Quantification of Cyber Crime Based on Entropy Weight Method and ARIMA-DID

Yijie Jiang [1, †, *], Xinze Li [2, †] and Pinhan Liu [3, †]

[1] School of Automation Science and Electrical Engineering, Beihang University, Beijing, China

[2] School of Physics, Beihang University, Beijing, China

[3] School of Space and Earth Sciences, Beihang University, Beijing, China

* Corresponding Author Email: 22377258@buaa.edu.cn

† These authors also contributed equally to this work

Abstract. This paper proposes an indicator assessment model based on the entropy weight method (EWM) and a combined ARIMA-DID analysis framework, focusing on the quantitative algorithm-driven research on the factors influencing the distribution of crime and the effectiveness of policies. Firstly, an assessment system containing core dimensions such as law and technology is constructed, the internal stopping rate (ISR) is introduced as an indicator of law enforcement effectiveness, and the EWM algorithm is used to objectively assign weights to the data from multiple sources, and the indicators related to the level of the economy and the efficiency of law enforcement are identified as the key influencing factors. Second, the ARIMA algorithm is used to model the time series data to capture the crime rate trend, and the DID (double difference method) is used to construct the policy effect assessment model, and typical cases are used to validate the effectiveness of interventions such as "technical defense enhancement" and to confirm the inhibitory effect of the policy on the crime rate. Finally, the Spearman rank correlation algorithm is used to analyze the correlation between the number of crimes and economic and technological characteristics, and to confirm the core influence of the level of economic development. Through the synergy of multiple algorithms, this study provides an algorithm-driven analytical paradigm covering indicator empowerment, trend prediction and policy evaluation in related fields.

Keywords: Spearman correlation analysis; ARIMA; DID; quantitative research; EWM.

1. Introduction

This paper focuses on the quantitative analysis of the influencing factors of global cybercrime distribution and prevention and control policies, aiming to reveal the intrinsic correlation between the crime phenomenon and the dimensions of economy, technology, and law through the multi-algorithm fusion framework [1]. First, an assessment system containing legal, technical and organizational measures is constructed based on the ITU Global Cybersecurity Index (GCI), and the internal stopping rate (ISR) is introduced to optimize the assessment of law enforcement effectiveness, and the multi-dimensional indicators are objectively empowered through the entropy weighting method (EWM) to identify the key influencing factors [2][3]. Second, combining the time series prediction ability of autoregressive integrated moving average model (ARIMA) and the causal inference advantage of double-difference method (DID) [4], select typical countries as cases to verify the actual effects of intervention strategies, such as "strengthening technological defenses" and "perfecting legal frameworks". Select a typical country as a case study to verify the actual effects of intervention strategies such as "strengthening technological defense" and "improving legal framework" [5]. Finally, the Spearman rank correlation algorithm is used to analyze the strength of the association between the number of crimes and demographic characteristics (e.g., GDP, Internet penetration), and to quantify the role of economic and social factors on the distribution of crimes [6].

The experimental results show that the economic level (characterized by $ln(GDP)$) and the efficiency of domestic law enforcement (ISR) have a significant impact on the number of cybercrimes, the ARIMA-DID model is effective in evaluating the actual effects of policy interventions, and the distribution of crimes is strongly correlated with the penetration of digital infrastructure.

2. Analysis of Crime Influencing Factors Based on Entropy Weight Method

2.1. GCI- and EWM- Based Screening of Crime Distribution Influence Factors

To analyze the distribution of cybercrime, the ITU's Global Cybersecurity Index (GCI) was introduced to assess national and regional commitment and capacity in cybersecurity. The GCI assessment framework is based on five core areas. Legal measures: whether laws and regulations related to cybersecurity are in place:

1. Legal measures: whether laws and regulations related to cybersecurity are in place.
2. Technical Measures: Whether technical capabilities and institutions are in place.
3. Organizational measures: Whether there is a clear national cybersecurity strategy and whether a dedicated coordinating body has been established to implement cybersecurity policies.
4. Capacity building: whether education and training on cybersecurity technology and knowledge are being conducted.
5. International cooperation: whether the country/region participates in international cybersecurity cooperation and agreements.

Other statistical indicators such as national GDP, annual education expenditure, Internet penetration rate and urban population rate are also introduced.

Based on the statistics collected on cybercrime incidents in countries around the globe, key observations can be made:

Regions with high rates of cybercrime are usually associated with high GDP and advanced digital infrastructure. As economically developed countries are more dependent on technology and the Internet, they tend to suffer more from cybercrime, which makes them attractive targets.

In countries/regions such as SA, GA, RS and FI, the success rate of cybercrime was high, approaching or exceeding specific thresholds (e.g. 0.9343%). In contrast, the success rates of cybercrime in countries such as MC, BW, UM and VN were significantly lower, remaining significantly below the threshold (e.g. 0.9343%).

To calculate the crime thwarting rate and crime reporting rate, statistics are provided. The “Discovery Method” field provides information on whether a cybercrime was discovered internally, reflecting effective cybersecurity measures and thwarted incidents. The cybercrime success rate is defined as the ratio of incidents discovered externally or through unknown methods to the total number of cybercrime incidents.

Based on the above results, the following conclusions can be tentatively summarized:

(1) Cybercrime tends to be more serious in developed countries with adequate economic and technological environments.

(2) Cybercrime shows a radiation effect among neighboring countries, indicating that international cooperation has an impact on cybercrime.

(3) The effectiveness of domestic law enforcement also affects cybercrime.

Based on the analysis, selected 4 dimensions of law, international cooperation economy, and technology as the indicators affecting cybercrime in each country.

2.2. Quantification of Crime Impact Factor Weights under the EWM Algorithm

The Cybersecurity Index includes quantifiable indicators such as Legal Measures (LM), Technical Measures (TM), and Organizational Measures (OM).

Legal measures can only assess the existence of relevant laws in the country and do not reflect the strength of domestic7 law enforcement. Therefore, a new indicator called the Internal Stopping Rate (ISR) was introduced.

$$ISR = \frac{\text{The number of cybercrimes discovered through internal legal measures}}{\text{The total number}} \tag{1}$$

The ISR provides a better description of the effectiveness of the implementation of legal measures than the LM itself. Since each country has a significantly higher GDP value than the other indicators,

and since the number of cybercrimes in countries/regions with a high GDP is much higher than in those with a lower GDP, the relationship between GDP and cybercrime is better represented by an exponential rather than a linear relationship. Use $\ln(\text{GDP})$ as a linear factor. The final linear equation is given as follows:

$$CRN = \beta_0 + \beta_1 LM + \beta_2 TM + \beta_3 OM + \beta_4 ISR + \beta_5 \ln(\text{GDP}) + \text{ò} \tag{2}$$

β_0 : The constant term (intercept), representing the baseline value of CRN when all variables are zero.

β_i : The regression coefficients for each variable.

ϵ : The error term, representing factors not included in the model or random noise.

To further determine the corresponding coefficients, the entropy weight method (EWM) is introduced for analysis.

The entropy weight method (EWM) is an objective approach for determining the weights of multiple indicators in a comprehensive evaluation. First, a decision matrix X is constructed, where each row represents an evaluation object and each column corresponds to an indicator:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \tag{3}$$

The entropy value e_j of each indicator is then calculated using the formula:

$$e_j = -k \sum_{i=1}^{m} p_{ij} \ln\left(p_{ij}\right), \tag{4}$$

Where $p_{ij} = \dfrac{x_{ij}}{\sum_{i=1}^{m} x_{ij}}$ represents the proportion of the i-th evaluation object on the j-th indicator, and $k = \dfrac{1}{\ln(m)}$. is a constant that ensures the entropy falls within the range $[0,1]$. Based on the calculated entropy values, the weight w_j for each indicator is computed as:

$$w_j = \frac{1 - e_j}{\sum_{j=1}^{n} \left(1 - e_j\right)} \tag{5}$$

A lower entropy value e_j indicates that an indicator provides more valuable information, thus receiving a higher weight. Finally, the comprehensive score S_i for each evaluation object is calculated as:

$$S_i = \sum_{j=1}^{n} w_j x_{ij} \tag{6}$$

The calculation results of EWM are shown in the Table 1.

Table 1. EWM result.

Factors	Information Entropy Value	Information Utility Value	Weight
LM	0.971	0.029	10.633%
TM	0.957	0.043	15.643%
OM	0.972	0.028	10.199%
Ln(GDP)	0.916	0.084	31.057%
ISR	0.912	0.088	32.467%

Based on the results, it can be seen that $ln(GDP)$ has a weight of 31.057%, which indicates that it has a significant effect on the number of cybercrimes in a country. That is, the more economically developed a country is, and the more likely it is to have cybercrime incidents. Therefore, it is crucial to strengthen the public's awareness of prevention.

The weight of ISR is 32.467% while LM is only 10.633%, which shows that it is not enough to have cybersecurity laws in place, but it is also crucial to ensure their effective implementation.

TM Has a weight off technological advancement and the amount of cybercrime.

OM Has the lowest weight of 10.199%, which could be attributed to the challenges of international cooperation in combating cybercrime. Such efforts require coordinated action by multiple Governments and are often less efficient than domestic law enforcement.

2.3. Presentation of Cybersecurity Theory

In the analytical summary, four cybersecurity theories are proposed:

1. Enhance the reporting and awareness of cybercrime in economically developed regions.
2. Establish clear legislative frameworks to define the roles, responsibilities, and responses of governments.
3. Advance technological solutions to enable precise tracking and intervention against cybercrime.
4. Swift responses to emerging cybercrime threats, with a focus on fostering international cooperation to address transnational cybercrime.

3. Policy Effectiveness Assessment Based on ARIMA-DID Modeling

3.1. ARIMA-DID Combined Model Construction

To prove the validity of the theory, a representative country (Singapore) is studied. The cybercrime rate trends are analyzed to illustrate the impact of these policies.

Short-term forecasting using a non-seasonal ARIMA model combines the properties of autoregressive (AR), integrated (I), and moving average (MA) components to model complex time series data. The autoregressive component describes the current value y_t as a linear combination of its p previous lagged values, expressed as:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t \tag{7}$$

Where ϕ_i are the autoregressive coefficients, and ϵ_t is the white noise error term with a mean of zero and variance σ^2. The differencing operation eliminates trends or nonstationarity by subtracting the previous value, defined as:

$$y_t' = y_t - y_{t-1} \tag{8}$$

For more complex trends, second-order differencing is applied, such as $y_t'' = y_t - 2y_{t-1} + y_{t-2}$.

The complete ARIMA model is expressed as:

$$\phi(B)(1-B)^d y_t = \theta(B)\epsilon_t \tag{9}$$

B Is the backshift operator, $(1-B)^d$ denotes the d-order differencing, and $\phi(B)$ and $\theta(B)$ are polynomials representing the AR and MA components, respectively. Once the model is fitted, future values are predicted using the formula:

$$\hat{y}_{t+h} = \mu + \sum_{i=1}^{p} \phi_i \hat{y}_{t+h-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t+h-j} \tag{10}$$

The core idea of Difference-in-Differences (DID) model is to compare the changes in outcomes between a treatment group (affected by the policy) and a control group (not affected by the policy) before and after the policy implementation. The standard DID model is expressed as:

$$Y_{it} = \alpha + \beta \cdot D_t + \gamma \cdot G_i + \delta \cdot (G_i \cdot D_t) + \epsilon_{it} \quad (11)$$

Y_{it} Represents the outcome variable for individual i at time t, G_i is a group indicator (1 for the treatment group, 0 for the control group), D_t is a time indicator (1 for post-policy, 0 for pre-policy), and $G_i \cdot D_t$ is the interaction term representing the treatment group after the policy implementation. The coefficient α is the baseline outcome for the control group pre-policy, β captures the time effect (changes over time in the control group), γ reflects the group effect (baseline differences between groups), and δ is the DID estimator that measures the policy's causal effect. The residual ϵ_{it} captures unobserved factors.

To compute the effect, the mean outcomes are compared across groups and time periods, this is calculated as:

$$\delta = (\Delta Y_{\text{treatment}} - \Delta Y_{\text{control}}) \quad (12)$$

$\Delta Y_{\text{treatment}}$ And $\Delta Y_{\text{control}}$ represent the changes in the treatment and control groups, respectively. The DID estimator δ indicates the policy's net effect after accounting for common time trends and group-specific differences. It is therefore possible to determine whether the policy has had a positive impact, a negative impact or a negligible impact, thus providing strong evidence of the effectiveness of the policy.

3.2. Validation of ARIMA-DID-based Policy Effectiveness Algorithm

As a result, the significance p-values and coefficients of the DID model analysis are presented in the Table 2 below. When the significance P -value is less than 0.05, the policy is considered effective; otherwise, it is considered ineffective. The coefficient can be used to determine whether the policy impact is positive or negative.

Table 2. DID result.

Nation	p-values	coefficients	Policy effective time
SG(POS)	0.014^{**}	-2.45	2020
SG(NEG)	0.088^{*}	1.3	2014

For Singapore, consider both positive and negative examples. In 2014, the “Smart Nation” was proposed. But Singapore's cybercrime incidents are increasing every year.

It has not made enhancing public cybersecurity education a primary goal. In addition, the allocation of responsibility for cybersecurity is not clear and the legal framework to address cybercrime is not up to date. As shown in Fig. 1 (a), the blue line indicates the raw data. This is not consistent with the first and second points of cybersecurity theory.

DID analysis also shows a significance P-value of $0.088*$, which is not significant at this level, leading to the failure to reject the null hypothesis. The coefficient is 1.3, suggesting that the policy intervention was ineffective but positive in direction.

In 2018, Singapore enacted the Cybersecurity Act, equipping the nation with a dedicated cybersecurity agency-the Cyber Security Agency (CSA). By 2020, Singapore introduced the Singapore Cybersecurity Strategy 2020, which aimed to identify and analyze malicious activities in cyberspace at the national level. The strategy also emphasized improving user literacy and cultivating cybersecurity habits. These measures are almost entirely consistent with the components of cybersecurity theory.

According to DID analysis, the significance P-value is 0.014^{**}, indicating significance at this level, leading to the rejection of the null hypothesis. The coefficient is -2.45, suggesting that the policy intervention was effective but negative in direction, as shown in the Fig. 1 (b).

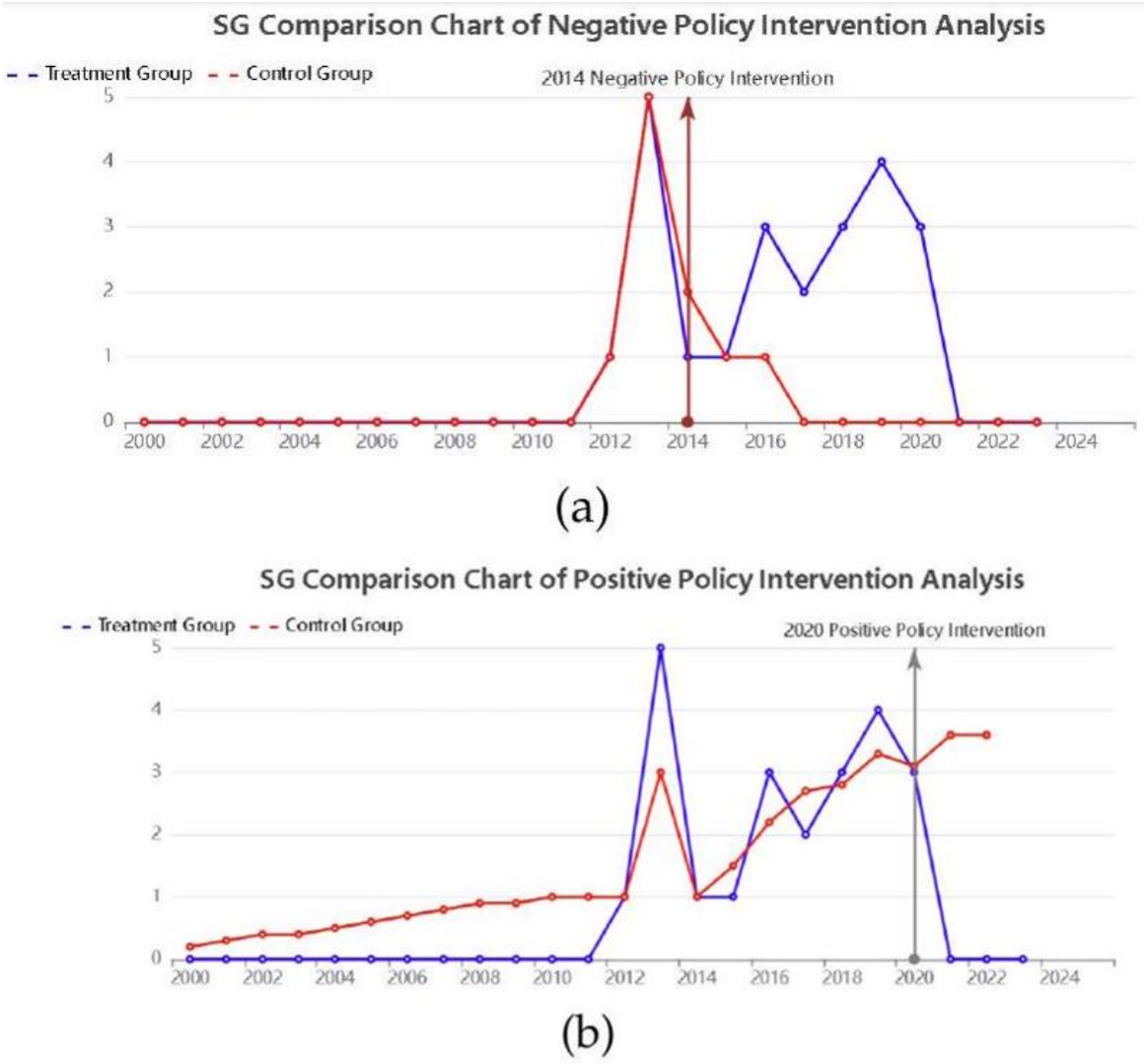

Fig 1. DID result of SG.

4. Criminal Association Profiling Based on Spearman's Algorithm

4.1. Establishing the Model

To explore the relationship between demographic characteristics and global cybercrime distribution, Spearman correlation analysis is performed. The Spearman correlation coefficient (ρ) is calculated as follows:

$$\rho = 1 - \frac{6\sum d_i^2}{n\left(n^2 - 1\right)} \tag{13}$$

Where:
d_i Is the difference between the ranks of each pair of variables?
n Is the number of paired observations?
The steps to compute the Spearman correlation are:
Compute the rank for each variable (x_i for cybercrime distribution, y_i for demographic features).
Calculate d_i as the rank difference for each pair $d_i = \text{Rank}(x_i) - \text{Rank}(y_i)$.
Square the differences and sum them $\sum d_i^2$.

Substitute the values into the Spearman formula to compute ρ, where $\rho = 1$: Perfect positive correlation $\rho = -1$: Perfect negative correlation $\rho = 0$: No correlation.

4.2. From Data to Insights: Modeling Results

To calculate the relationship between demographic data and cybercrime, data on several demographic characteristics were collected, including Internet penetration, level of urbanization, percentage of education expenditures (used as an indicator of educational attainment), and national GDP. In addition, internal cybercrime deterrence rates collected during the cybercrime analysis, as well as data on the distribution of cybercrime, were also collected.

Correlation analysis using the Spearman model yielded the correlation coefficients summarized in Fig. 2 of the table below:

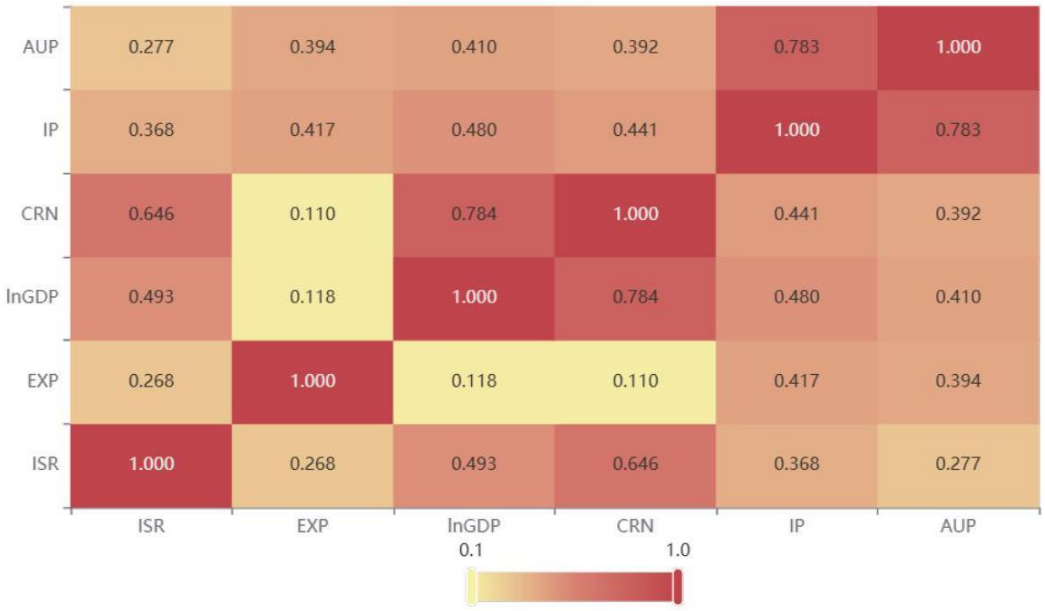

Fig 2. Spearman correlation heatmap.

4.3. Correlation Insights

Based on the correlation heatmap, the relationships between Cybercrime Number (CRN) and demographic factors are as follows:

1. Strong Positive Correlations:

CRN and GDP ($lnGDP$): Higher levels of GDP, with a correlation of 0.784, are strongly associated with increased cybercrime.

CRN and Internal Stopping Rate (ISR): a correlation of 0.646 shows that better detection systems can reduce the impact of cybercrime.

2. Weak Positive Correlations:

CRN and Internet Penetration (IP): The correlation of 0.441 indicates that increased Internet access increases the risk of cybercrime.

CRN and level of urbanization (AUP): with a correlation of 0.392, urbanization has a limited impact, but higher connectivity in urban areas increases vulnerability.

CRN and Education Level (EXP): a correlation of 0.118 indicates that education level has the least direct impact on cybercrime rates, while digital literacy has a limited indirect impact.

5. Conclusion

This paper proposes a multi-algorithm fusion framework based on entropy weight method (EWM), ARIMA-DID model and Spearman correlation analysis, aiming to reveal the key influencing factors of crime distribution and the actual effects of policy interventions through quantitative analysis. The

framework integrates multidimensional data, and is capable of objective weighting of indicators, time series prediction, causal inference of policies, and correlation analysis, providing a systematic algorithm-driven paradigm for crime research. First, the EWM model calculates the weights of core indicators, such as legal measures and technical measures, and identifies economic level (ln (GDP)) and domestic law enforcement efficiency (ISR) as the main influencing factors of the number of crimes, which realizes the scientific ranking of the importance of multiple indicators. Second, the ARIMA-DID combined model accurately quantifies the inhibitory effect of interventions such as "technological defense enhancement" on crime rates (e.g., the effect is significant after the implementation of a specific policy) by separating the time trend from the policy effect, which provides statistically rigorous evidence for the evaluation of policy effectiveness. Finally, Spearman correlation analysis reveals the strong correlation between crime distribution and GDP, Internet penetration rate and other characteristics, which further validates the centrality of economic and technological factors in crime formation. Future research can further expand the spatial dimension of the model, incorporate more regional data to enhance the generalizability of the conclusions, or combine machine learning algorithms to optimize the prediction accuracy.

References

[1] Liu Donghua. An empirical test of inflation targeting to "anchor" inflation expectations and its policy implications for China [J]. Economic Science, 2009, (05): 19-32.DOI: 10.19523/j.jjkx.2009.05.002.

[2] Jiao Xuejun, Wang Wenjie, Hu Yinglei, et al. Evaluation of Geological Environmental Hazards on General Highways Using Entropy Weight Method [J]. Geospatial Information, 2025, 23(05):129-131.

[3] Wang Caixia. Forecasting disposable income of urban and rural residents in Haikou city and analyzing influencing factors [D]. Hainan Normal University, 2022.DOI: 10.27719/d.cnki.ghnsf.2022.000133.

[4] Zhong Cheng. Research on reservoir dam safety prediction method based on ARIMA [J]. Water Resources Science and Cold Region Engineering,2025,8(04):111-114

[5] Lin Xiushui. Research on the empowering effect of digital China construction on enterprises' new quality productivity--a quantitative analysis based on multi-temporal double difference model [J]. Journal of Southwest University (Social Science Edition), 2025, 51(01):148-164+304.DOI:10.13718/ j. cnki.xdsk. 2025.01.012.

[6] Pan Ruiping, Yang Hua, Xin Boxiang, et al. Early warning method for monitoring the risk of power outage in station area based on Spearman's correlation coefficient [J]. China New Technology and New Products, 2025, (05): 137-139.DOI: 10.13612/j.cnki.cntp.2025.05.026.