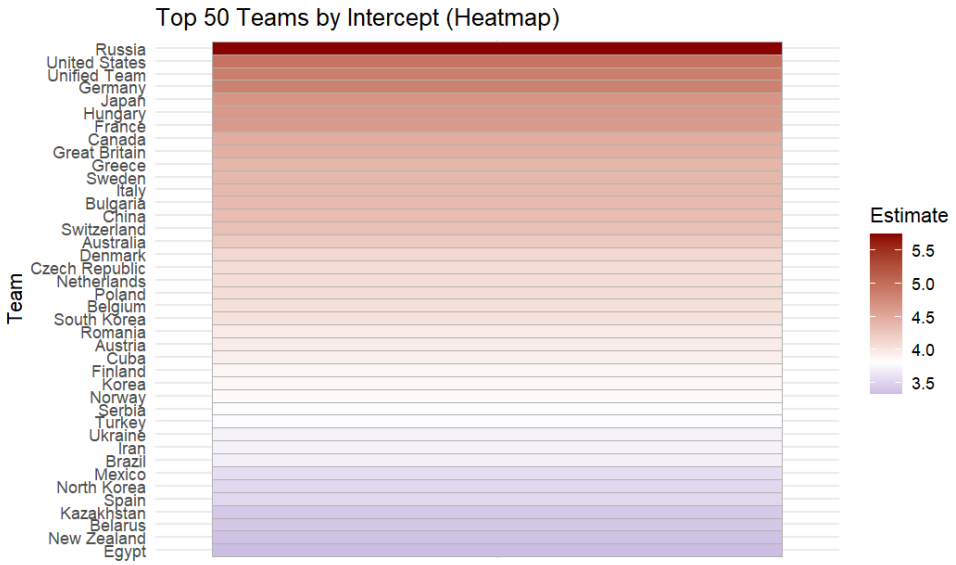

Figure 12: Bayesian model intercept for different countries.

Since Eastern European countries, including Russia, (East) Germany, Hungary, and Bulgaria pile up at the top of Figure 12, we suspect this graph could provide evidence that these nations are affiliated

with illegal sports substances. However, the United States is also at the top of the graph, which could hint at potential doping in the US as well.

10.2 US-China Disparity

Despite the United States and China having similar gold medals in many of the recent Olympics, our models indicate the two countries earn gold medals from different Sports classes. Particularly, the estimated number of gold medals China earns from the "stable" Sports is 44, compared to the US's 27. On the contrary, the US is expected to earn 39 medals from "unstable" Sports, compared to China's 10. China's fixation and attention to "stable" Sports could be due to its bureaucratic system, in which Sports authorities are operating under notoriously high pressure. This could lead to their unwillingness to allocate budget to riskier and previously unexplored "unstable" Sports. The US, under less authoritative pressure to perform in the Olympics, accomplishes exceptionally well in "unstable" and "new" Sports while maintaining good competition for China in the "stable" Sports.

References

[1] Cai, Z., Fan, J., & Yao, Q. (2000). Functional-Coefficient Regression Models for Nonlinear Time Series. Journal of the American Statistical Association, 95(451), 941–956.

[2] Chen, R., & Tsay, R. S. (1993). Functional-Coefficient Autoregressive Models. Journal of the American Statistical Association, 88(421), 298–308.

[3] Schlembach, C., Schmidt, S. L., Schreyer, D., Wunderlich, L. (2022). Forecasting the Olympic medal distribution – A socioeconomic machine learning model. Technological Forecasting and Social Change, Volume 175.

[4] Fan, J., & Gijbels, I. (1996). Local Polynomial Modelling and Its Applications. Chapman and Hall, London.

[5] Fan, J., & Yao, Q. (2005). Nonlinear Time Series: Nonparametric and Parametric Methods. Springer.

[6] Franke, W. (1997). ”Hormonal doping and androgenization of athletes: a secret program of the German Democratic Republic government”. Clinical Chemistry. 43 (7): 1262–1279.