Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 12, No. 3, 2025

Prediction of Cybercrime and Policy Evaluation Using a Combination of PageRank and ARIMA-DID Algorithms

Jiayu Gao *, Yijia Suo, Wanyun Xing

Northeast Agricultural University, Harbin, Heilongjiang, China

*** Corresponding author:** Jiayu Gao (Email: 15765510395@163.com)

Abstract: This paper focuses on the distribution characteristics of global cybercrime and the effectiveness of policy interventions, constructing a multi-model analysis framework that integrates PageRank and ARIMA-DID. Firstly, the study uses the PageRank algorithm to calculate the cybercrime connectivity of countries, identifying the United States as the core risk node. Based on indicators such as cybersecurity preparedness and the number of cases, countries are categorized into high, medium, and low-risk groups. It is found that high-risk countries exhibit the characteristics of 'high attack volume, high losses, and low cybersecurity preparedness.' Further combining the ARIMA model with the difference-in-differences (DID) method, using the United States' 2013 cybersecurity policy as an example, the study constructs a counterfactual prediction scenario, confirming that policy implementation significantly reduces actual attack volumes compared to predicted values, but long-term effects are influenced by external factors. Cross-national comparisons show that low-risk countries like Japan and Australia have better policy sustainability, and preventive measures (like vulnerability management) have significantly higher cost-effectiveness than post-incident responses. The study provides a risk assessment model and policy effectiveness quantification tools for global cybersecurity governance.

Keywords: PageRank Algorithm; ARIMA Model; Difference-in-Differences (DID) Method; Global Cybercrime.

1. Introduction

5G and artificial intelligence are driving the deep interconnection of global networks, but they have also given rise to rampant transnational cybercrime, with frequent incidents of ransomware and data breaches. Against this backdrop, gaining a scientific understanding of the intrinsic connections within the global cybercrime ecosystem and accurately quantifying the actual effectiveness of policy interventions have become critical issues in building cybersecurity defenses [1].

This paper breaks through the limitations of traditional research's single perspective, innovatively integrating the PageRank algorithm with the ARIMA-DID model to establish a composite research framework that combines network structure analysis with time series prediction capabilities [2]. Using the PageRank algorithm to analyse the cybercrime networks of various countries, the study precisely identifies core risk nodes such as the United States. Based on multi-dimensional indicators including cybersecurity preparedness and case numbers, it categorizes countries into high-, medium-, and low-risk clusters, revealing the intrinsic connections between risk levels and attack scale, loss severity, and defensive capabilities [3]. Additionally, using the ARIMA model and the DID method, the study takes the United States' 2013 cybersecurity policy as a typical case to construct a counterfactual analysis scenario, quantitatively assessing the policy's inhibitory effect on the volume of cyberattacks [4-5]. Through cross-national comparative research, the study further analyses the advantages of low-risk countries like Japan and Australia in terms of policy sustainability and preventive mechanisms. The research findings not only provide a dynamic assessment model for global cybersecurity risk warning but also offer policymakers a quantifiable tool for evaluating governance effectiveness, holding significant theoretical and practical value for improving the international cybersecurity governance system.

2. Patterns and Distribution of Cybercrime

Cybercrime, including hacking, fraud, identity theft, ransomware attacks, and phishing, is a global issue, but its distribution is not uniform across the world. The prevalence and success of cybercrimes vary depending on multiple factors, such as technological infrastructure, regulatory measures, law enforcement capabilities, internet penetration, and economic conditions.

2.1. Network Crime Risk Distribution Modeling

In the field of cybersecurity, in-depth analysis of the source region of an attack is a critical step in understanding and preventing cybercrime. The application of the PageRank algorithm in cybercrime source analysis benefits from its unique link analysis and centrality metrics, which can accurately identify the core nodes in a criminal network and their influence spreading paths. Through iterative computation and denoising techniques, the algorithm can deeply sort out criminal clues and eliminate interfering information. In addition, the algorithm's community detection function helps to distinguish different criminal groups, while its dynamic update feature ensures continuous monitoring of changes in the criminal network. With the comprehensive credibility assessment, PageRank provides a clear direction and prioritization for the investigation work, which makes the task of exposing and combating the source of cybercrime more efficient and smoother.

This paper therefore calculated the PageRank score for each country using the PageRank algorithm, which reflects the relative importance of the country in the cybercrime graph. The advantage of the PageRank algorithm is that it can

identify nodes that have a high-quality of connections despite a small number of connections, which is particularly important in cybersecurity analysis.

Using the PageRank algorithm [6], this paper calculated the PageRank score for each country, which reflects the relative importance of the country in the cybercrime graph. The advantage of the PageRank algorithm is that it identifies nodes that have a small number of connections but a high quality of connections, which is especially important in cybersecurity analysis.

$$PR(p_i) = (1-d) + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)} \quad (1)$$

The results are shown in the Table 1, where a list of countries in descending order of PageRank score reveals the main sources of cybercrime attacks, which may play a key role in cybercriminal activities. This paper initialized each country's PageRank to $1/N$. For each country A, this paper computed its PageRank, this paper iterated through all the countries B that are linked to A, and this paper updated the PageRank value of A according to the above formula. Each point in the PageRank graph represents a country, and the connecting line represents the relationship between the aggressor and the victim country.

Table 1. Country's PageRank score

Country	PageRank Score	Country	PageRank Score
US	0.593	GB	0.043
CA	0.033	IN	0.022
AU	0.015	NZ	0.014
IE	0.0077	DE	0.0071
CN	0.0061	KR	0.0059

This paper found the optimal number of groups by looking at the sum of squared errors using the elbow rule. Then this paper performed an analysis of variance based on the fields, including the results for mean and standard deviation, F-test results, and significance p-value. The p-value of each analyzed item was analyzed to see if it was significant ($p < 0.05$). Table 2 shows that the four variables do not differ significantly between the categories classified by the clustering classification.

Table 2. Clustering results

	Category 1 ($n = 72$)	Category 3 ($n = 63$)	Category 2 ($n = 58$)	F	P
Cybersecurity Readiness (%)	65.89±16.869	63.925±18.555	61.692±18.456	0.883	0.415
Estimated Total Cases	17938.528±9117.02	50976.698 ±9578.519	85570.828±7762.486	931.272	0.000***
Loss Amount (Million USD)	469.598±297.955	504.01±269.969	506.463±283.31	0.354	0.702
Cyber Threat Index	48.558±22.324	47.542±22.214	53.738±20.133	1.42	0.244

2.2. Results of the Classification

The cluster analysis allows this paper to categorize the countries into three categories high, medium, and low risk, categories II, III, and I, respectively. The characteristics of these countries are shown in the Table 3:

Table 3. The characteristics of these countries

Country Type	Number of Cases	Number of Losses	Cyber Threat Index	Cyber Security Readiness
High-Risk Countries	High	High	High	Low
Low-Risk Countries,	Low	Low	Low	High
Medium-Risk Countries	Medium	Medium	Medium	Medium

High-risk countries, such as Russia and the United States, need to strengthen cybersecurity protection measures and improve cybersecurity readiness to cope with a high number of cyberattacks and a high risk of loss.

Medium-risk countries, such as China, Germany need to continuously monitor their cybersecurity posture, optimize their cybersecurity strategies and improve their cybersecurity readiness to reduce potential risks.

Low-risk countries, such as Australia, Japan, need to remain vigilant, continuously monitor cybersecurity threats and ensure the effectiveness of cybersecurity protection measures.

2.3. Distribution Patterns of Cybercrime Eigenvalue Characteristics

Based on the data in the Global Cybersecurity Index 2024, this paper has compiled the scores of countries around the world on the five Measures: Legal, Technical, Organization, Capacity, and Cooperation. This paper will model and combine the existing data to make reasonable predictions and calculations on the success rate, block rate, reporting rate and prosecution rate of cybercrime in each country.

(1) Success rate vs. blocked rate

According to the Global Cybersecurity Index report released annually by the International Telecommunication Union (ITU), it can be inferred that cybercrime is more likely to be successful in countries that perform poorly in terms of legal sophistication, level of technological development, organization, capacity-building and cooperation, and on the contrary, cybercrime has a greater likelihood of being prevented in more developed countries. Therefore, this paper will build a mathematical model about the success or failure of cybercrime with the five Measures of Legal, Technical, Organization, Capacity, and Cooperation, to calculate the success and deterrence rates of cybercrime in each country, and to visualize the distribution characteristics of the data based on the obtained data.

The model was developed as follows:

$$B_i = 0.5 \times L_i + 0.5 \times T_i + 0.5 \times O_i + 0.5 \times C_i + 0.5 \times P_i \quad (2)$$

$$S_i = 100 - (0.5 \times L_i + 0.5 \times T_i + 0.5 \times O_i + 0.5 \times C_i + 0.5 \times P_i) \quad (3)$$

Substituting the data to get the cybercrime success rate and blocked rate.

(2) Reporting and appeal rates

If the country has a strong judicial capacity and the relevant laws are well established, then victims of cybercrime will have greater confidence in recovering their losses through the law, and therefore the reporting rate of cybercrime will be higher. Therefore, this paper collected the Corruption Perceptions Index (CPI) from Transparency International and modeled it by combining it with the legal scores in the GCI.

$$R_i = w_1 \times CPI_I + W_2 \times L_i + c \tag{4}$$

R is the reporting rate, w is the weight, L is the legal score and c is the constant

A high reporting rate usually implies a higher level of public trust in law enforcement agencies and a more efficient judicial system. Also, countries with high reporting rates may have relatively well-developed and transparent judicial systems, leading to fewer cases of dissatisfaction and thus relatively low appeal rates.

Low reporting rates, on the other hand, may hide problems, and only serious or controversial cases may be reported, which are more likely to be appealed. This paper therefore believes that there is a correlation between the appeal rate and the reporting rate.

3. Analysis of the Effects of Policy Implementation

To test the effectiveness of countries' cybersecurity policies, it is not possible to simply compare cybercrime rates before and after policy implementation, as fluctuations in crime rates are affected by a variety of factors. This paper used a double-difference (DID) approach to introduce a control group that is not affected by the policy, to exclude the interference of external factors and more accurately assess the effect of the policy.

To find a control group unaffected by the policy, use a time series ARIMA model to predict the crime rate in the absence of the policy intervention based on the time continuity and regularity of the crime rate.

Based on the previous k-means clustering results, this paper selected high-risk countries (e.g., the U.S., Russia) and low-risk countries (e.g., Australia, Japan) for time series analysis. Due to the prominent representation of cybersecurity issues in the United States, this paper took the United States as an example and focus on analyzing the comparison before and after the implementation of its policies.

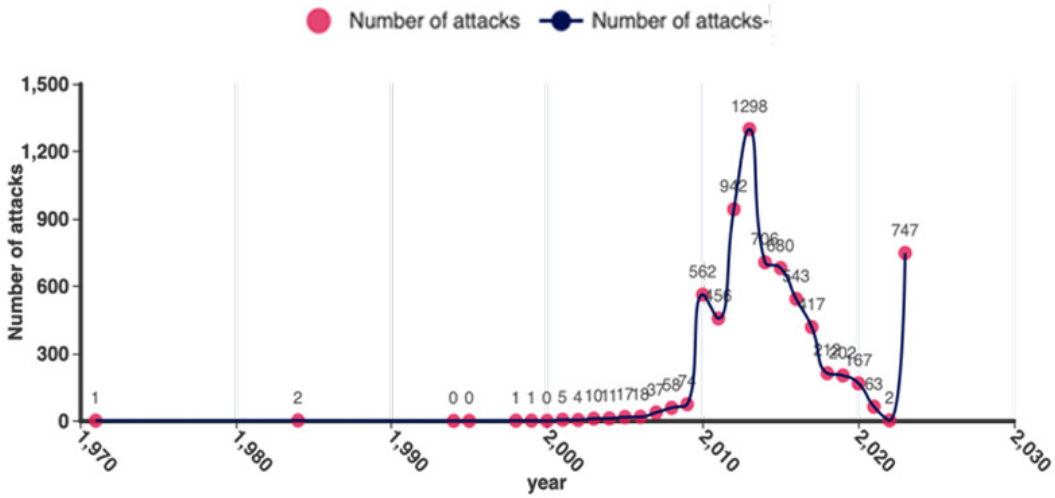

Figure 1. Number of cyberattacks on the United States

As shown in Figure 1, the number of cyberattacks in the United States peaked in 2013 and then declined rapidly between 1970 and 2023.It is hypothesized that the U.S. implemented an effective cybersecurity policy in 2013. Therefore, this paper used the 1970-2013 data as a timeseries sample to predict the cybercrime rate that is not affected by the policy in the next 10 years.

3.1. ARIMA-based Prediction of Crime Rates Without Policy Intervention

The ARIMA model is a statistical analysis tool for predicting future trends, combining autoregressive (AR), differencing (I) and sliding average (MA) components, and is suitable for dealing with trending and seasonal data. In predicting cybercrime rates, the smoothness of the data is tested, the non-smooth data is differenced to make it smooth, and then the model parameters (p,d,q) are determined by the AIC and BIC methods.

(1) Smoothness testing and prediction

The unit root test (ADF) was conducted using SPSSPRO and the results of the test are shown in the Table 4 below for the United States.

Table 4. ADF inspection form

ADF Inspection Form							
Variant	Difference in Order	t	P	AIC	threshold value		
					1%	5%	10%
Year	0	-0.77	0.828	-738.373	-3.889	-3.054	-2.667
	1	-5.126	0.000***		-3.859	-3.042	-2.661
	2	-4.054	0.001***		-3.889	-3.054	-2.667
Note: ***, **, * represent 1%,5%, and 10% significance levels, respectively.							

The results of the test of this series show that based on the variable $Year$, the significance p-value is less than 0.05 at the level of difference of order 1, which presents significance and rejects the original hypothesis that the series is a smooth time series.

(2) Time series ARIMA model

After passing the smoothness test, this paper needs to get the values of model p and model q. This paper used AIC and BIC methods to estimate the parameters p, q.

$$AIC = 2k - 2\ln(L) \tag{5}$$

$$BIC = k\ln(n) - 2\ln(L) \tag{6}$$

In the above equation, k is the number of model parameters, L is the likelihood function of the model, n is the number of samples.

According to the AIC and BIC results, the values of p and q are $(3,0)$ and $(0,0)$, respectively. To avoid p and q being too large, the parameter that minimizes d is chosen, which is finally determined as $\mathrm{p} = 0$ and $\mathrm{q} = 0$. Therefore, the model is ARIMA $(0,1,0)$. To ensure that the parameters were appropriate, this paper performed a residual test. The residuals are normalized to be close to normal distribution, their histograms conform to the ideal noise distribution, and most of the residuals are distributed near the normal line, further indicating that the residuals are close to normal distribution. The ARIMA (0,1,0) model was fitted to the U.S. cybercrime rate data from 1970-2013 to obtain the fitted model. Based on this model, the time variable was shifted back to 10 units to predict the cybercrime rate for the next 10 years. The specific results are shown in Figure 2.

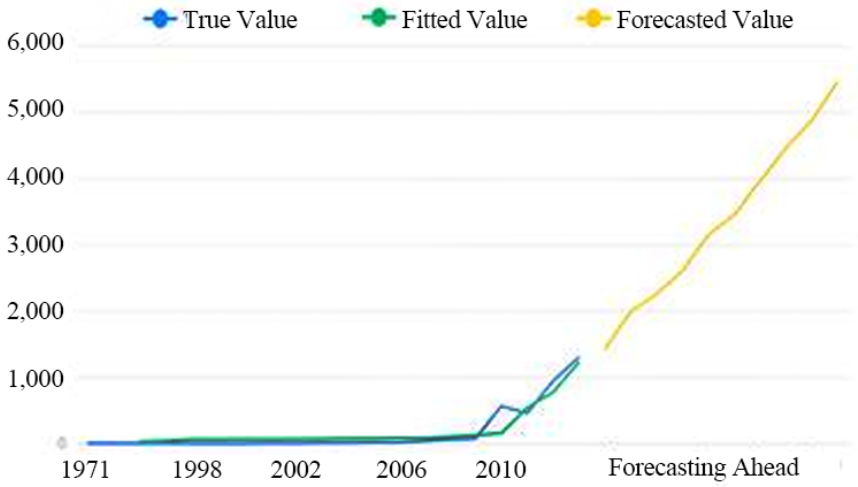

Figure 2. Time series chart

Up to this point, this paper has obtained a control group unaffected by the policy.

3.2. Difference-in-differences based Policy Effect Test

Difference-in-difference is a quasi-experimental method used to assess the effects of a policy or intervention. It estimates the causal effect of a policy by comparing the difference between the changes in the treatment group (the group affected by the policy) and the control group (the group not affected by the policy) before and after the implementation of the policy. The core idea of the difference-in-differences approach is to utilize a natural experimental setup that controls for other potential confounders to more accurately assess the true impact of a policy.

(1) Modeling DID

This paper examined whether there was a significant change in the number of cybercrimes in the United States after the implementation of the 2013 cybercrime policy. Because there are predicted and actual values for data before and after the policy implementation, this paper used a double-difference (DID) approach to assess the net effect of the policy.

$$\mathrm{Diff}_{it} = \alpha + \beta_1 \cdot \mathrm{Post}_t + \epsilon_{it} \tag{7}$$

Where, $Diff_{it}$ is the difference between the true value and the predicted value in a given year; $Post_t$ is the 1 for the year after the policy is implemented and 0 before.

(2) Solution results

According to the model estimation results, $1 = -2709.45, p\text{-value} < 0.001$. this indicates that the actual number of cyber attacks after the policy implementation has been reduced by an average of about 2709 compared to the predicted number, and the results are statistically significant, and the implementation of the policy played a significant role in reducing the number of cybercrimes. The Figure 3 shows the change in the predicted and real values after the policy implementation in 2013.

The green line in the Figure 3 shows the predicted values, while the blue line shows the actual values. After the implementation of the policy in 2013, the number of cyberattacks declined significantly from 2014-2021 but rise again in 2022. This suggests that the policy is effective in the short term, but the long-term effects may be influenced by other factors.

Similarly, this paper get a graph of predicted vs. actual values for Russia, Japan, and Australia.

As shown in Figure 4, it can be seen that the number of cybercrimes in the United States, Russia, Japan, and Australia declined rapidly in 2013, 2014, 2013, and 2016, respectively.

In 2013, the United States implemented Executive Order 13636. Russia has adopted legislation to build a cybersecurity "firewall," emphasizing the protection of cyber sovereignty jurisdiction and developing a cybersecurity risk assessment system.

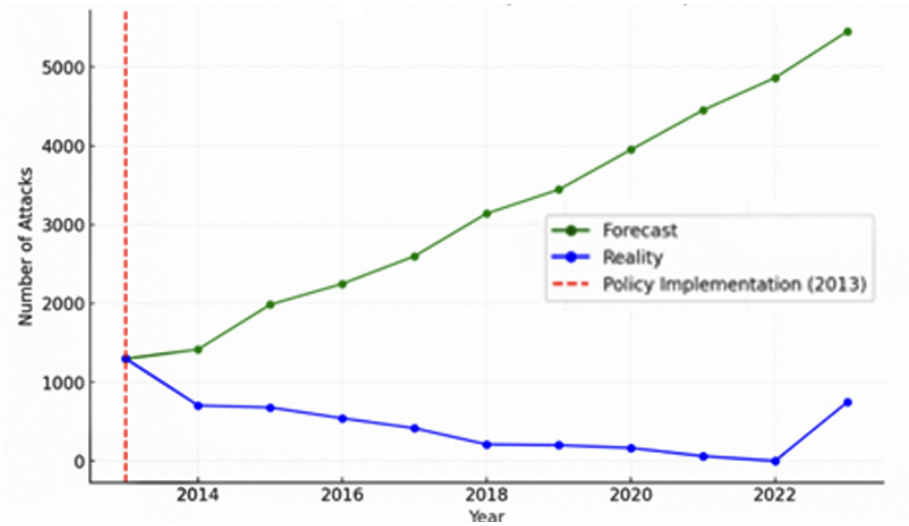

Figure 3. Reality and forecast of the US (2013 and after)

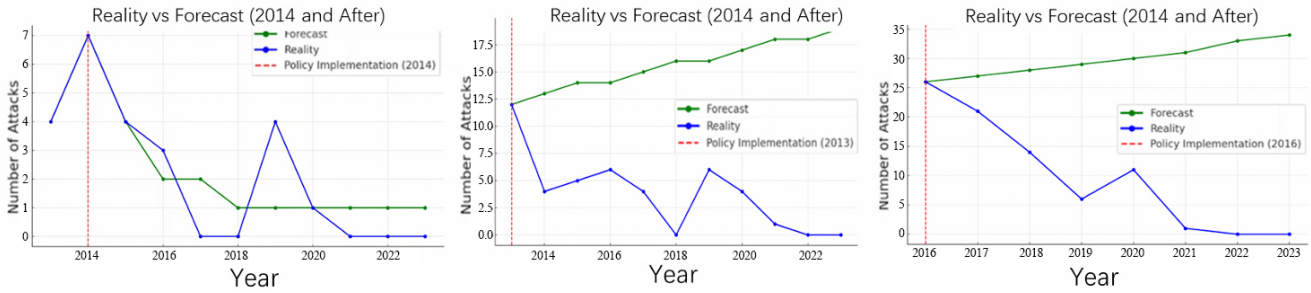

Figure 4. Comparison of projected and real values for three countries

Japan released its Cybersecurity Strategy in June 2013 to build a strong cybersecurity nation through cyber defense capacity building, and government-civilian cybersecurity cooperation. Australia released the Australian Cybersecurity Strategy in 2016 to build a strong "cyber defense network" that can more effectively monitor, block and respond to cyber threats.

Among high-risk countries, the U.S. and Russia's cybersecurity policies are effective in the short term, but their long-term effects are inconsistent, and their policies need to be continually strengthened. In contrast, low-risk countries have more effective policies, particularly Japan, where the number of cyberattacks dropped significantly and remained low after implementation.

High-risk countries emphasize private sector involvement and international cooperation, while low-risk countries focus on government-civilian cybersecurity cooperation. The analysis found that preventive policies, especially vulnerability management systems, showed the best results and that in the area of cybersecurity, preventive inputs are more cost-effective than remediation, with more stable and sustainable results.

3.3. Sensitivity Analysis

To assess the applicability and reliability of the model, this paper used the model to make step-by-step predictions on the original data, i.e., based on the cybercrime rate of the previous N years, this paper used the model to predict the cybercrime rate of $N+1$ years, and then this paper predicted the cybercrime rate of $N+2$ years based on the cybercrime rate of $N+1$ years, until the last year. Taking the United States as an example, the results are shown in the Figure 5.

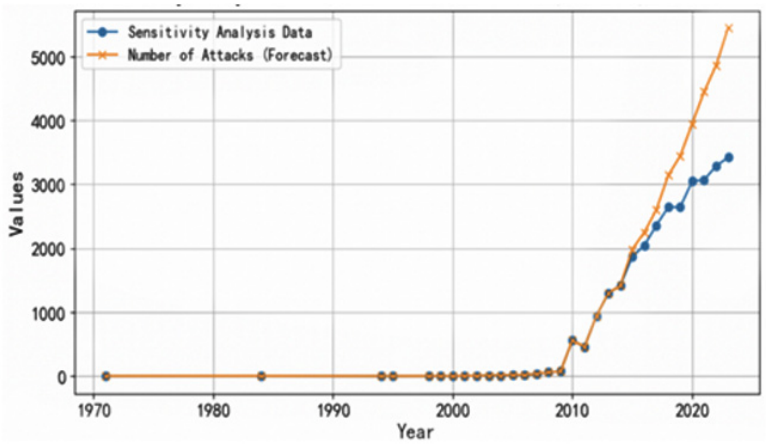

Figure 5. Sensitivity analysis and number of attack (forecast) over years

The various error indicators are listed in Table 5:

Table 5. Error index

Error Index	Value
Mean Square Error	25453.41
Root Mean Square Error	159.54
Mean Absolute Error	134.68
R^2 score	0.73

References

[1] Jin Gaofeng, Zhang Yongding, Zhao Hongyang. Analysis and Forecast of China's Crime Situation in 2023–2024 [J]. Journal of the People's Public Security University of China (Social Sciences Edition), 2024, 40 (03): 1–10.

[2] Zhou Feng. Taxi Demand Forecasting Based on the PageRank Algorithm [J]. Microcomputer Applications, 2019, 35 (04): 8-11+19.

[3] Song Min, Zhang Xueren, Nie Cong. A Study on the Influence of Chinese and American Patents: Based on the PageRank Algorithm [J]. Research on Science and Technology, 2024, 42 (04): 721-732.

[4] Yan Xun, Tie Chengcheng, Yan Wei, et al. Global Temperature Prediction Analysis Based on ARIMA Model and CNN-LSTM Combined Model [J]. Science and Innovation, 2024, (02): 19-22.

[5] Cai Jun, Yang Lan, Zhou Yahong. Current Status and Improvement Methods of PSM-DID in Policy Evaluation [J]. Journal of Management Science, 2024, 27 (02): 30-48.

[6] Shi Fenghao, Wang Xin, Pan Wenlin. Identification of Key Nodes in Hypernetworks Based on an Improved PageRank Algorithm [J]. Information Technology, 2024, (03): 22-27.