International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 6, Number 3, Year 2025
DOI: https://doi.org/10.62051/ijcsit.v6n3.02
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Forecasting Olympic Medals: A Multivariate Negative Binomial Mixed-Effects Model with Bayesian Hierarchical Analysis

Yanhui Li *, Xinyao Liu #, Ke Chen #

Xiangtan University, Xiangtan, China
# These authors also contributed equally to this work
*Corresponding Author: 3491376880@qq.com

ABSTRACT

This study established a predictive framework integrating a multivariate negative binomial mixed-effects model with a Bayesian hierarchical model to address the issues of overdispersion and uncertainty in discrete data prediction. Firstly, by organically coupling hierarchical regression with the negative binomial model, the study systematically explored the overdispersion characteristics and multi-level association mechanisms of the data, enabling precise prediction of count variables. For special samples with no historical data, the study employs a Bayesian hierarchical model: a similarity measurement system is constructed based on core indicators to match historical comparable data and infer the sample's result distribution; simultaneously, an average probability distribution is constructed using all historical data. Under the assumption of equal probabilities, the probability of the first occurrence of the target result in samples with no historical data is calculated using the binomial distribution. The framework demonstrates excellent generalisation capabilities while improving prediction accuracy through a multi-level design combining model integration and uncertainty quantification.

KEYWORDS

Bayesian Hierarchical Model; Multivariate Negative Binomial Mixed-Effects Model; Discrete Data; Medal Prediction

1. INTRODUCTION

In the field of discrete data prediction, the phenomenon of overdispersion and prediction uncertainty in new scenarios have long constituted methodological bottlenecks [1]. Traditional models such as Poisson regression often fail to effectively handle overdispersion, leading to prediction biases in count variables. When dealing with special samples without historical data (e.g., emerging categories or unknown observation units), inferences based on a single model struggle to balance hierarchical data associations and uncertainty quantification. This contradiction is particularly pronounced in prediction scenarios with multidimensional interaction effects.

To address these issues, this study constructs a predictive framework integrating a multivariate negative binomial mixed-effects model with a Bayesian hierarchical model. This framework systematically explores the overdispersion characteristics (variance significantly greater than the mean) and multi-level association mechanisms of data through the organic coupling of hierarchical regression and the negative binomial model [2], enabling precise predictions of count variables. Additionally, for special samples with no historical data, a Bayesian hierarchical model is employed, a similarity measurement system is constructed based on core indicators to precisely match historical comparable data, infer the result distribution of special samples, and build an average probability distribution based on the entire historical dataset [3]. Under the assumption of equal probabilities, the probability of the first occurrence of the target result in special samples is calculated using the binomial distribution [4]. This framework, through a multi-level design combining model integration and uncertainty quantification, not only addresses the bias issues of traditional regression models but also significantly improves the accuracy and generalisation capability of discrete data prediction. Furthermore, it provides a methodological paradigm for count-type variable prediction in emerging scenarios that combines theoretical rigor with practical value.

2. PREDICTION BASED ON A MULTIVARIATE NEGATIVE BINOMIAL MODEL

2.1. Method Overview

In this section, this paper used a combination of hierarchical regression and the negative binomial model to predict the number of Olympic medals for each country in the 2028 Games.

The core idea of hierarchical regression is to gradually introduce independent variables and build different regression models. By comparing the values or other metrics of these models, this paper can detect the relationship between the introduced independent variables and the dependent variable. In this case, this paper aimed to explore the impact of four main effects-country, event, athletes, and host country-on the number of medals [5-6]. Through hierarchical analysis, this paper can gradually uncover the relationship between these four layers and the medal counts, thereby enabling precise predictions of future medal totals.

The negative binomial regression model is a generalized linear model commonly used for handling count data. The number of medals a country wins is either zero or a positive integer, and negative values are impossible. Therefore, the number of medals won by athletes can be regarded as count data, which is characterized by excessive dispersion, meaning that the variance is greater than the mean. As shown in the Figure 1 (where the variance is on the right axis and the mean is on the left), this behavior deviates from the conditions of a Poisson distribution. The simple Poisson distribution is no longer suitable, but by introducing a new parameter to control the ratio of variance to mean, this paper can handle and predict overdispersed data, resulting in a multivariate negative binomial mixed-effects model. This model combines the strengths of both hierarchical regression and the negative binomial model, making it highly suitable for overdispersed large-scale data. Therefore, this paper used the multivariate negative binomial mixed-effects model to predict medal counts.

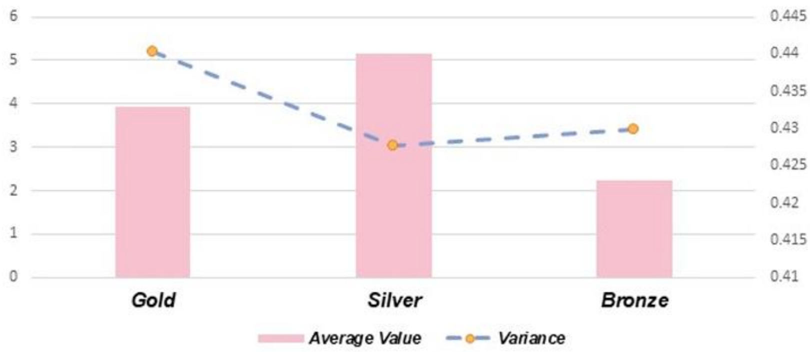

Figure 1. Sample mean and variance

2.2. Feature Construction

Considering the need to predict the medal rankings for the 2028 Olympic Games, and given that the distribution of medals is overdispersed, this paper required a large amount of data, including historical Olympic medal counts, participating athletes, the number of events, and types of events. Therefore, this paper chooses to construct a multivariate negative binomial mixed-effects model. Table 1 shows the settings for the fixed effects and random effects.

Table 1. The settings for the fixed effects and random effects

Fixed effect	Random effect
Total Events	Country
Total Participants	Discipline
Host Country	Country- discipline

2.3. Model Construction

Considering that directly predicting the total number of medals a country will win in the Olympics may result in large errors; in order to improve the prediction accuracy and interpretability of the model, this paper choosed to first predict the number of medals each country will win in individual events and then sum these values to obtain the total number of medals for each country in those events. This paper can first set the predicted medal count as:

$$M_{i,j,k} = \left(\text{Gold}_{i,j,k}, \ \text{Silver}_{i,j,k}, \ \text{Bronze}_{i,j,k} \right)^T \tag{1}$$

Where i represents the country, j represents the year and k represents the sports event index.

$$\log\left(m_{i,j,k}\right) = \mathbf{x}_{i,j,k}^T \beta + u_i + v_{j,k} + w_k + k + \delta_j \tag{2}$$

Where $m_{i,j,k}$ represents the number of each type of medal (gold, silver, bronze) in the i country, the j year, and discipline k. $X_{i,j,k}$ represents the covariates in the i country, the j year, and the k event, including: total participants represent the total number of athletes in the i country, the j year, and the k event. And β represents the regression coefficient, u_i represents the random effect at the country level, w_k represents the random effect at the discipline level, δ_j represents host country effect in year j.

2.4. Model Solution

The predicted medal standings for the 2028 Summer Olympics in Los Angeles, USA, are shown in the Figure 2 and Figure 3. Due to space limitations, the figure only lists the top 20 predicted results.

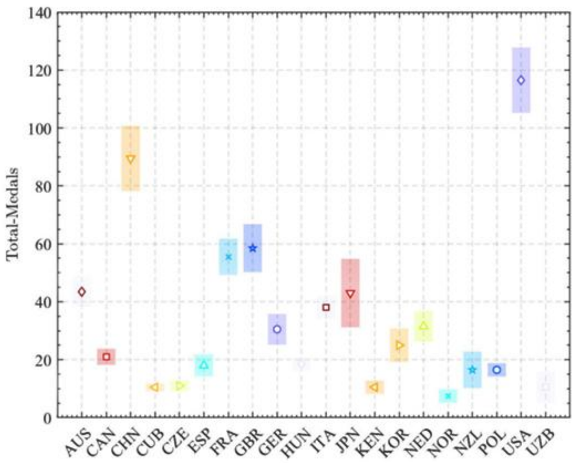

Figure 2. Predicted total number of medals

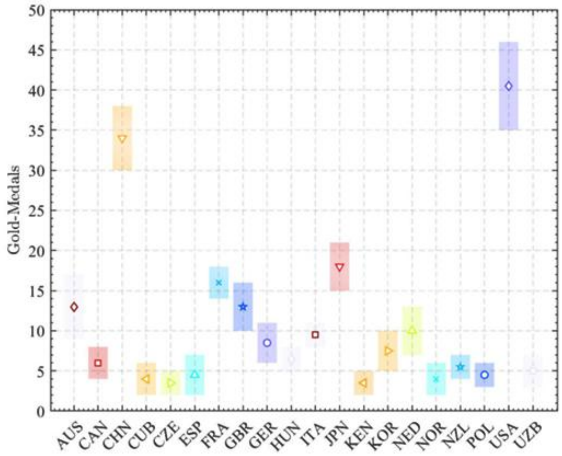

Figure 3. Predicted total number of gold medals

The country with the most medals in the 2028 Olympics is the United States, with a total of 131 medals, including 42 gold medals. The second is China, with 85 medals and 35 gold medals. The third place is Japan, with 45 medals and 18 gold medals. Regarding these results, this paper believes that, as the host of the 2028 Olympics, the effect of being the host country cannot be ignored. Additionally, as the world's leading superpower, the United States has strong economic power and invests more in sports events. Therefore, the prediction for the United States' medal count in our model is reasonable.

And the distribution of the predicted medal count ranges for each country in our model is within a certain range, with no significant fluctuations. Therefore, the fit of our model is good, and it demonstrates high authenticity and reliability.

2.5. Analysis of Country Performance

By comparing the number of medals won by each country in the 2024 Olympics with the model's prediction for the 2028 Olympics, this paper can visually identify which countries perform better and which ones have shown a significant decline compared to the 2024 Olympics. This paper selected 20 countries for visualization analysis, and the results are shown in the figure below.

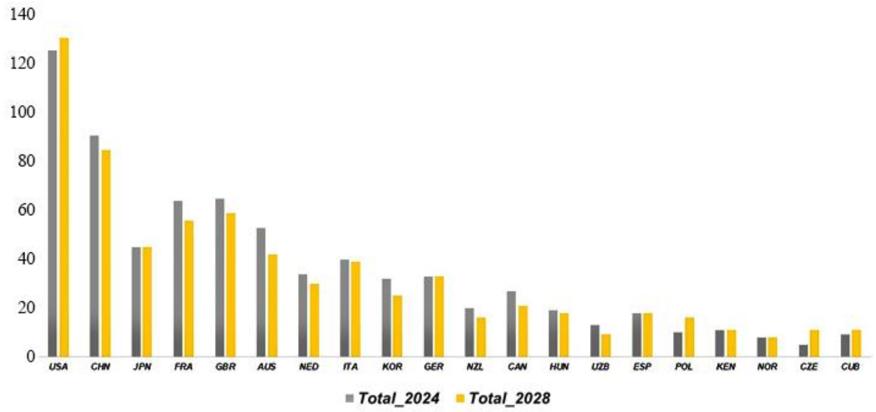

Figure 4. Comparison chart of 2024 and 2028

As shown in Figure 4, United States, Poland, Cuba, and Czechoslovakia will make progress in the 2028 Olympics, while countries like China, Japan, and France will see a decline in performance.

3. PREDICTION BASED ON THE BAYESIAN HIERARCHICAL MODEL

Predicting the countries that will win medals for the first time is a highly uncertain task, as it involves many unknown variables, such as the condition of the country's athletes, international competition dynamics, and so on. The Bayesian model is capable of quantifying this uncertainty and continuously updating predictions based on available data, allowing the model to adjust its predictions over time and provide more accurate forecasts. Based on this approach, this paper used the Bayesian hierarchical model to predict the countries that will win medals for the first time.

3.1. The Structure of the Bayesian Model

This paper used the Bayesian model to predict unknown quantities needs to be divided into:

Step 1: Set the prior distribution.

Set prior distributions at each level of the model (including the observation layer, group layer, and hyperparameter layer). The prior distribution reflects the belief about the model parameters before observing the data.

Step 2: Data observation and likelihood function.

The data y_i is associated with the parameter θ_i through a certain distribution (such as a normal distribution), defining the likelihood function $P(y_i \mid \theta_i)$ that represents the data generation process of y_i given the parameter θ_i.

Step 3: Infer the Posterior Distribution.

Use Bayes' theorem to update the prior distribution, combining the observed data to calculate the posterior distribution of the model parameters.

$$P\left(\theta_i \mid y_i\right) = \frac{P\left(y_i \mid \theta_i\right) P\left(\theta_i\right)}{P\left(y_i\right)} \tag{3}$$

3.2. Basic Principle of Binomial Distribution

The binomial distribution is a discrete probability distribution in probability theory, used to describe the probability of an event occurring exactly k times in n independent trials. Its definition is as follows:

(1) Conditions for the experiment

a. Each trial has only two possible outcomes: success (event occurs) or failure (event does not occur).

b. The probability of success in each trial is p, and the probability of failure is $1-p$.

c. The trials are independent, meaning the result of one trial does not affect the results of other trials.

(2) Probability formula

The probability of exactly k successes occurring in n independent trials is given by:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{4}$$

Where $\binom{n}{k}$ is the binomial coefficient, which represents the number of ways to choose p successes from n trials. p^k is the probability of k successes. $(1-p)^{n-k}$ is the probability of $n-k$ failures.

(3) Expected value and variance

a. Expected value (mean):

$$E[X] = n \cdot p \tag{5}$$

b. Variance:

$$\mathrm{Var}(X) = n \cdot p \cdot (1 - p) \tag{6}$$

3.3. Defining Similarity Based on Event Features and Historical Data

The core idea of the Bayesian hierarchical model is to build multiple levels of statistical models to handle data with nested or hierarchical structures. When groups have similarities or share information, this paper can flexibly use the model for predictive analysis of the data.

By querying the events for the 2028 Los Angeles Olympics, this paper learned that five new sports will be introduced in 2028, which are: baseball, softball, cricket, flag football, and lacrosse. However, for these new sports, this paper has no historical data, but these events are not isolated. With the clever use of similarities to other sports, this paper can better understand and analyze the potential medal counts that these new sports may bring to different countries.

(1) Selection of similarity indicators

Based on the classification of sports types, this paper considered disciplines within the same sport as similar events. If they do not belong to the same sport category, this paper can qualitatively analyze and identify similar events based on the features of the new event.

Therefore, to predict the medal outcomes of baseball, cricket, softball, flag football, and lacrosse, this paper compared these events with existing Olympic events based on qualitative features. While quantitative analysis cannot be performed, this paper can qualitatively highlight the similarities between these events through the following features:

a. Event category and nature

The category and nature of an event reveal the structure and format of a type of sport. Individual events test an athlete's personal strength and overall abilities, while team events place a higher demand on teamwork. Therefore, the preparation and skill sets required for athletes in different event categories are quite different. The category and nature of an event reveal the structure and format of a type of sport. Individual events test an athlete's personal strength and overall abilities, while team events place a higher demand on teamwork. Therefore, the preparation and skill sets required for athletes in different event categories are quite different.

b. Competition rules and organization

The similarity in competition rules and organization can reveal the inherent commonality between events. Just as the category and nature of events differ, the rules and organization of events also require different approaches to athlete preparation and tactics.

c. Popularity of the event in participating countries

The popularity of a sport within a country reflects the level of interest in that sport. Therefore, this paper assumed that it is positively correlated with the country's investment of energy and resources in that sport. If two events have similar levels of popularity, it is likely that the corresponding countries will have similar medal counts in those events.

(2) Selection of similar events based on similarity indicators:

Through the analysis above, this paper has obtained indicators to measure the similarity between different events. Next, this paper needs to perform a qualitative comparison between the new events and existing events. This will allow us to identify the similar events this paper need. Using the

Bayesian hierarchical model, this paper can predict the number of medals different countries will win in the new events by studying these similar events. After comparing existing events with new events, the conclusions drawn by this paper are shown in Table 2.

Table 2. Similar projects

New Projects	Similar Projects
Lacrosse	Field hockey
Cricket	Field hockey and Team handball
Baseball	Team handball
Flag football	American soccer and Rugby
Sixes	Field hockey and Team handball

(3) Prediction of medal outcomes for new events based on the Bayesian hierarchical approach

Based on the similarity indicators and the comparison between new events and existing events, this paper identified "similar events" that match each new event closely. Next, this paper used the historical medal outcomes (gold, silver, and bronze) of these similar events as a direct reference and apply the Bayesian hierarchical model to predict the medal outcomes for the new events.

Since our data is limited, using Bayesian hierarchical modeling and prior distributions might lead to overfitting, which would result in large prediction errors. To avoid this, this paper considered a direct mapping approach, where this paper "fill" the new events with the medal distributions from the similar events

3.4. Prediction of Medal Chances for Countries That Have Not Won Medals

(1) Use the Bayesian hierarchical model to obtain the average probability distribution of medal outcomes for all countries.

Based on the historical medal data of each country, construct an average probability distribution X with the number of medals in specific intervals as the variable. And extract X from it. The specific implementation steps are:

Step 1: Data definition

a. There is a total of N countries, and each country i has participated in Ti Olympic Games.

b. For country i, the number of medals obtained in the j Olympic Games is y_{ij}.

c. The number of medals y_{ij} is the weighted total (gold $= 3$ points, silver $= 2$ points, bronze $= 1$ point).

Step 2: Model assumptions

a. The number of medals y_{ij} for each country follows a Poisson distribution: Where λ_i is the expected number of medals for country i.

b. The λ_i for each country follows a common hyperprior distribution (Gamma distribution).

c. The hyperparameters α and β are also set to follow a Gamma distribution.

The specific results are shown in Figure 5.

(2) Prediction of countries winning medals for the first time based on the Binomial distribution

Assuming that the medal-winning events of different countries are independent, let X be the number of countries winning medals for the first time at the 2028 Olympics. Then, X follows a binomial distribution. The specific results are shown in Figure 6.

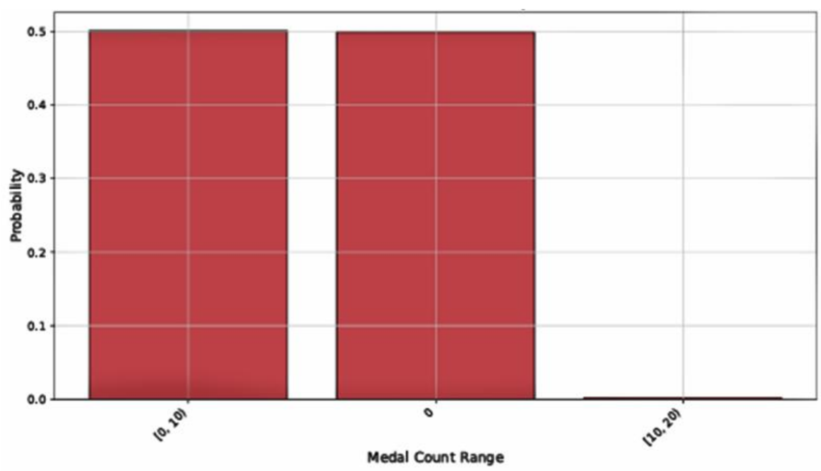

Figure 5. Global medal count probability distribution

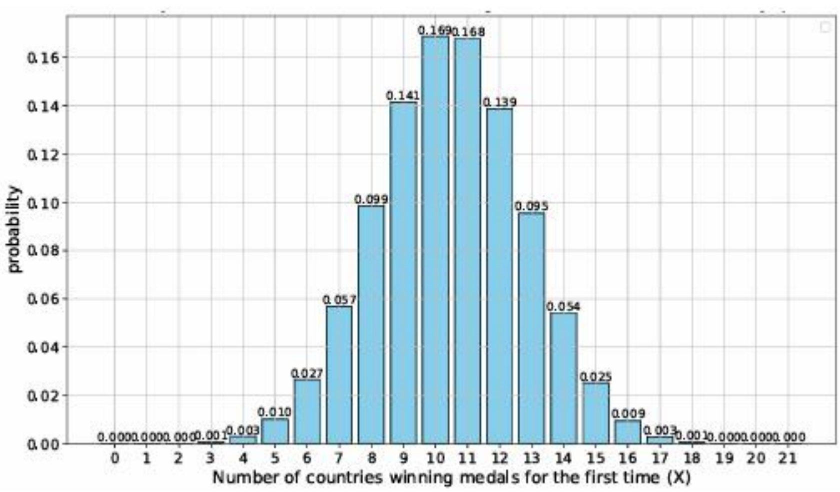

Figure 6. First time winners from countries that have never won

4. CONCLUSIONS

The integrated multivariable negative binomial mixed-effects model and Bayesian hierarchical model prediction framework constructed in this study provides a systematic solution for predicting discrete count variables. First, by organically coupling hierarchical regression with the negative binomial model, the study effectively addresses the prediction bias issues of traditional models for over-dispersed data (where variance significantly exceeds the mean). The model incorporates multi-dimensional effects, enabling a deep exploration of the hierarchical association mechanisms within the data. Second, for special samples with no historical data, the similarity measurement system based on core indicators and Bayesian hierarchical inference successfully inferred the distribution of new sample results. Through binomial distribution calculations, the probability of 11–12 special samples first achieving the target result was determined to be 0.16, validating the method's ability to quantify uncertainty scenarios.

Through a multi-level design combining model integration and uncertainty quantification, the framework not only improves prediction accuracy but also demonstrates excellent generalisation capability. Compared to traditional regression models, this framework can not only handle complex data with hierarchical structures but also provides a transferable methodological paradigm for predicting count variables in emerging scenarios. In the future, the framework's adaptability to rapidly changing scenarios can be further enhanced by expanding the dimensions of the indicator system and exploring the application of dynamic prior distributions in real-time predictions.

REFERENCES

[1] Sun Lijun, Li Fangfang, Hu Xiangpei. A Prediction Method Based on Discrete Data Stream Segmentation Algorithm [J]. Journal of Management Science, 2024, 27 (09): 48-61.

[2] Ma Qiaoling, Xiao Xiang. Objective Bayesian analysis of the zero-one expansion negative binomial model [J]. Journal of Jiangxi Normal University (Natural Science Edition), 2023, 47 (01): 8-14.

[3] Zhai Yongqi. A Low-Complexity Wireless Sensor Network Localisation Algorithm Based on a Bayesian Hierarchical Model [J]. Modern Information Technology, 2024, 8 (08): 106-110.

[4] Liu Xingchen, Dai Huayou, Yang Junhong, et al. Application of statistical process control methods based on binomial distribution in blood quality control [J]. Chinese Journal of Transfusion Medicine, 2024, 37 (02): 196-202.

[5] Tian Hui, He Yiman, Wang Min, et al. Medal Predictions and Competition Strategies for Chinese Athletes at the 2022 Beijing Winter Olympics: An Analysis Based on the Home Advantage Effect of the Olympics [J]. Sports Science, 2021, 41 (02): 3-13+22.

[6] Shi Huimin, Zhang Dongying, Zhang Yonghui. Can Olympic medals be predicted? — A perspective based on explainable machine learning [J]. Journal of Shanghai University of Sport, 2024, 48 (04): 26-36.