Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Data Prediction and Intervention Effect Analysis Based on Random Forest and DID Algorithm

Yue Zhong, Bin Wu *

School of Big Data and Basic Science, Shandong Institute of Petroleum and Chemical Technology, Dongying, China

* Corresponding Author Email: 2983731827@qq.com

Abstract. In this paper, negative binomial regression model, random forest model, PageRank algorithm and PSM-DID algorithm are proposed, focusing on the application of multi-models in the prediction of count data, the assessment of importance of feature variables, the analysis of association networks and the quantification of intervention effects. First, to address the over-dispersion of count data, a negative binomial regression model is used, which solves the limitations of traditional Poisson regression and realizes the effective prediction of count variables by introducing negative binomial distribution modeling. Second, the random forest model is constructed and combined with the SHAP method, based on the principle of additive feature attribution, the marginal contribution of the feature variables is weighted to quantify the degree of influence of each variable on the prediction results, to realize the modeling of the complex nonlinear relationship and the importance ranking of the features. Then, the node association network is constructed based on the PageRank algorithm, and the ordering of node potential in the network is realized by defining a stochastic wandering model with a damping factor and iteratively calculating the smooth distribution value of the nodes; finally, the PSM-DID algorithm is utilized to quantify the net effect of interventions by eliminating the selection bias through the propensity score matching and stripping the influence of the temporal trend in conjunction with the double-difference method. These methods can effectively handle count data, nonlinear relationships and network structure data, enhance the stability of analysis results through complementary validation between models, and provide a structured quantitative analysis framework for data modeling and causal inference in multiple fields.

Keywords: SHAP; PageRank; PSM; DID; random forest model; negative binomial regression model.

1. Introduction

This paper focuses on the application of multi-models in data modeling and analysis in sports, aiming to explore the methodological paths of count data prediction, correlation analysis, and quantification of the effects of intervention factors through multiple algorithms [1]. First, a negative binomial regression model is introduced to effectively solve the problem of over-dispersion of data, which can model the prediction of domain-specific count variables and analyze the influence mechanism of different factors on the outcome variables [2]. Second, the random forest model is constructed and combined with the SHAP interpretation method, which decomposes the contribution of each feature variable to the prediction results into marginal effects through the additive feature attribution framework, and realizes the portrayal of complex nonlinear relationships and the importance ranking of feature variables [3]. Then, based on the PageRank algorithm, the medal potential association network between countries is constructed, and the smooth distribution values of nodes are iteratively calculated to quantify the strength and ranking of medal potential association between countries by defining the random wandering process with damping factor [4]. Finally, the PSM-DID algorithm is applied to eliminate sample selection bias through propensity score matching, combined with the double difference method to strip out the time trend interference, and assess the net influence effect of specific intervening factors on the outcome variables [5] [6].

The experimental results show that the model exhibits good adaptability to overdispersion in count data prediction, provides a reliable framework for causal inference, and the synergistic application of

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

multiple models provides a systematic quantitative tool for data analysis and decision-making in related fields.

2. Count Data Modeling and Characterization Based on Negative Binomial Regression and Random Forests

2.1. Negative Binomial Regression Model

2.1.1. Negative binomial distribution

Suppose there is a series of independent Bernoulli experiments. In each experiment, the probability of success is ρ and the probability of failure is $(1-\rho)$. This sequence is observed until a predefined number of successes r occurs. Then the resulting random number of failures X will have a negative binomial distribution:

$$X \sim NB(r, p) \tag{1}$$

2.1.2. Negative binomial regression model

When the dependent variable is a counting model, the linear regression model $y = x\beta + \varepsilon$ is no longer applicable. Because when the log-linear model is used, an implicit assumption is that y obeys the normal distribution, but the y here can obviously only be non-negative.

Poisson regression assumes that the mean and variance of the dependent variable are equal, but for over-dispersed data such as the number of medals, the assumption of Poisson regression is no longer valid. The model can be corrected in the following two ways:

1. Use Quasi-Poisson distribution: A method to improve Poisson regression by weighting the variance.

2. Use Negative Binomial Regression: Directly use the negative binomial distribution to model, which can adapt to the overdispersion of the data.

Assume that the probability of event A occurring in a certain experiment is p, and n independent random experiments are conducted. Let the number of times A occurs be Y, then the probability of $Y = y$ is:

$$P(Y = y) = C_n^y p^y (1-p)^{n-y} \tag{2}$$

It can be seen that when n is very large and p is very small, $\lambda = np > 0$, the binomial distribution can be approximated by the Poisson distribution.

Applied to the count model, for individual i, the probability of $Y_i = y_i$ is:

$$P(Y_i = y_i \mid X_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \tag{3}$$

Where λ_i is the average number of occurrences of an event.

$$P(Y_i \mid X_i) = Var(Y_i \mid X_i) = \lambda_i \tag{4}$$

Since $E(Y_i|X_i)$ is non-negative, assume that:

$$E(Y_i|X_i) = \lambda_i = e^{X_i\beta} \tag{5}$$

Taking the logarithm of both sides gives:

$$\ln \lambda_i = X_i \beta \tag{6}$$

The form of the negative binomial regression model is very similar to Poisson regression:

$$\log(E(y_i)) = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \mu_i \tag{7}$$

2.1.3. Solving of negative binomial regression models

Considering the host effect, the strong country effect, and the factors of small countries obtaining medals, the number of gold medals and medals won by some countries is shown in Fig. 1:

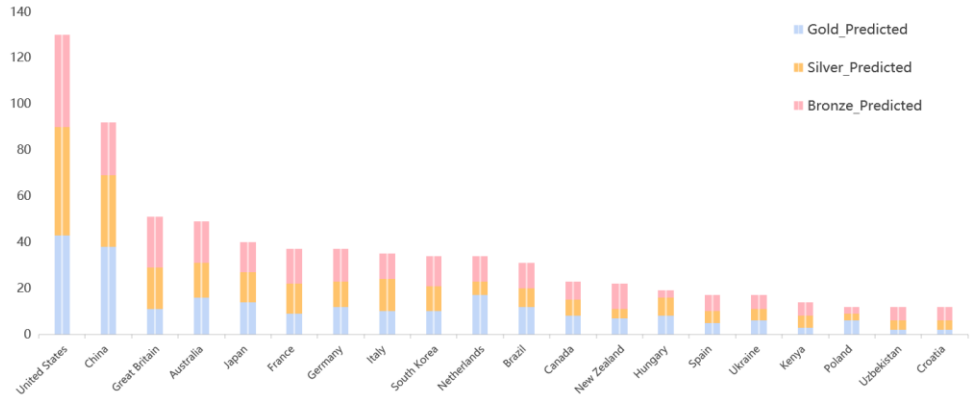

Fig 1. 2028 Los Angeles Olympic Medal Table Forecast (Top 20 Countries).

According to the Fig. 2, in the 2028 Los Angeles Olympics medal table forecast, the countries that performed well include the United States, Germany, the Netherlands and Brazil. The United States remains strong and is expected to win 130 medals, an increase of 4 medals compared to 2024. Although China's gold, silver and bronze results have fluctuated, it is expected to maintain a high performance.

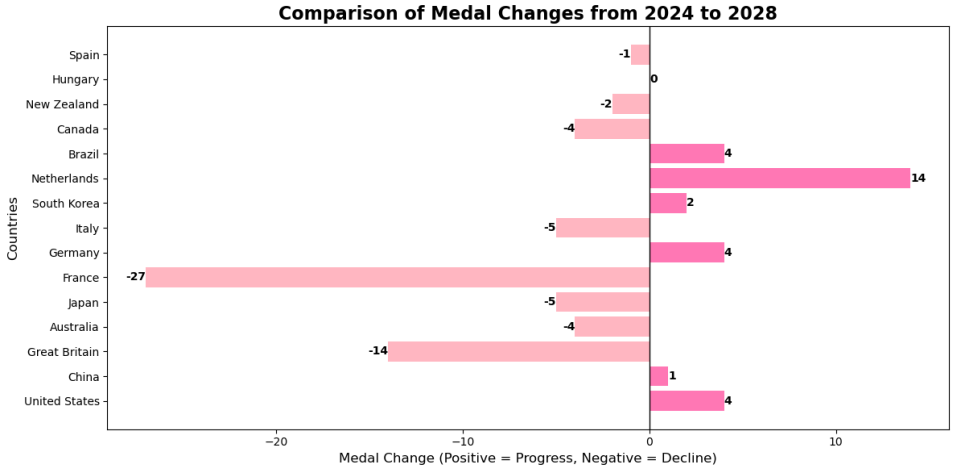

Fig 2. Comparison of Medal Changes.

2.2. Nonlinear Feature Importance Assessment Based on Random Forest Modeling

2.2.1. Random forest model construction

This paper uses the SHAP method as an interpretable method to evaluate and compare the importance of different feature variables through the Shapley value. The key method used in this paper is the additive feature attribution method. This method interprets the contribution of each feature variable

to the model prediction result as "the contribution of this variable (x) to the final prediction result (y) when it participates in the model prediction". The "total prediction contribution" of a prediction model can be expressed as:

$$g(x)=\phi_0+\sum_{i=1}^{M}\phi_i(x)\mathbb{k}(x_i) \tag{8}$$

Where $x=(x_1,\cdots,x_M)'$ is an M-dimensional explanatory or characteristic variable, $\mathbb{k}(x_i)$ is a binary indicator variable; $g(x)$ represents the final prediction result, ϕ_0 represents the predicted mean, and $\phi_i(x)$ represents the marginal contribution of the characteristic variable x_i to the prediction result.

Here, $g(x)=\phi_0+\sum_{i=1}^{M}\phi_i(x)$ represents "the logarithm of the number of awards/gold medals won by a certain team in a certain event in a certain year", ϕ_0 represents the mean number of awards/gold medals won by all teams in this event, and x_i represents the value of the i-th characteristic variable. By measuring $\phi_i(x)$, the impact of the change of x_i on the number of awards/gold medals won can be got, so as to find the features that make a greater contribution to predicting the change of medals.

The Shapley value of the feature variable x_i is its contribution to the prediction result $g(x)$, which is calculated by weighted summing the marginal contribution of the feature variable to the model prediction result.

$$\phi_i(x)=\sum_{S\subseteq F\setminus\{i\}}\frac{|S|!(|F|-|S|-1)!}{|F|!}[f_{S\cup\{i\}}(x_{S\cup\{i\}})-f_S(x_S)] \tag{9}$$

Among them, i represents the i-th independent variable, $\phi_i(x)$ is its contribution to the model prediction result, F is the feature variable set used by the model, S is a subset of $F\setminus\{i\}$, x_s is all the feature variables contained in S, $x_{S\cup\{i\}}$ contains x_s and x_i.

2.2.2. Solution of random forest model

The SHAP method is used in the random forest model to identify the predictive variables that affect the final number of awards/gold medals. The specific identification process is as follows:

1. The random forest model is used to train the number of awards/gold medals by project. The training set is the data 1896 to 2021, and the test set is the data of the 2024 Paris Olympics.

2. A random forest model is constructed for the prediction of the number of awards/gold medals for each project.

3. The SHAP method is used to analyze the model training results and identify the characteristic variables in each project that are most closely related to the predicted number of awards/gold medals.

2.2.3. Empirical results and analysis

Using the random forest model to predict the performance of each Olympic team in each event based on the performance data of the Summer Olympic Games teams from 1896 to 2024, the differences in the predictability of performance in different Olympic events are compared. The prediction results are shown in table 1:

Table 1. 2028 Los Angeles Olympic Medal Table Prediction.

Country	Gold_SPre	Total_SPre	Gold_LPre	Total_LPre	Contrast_2024
U. S.	42	129	43	130	+3
China	40	89	38	92	-2
Great Britain	18	63	11	51	-2
Australia	16	62	16	49	+9
Japan	18	52	14	40	+7
France	22	51	9	37	-13
Germany	11	37	12	37	+4
Italy	17	53	10	35	+13

From the Fig. 3, it can be seen that there are certain differences in the predictions of gold medals and total medals by different methods. At the same time, the changes in the total medals in 2028 and 2024 reflect the dynamic development trend of the competitive sports strength of various countries. The United States is still the leader in gold medals and total medals. The prediction results of random forest and negative binomial regression methods are not much different, which further confirms the stability of its comprehensive strength.

The prediction results of the United States and China's gold medals and total medals are relatively close under the two methods, indicating that its position as a sports power remains stable.

The performance of European countries shows a trend of differentiation. The total number of medals in the United Kingdom and France is expected to decrease. In contrast, Germany and Italy performed more impressively.

Japan, Australia and Canada are expected to increase their medals. In addition, the random forest method has higher results than the negative binomial regression method in medal prediction for some countries due to its more accurate characterization of complex nonlinear relationships.

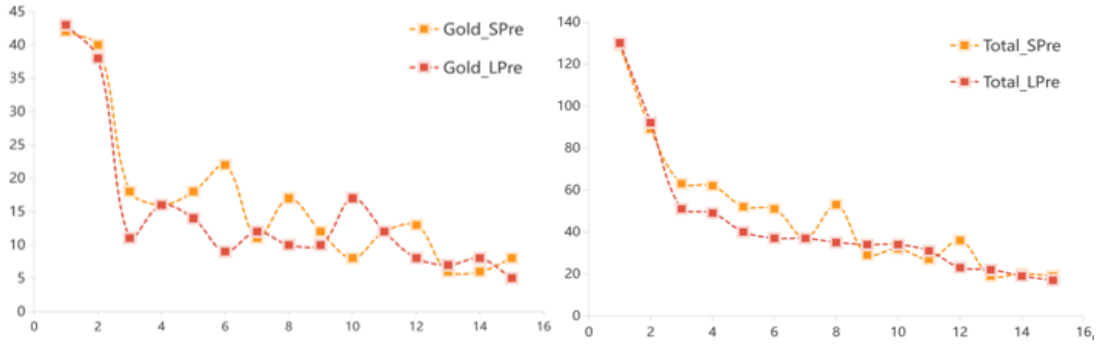

Fig 3. Line chart of the forecast of the number of gold medals and all medals in 2028.

2.3. Medal Potential Research Based on PageRank Algorithm

2.3.1. Definition of PageRank

Given an arbitrary directed graph with n nodes, a general random walk model is defined on the directed graph, namely, a first-order Markov chain. The transfer matrix of the general random walk model consists of a linear combination of two parts. One part is the basic transfer matrix M of the directed graph, which means that the transfer probability from one node to all the nodes connected to it is equal. The other part is a completely random transfer matrix, which means that the transfer probability from any node to any other node is $1/n$. The linear combination coefficient is the damping factor d $(0 \le d \le 1)$. This general random walk Markov chain has a stationary distribution, denoted by R.

$$R = dMR + \frac{1-d}{n}1 \tag{10}$$

2.3.2. PageRank algorithm implementation

Given a directed graph with n nodes and a transfer matrix M, the general PageRank of the directed graph is determined by the limit vector R of the iterative formula. The iterative algorithm of PageRank is to perform iterations according to this general definition until convergence.

2.3.3. Medal potential prediction based on PageRank algorithm

Network construction: Each country is considered a node in the network, and the medal potential relationship between countries is considered an edge. If two countries have similar medal distribution in a certain sport, a connecting edge is established and a weight is assigned.

PageRank calculation: Based on the correlation matrix, the PageRank algorithm is used to rank the medal potential of each country. The algorithm iteratively calculates the PR value of each country and finally obtains a ranking that includes the medal potential of each country.

It can be calculated that Belize has the highest probability of winning the award, 17.5%, indicating that it is expected to have a more positive performance at the 2028 Olympic Games; Angola, Mali and Vietnam have a probability of winning the award of 8.5%, 4.6% and 3.2%, respectively.

3. Causal Inference Modeling of Intervention Effects Based on PSM and DID

3.1. Propensity Score Matching Method

The PSM algorithm constructs a propensity score function to calculate the probability value of each country receiving guidance from an "excellent coach" based on the characteristics of the country. Then the countries that receive guidance and the countries that do not receive guidance are matched according to the propensity score, so that the two groups of countries after matching are as similar as possible in these characteristics, thereby eliminating the selection bias.

First, the propensity score is calculated using the logistic regression model:

$$\log it(P(D=1|X)) = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_k X_k \tag{11}$$

Among them, X is a covariate of whether the country accepts "great coaching", $P(D=1|X)$ is the calculated propensity score, and D is a binary variable indicating whether it accepts guidance.

To obtain the contribution of "great coaching", the average intervention effect (ATT) of the intervened countries is calculated to quantify the contribution of "great coaching"

$$\tau_{ATT} = E(\tau | D=1) = E(Y(1) | D=1) - E(Y(0) | D=1) \tag{12}$$

Then calculate the propensity score, which is the probability that a country will be coached by a “great coach”.

$$P(X) = P(D=1|X) \tag{13}$$

Conditional Independence Assumption CIA: Given a set of observable covariates, the potential outcome and the mean intervention assignment are independent of each other.

$$(Unconfoundedness)\ Y_0, Y_1 IID|\ X, \forall X \tag{14}$$

All variables that affect intervention allocation and potential outcomes can be considered to be observed simultaneously. If the above equation holds, then intervention allocation and potential outcomes are also conditionally independent based on $P(X)$.

When CIA is satisfied, ATT can be estimated:

$$\tau_{ATT}^{PSM} = E_{P(X)|D=1}\left\{E[Y(1) \mid D=1, P(X)] - E[Y(0) \mid D=0, P(X)]\right\} \tag{15}$$

3.2. Difference-in-Difference Method

In the analysis of the "great coach" effect, the time dimension is used before and after the coaching, and the countries or projects that received and did not receive coaching are used as the grouping dimension. The "great coach" effect is evaluated by comparing the changes in the number of medals of the two groups at different times. The double difference method can be understood as a simulation of a random assignment experiment, verifying the causal relationship without a random experiment.

The specific steps are as follows:

Step 1: Grouping. For a natural experiment, all sample data are divided into two groups: one group is affected by the intervention, namely the experimental group; the other group is not affected by the same intervention, namely the control group;

Step 2: Target selection. Select a target indicator to be observed, such as purchase conversion rate and retention rate, which is generally the KPI that you want to improve;

Step 3: First difference. Perform two differences (subtraction) before and after the intervention to obtain two sets of differences, representing the relative relationship between the experimental group and the control group before and after the intervention;

Step 4: Second difference. Perform a second difference on the two sets of differences, thereby eliminating the original differences between the experimental group and the control group, and finally obtaining the net effect brought by the intervention. The final formula for the "great coach" effect is:

$$q = \Delta Y_{treatment} - \Delta Y_{control} \tag{16}$$

Among them, q is the score of "great coach", $\Delta Y_{treatment}$ represents the number of medals with coaching guidance, and $\Delta Y_{control}$ represents the number of medals without coaching guidance.

3.3. Solving the Model

The implementation of the common support test is a crucial step when matching propensity scores, as it significantly improves the validity of the matching process. The common support test ensures that the selected matching samples are comparable in all respects, thus reducing the risk of subset effects, which may negatively affect the accuracy of the matching results. Therefore, the validation of common support ensures the quality of matching and avoids analytical errors caused by uneven sample characteristics. From the kernel density Fig. 4, it can be seen that the deviation of the kernel density curves between the two groups before matching is relatively large, while the kernel density curves after matching are relatively close to each other, indicating that the common support assumption is satisfied.

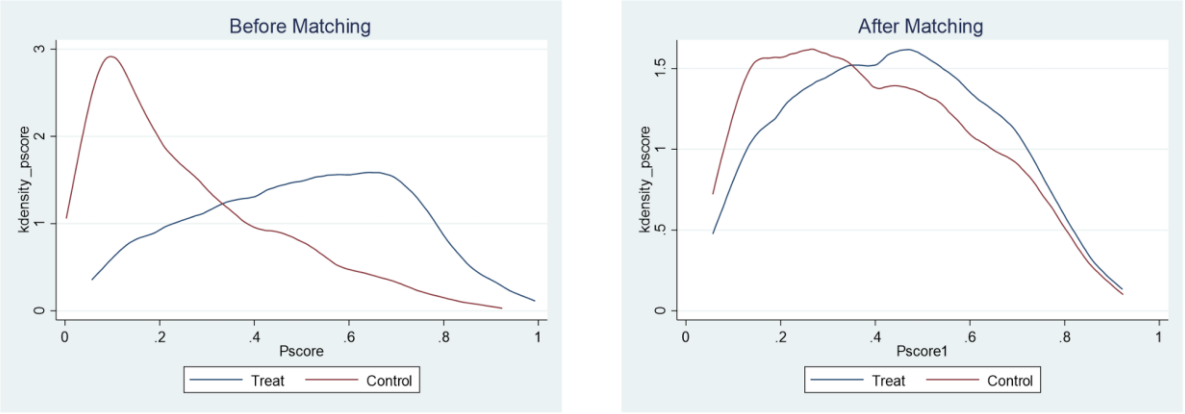

Fig 4. Kernel density plot of scores before and after matching.

After calculating the propensity score, the countries with a higher probability of being influenced by a “great coach” are shown in the table 2:

Table 2. Countries and probability of being affected by “great coaches”.

Country	NOC	Probability	Sports
United States	USA	0.6542	Gymnastics
China	CHN	0.5118	Weightlifting
Japan	JPN	0.4835	Badminton

It can be found that the United States, China, and Japan have a probability of 0.6542, 0.5118, and 0.4835 respectively to receive the guidance of the "great coach", and the probability is greater than or equal to 50%. These three countries have shown great potential in gymnastics, weightlifting, and badminton. Therefore, it was decided to continue to select the United States, China, and Japan as the countries directed by the "great coach". After the double difference method, it was concluded that after these three countries receive the guidance of the "great coach", it is expected that by the 2028 Los Angeles Summer Olympics, their probability of winning the projects shown in the figure will increase, which further verifies the potential and influence of the "great coach" in these countries.

4. Conclusion

In this paper, a multidimensional quantitative analysis framework is constructed, and negative binomial regression model, random forest model, PageRank algorithm and PSM-DID algorithm are introduced to show the advantages and application value of the method for count data, nonlinear relationship, network association and causal inference. The negative binomial regression model is based on the negative binomial distribution, solves the problem of over-dispersion of count data, breaks through the Poisson regression assumption, accurately fits the data distribution, and provides a tool for the prediction of count variables and reveals the influence mechanism of variables. Random forest model combined with SHAP explanation method decomposes the prediction results by additive feature attribution logic, realizes high-precision prediction and model explanation, and systematically identifies the key influencing factors. The PageRank algorithm transforms national medal potential associations into smooth distribution calculation of network nodes through the stochastic wandering model with damping factor, explores the implicit association structure and potential ranking, and provides a new perspective for competitiveness assessment. The PSM-DID algorithm eliminates the selection bias through the propensity score matching, and combines with double-difference method to strip away the temporal trend, which provides a framework of causal inference for the assessment of the effect of intervention factors, and enhances the reliability of the conclusions. Enhance the reliability of conclusions. Future research can expand the application of the model in dynamic data tracking and cross-domain correlation analysis, and explore the path of multi-model fusion to enhance the depth and breadth of system analysis.

References

[1] Zhu Yin. An empirical analysis of the factors affecting the Olympic medal table - taking the 31st Olympic Games as an example [J]. Journal of Chifeng University (Natural Science Edition), 2017, 33(03): 123-127. DOI: 10.13398/j.cnki.issn1673-260x.2017.03.048.

[2] Wang Qiaoyu. Zero-expansion Poisson-negative binomial mixed counting model and its application[D]. Northwest Normal University, 2024.DOI:10.27410/d.cnki.gxbfu.2024.002528.

[3] Wu Yan. A malicious encrypted traffic prediction model integrating random forest and SHAP[J]. Journal of Harbin University of Commerce (Natural Science Edition),2024,40(02): 167-178.DOI: 10.19492 /j.cnki.1672-0946.2024.02.014.

[4] Zhang Bingtao,Wei Dan,Shen Yu,et al. An improved K-mean clustering algorithm based on PageRank[J]. Journal of Beijing University of Posts and Telecommunications,2025,48(02): 18-27.DOI:10. 13190/j. jbupt.2023-098.

[5] Guo Xuchang. A study on the coaching behavior of competitive sports coaches [D]. Fujian Normal University, 2008.

[6] Xia Xulan. An empirical analysis of the impact of the eastern region's leading development strategy on economic development based on PSM-DID [D]. Heilongjiang University, 2024. DOI: 10.27123/ d.cnki.ghlju.2024.001171.