ISSN: 2664-9640

Research on Cybercrime based on K-means Clustering Algorithm and Linear Regression Modelling

Jingcheng Yang[†,*], Zeyu Wu[†], Luming Wang[†], Hongkai Zeng[†]

Manchester Metropolitan Joint Institute, Hubei University, Wuhan, China

[†]These authors also contributed equally to this work

[*]Corresponding author: 2324837899@qq.com

Abstract

This paper focuses on research related to cybercrime governance, aiming to analyse the distribution of cybercrime globally, the factors influencing it and the role of different policies in curbing it. Through data collection and visual analysis, significant differences were found between countries on five dimensions of cybercrime governance, namely legal, technological, organisational, capacity development and cooperation measures. The K-means clustering algorithm was used to classify countries into four risk clusters, which showed that some African countries were in the high-risk group, while developed countries were mostly in the low-risk group. After Pearson correlation analysis, a strong positive correlation was observed among the governance measures. Linear regression modelling using the least squares method, with cybercrime incidence as the dependent variable and policy scores and pattern scores as the independent variables, yielded that both pattern scores and policy scores were negatively correlated with cybercrime incidence, with pattern scores having a stronger impact. The study shows that most developing countries are less capable of curbing cybercrime, while developed countries have more comprehensive measures, and that cybercrime rates are negatively correlated with response scores, findings that provide an important basis for the formulation of cybersecurity policies.

Keywords

Cybercrime; Global Distribution; Policy and Pattern; Demographic Characteristics.

1. Introduction

In the rapid development of science and technology, intelligent transport systems, with advanced vehicle dynamic control and data analysis technology, are gradually changing the way people travel, significantly improving travel efficiency and safety. Relying on real-time data collection and processing, these technologies accurately achieve personalised trajectory prediction and risk avoidance, injecting new vitality into urban transport [1].

However, technological advances have introduced new risks while bringing convenience. With the popularity of ITS, the cybersecurity challenges it faces are becoming increasingly severe. Cybercrime methods are endless, and attacks on ITS are harmful. By attacking the data transmission link, malicious elements can arbitrarily manipulate system functions and interfere with vehicle decision-making, which can lead to serious security risks and even threaten the normal operation of the city and public safety. For example, interfering with the navigation commands of self-driving vehicles may cause the vehicles to deviate from the established routes, endangering the lives of passengers and disrupting the traffic order [2].

In the face of such a serious situation, it is urgent to integrate cybersecurity measures into the design of intelligent transport solutions and build a complete and sustainable security

ISSN: 2664-9640

framework. This is not only the key to safeguard data integrity and privacy, but also a necessary condition to promote the healthy and sustainable development of intelligent transport technology, which is of far-reaching significance to the maintenance of urban transport stability and security.

2. Global Distribution of Cybercrime and Factors Affecting It

2.1. Data Processing

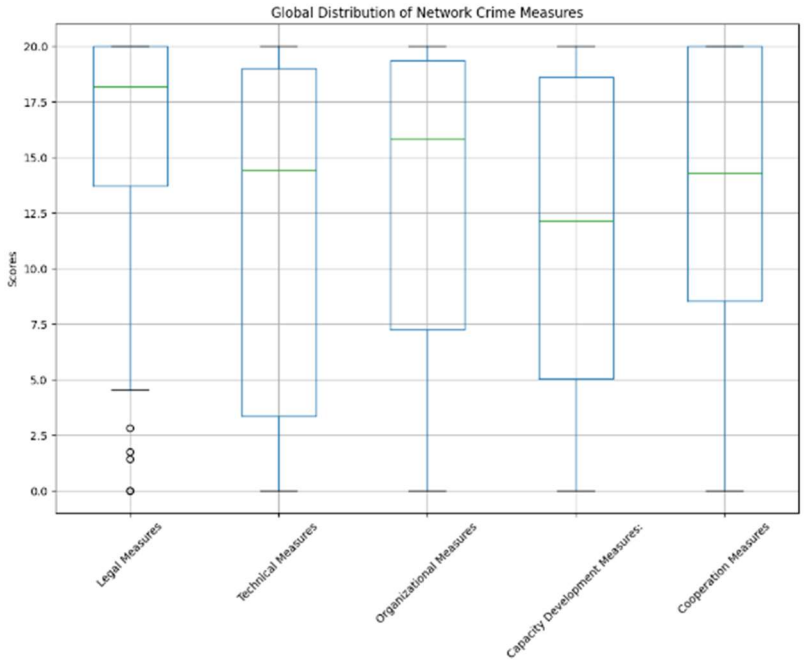

Figure 1. Data visualization

Figure 1 shows a visualisation of the data we have collected, which gives an intuitive picture of the distribution of scores for each country on the five main dimensions:

Legal Measures: The median score for countries is 18.18, with a maximum of 20 and a minimum of 0, indicating that the legal measures in most countries are relatively sound. Although there are still a few countries with almost zero investment in legal aspects.

Technical Measures: The median score for this dimension is 14.44, with a relatively even distribution. The maximum score is 20, and the minimum is 0, which indicates that many countries still face deficiencies in technical measures.

Organizational Measures: Similar to technical measures, the distribution is also quite broad, with a median score of 15.81, a minimum of 0, and a maximum of 20, indicating significant disparities in organizational management capabilities.

Capacity Development Measures: The median score is 12.12, the lower standard deviation (6.98) suggests that scores for capacity development are relatively concentrated among countries.

Cooperation Measures: The median score for cooperation is 14.28, close to the maximum of 20. However, there are still countries that score near zero in international cooperation.

These statistics indicate that most countries have implemented certain measures for the prevention and governance of cybercrime. However, significant gaps remain, especially in Africa and other developing countries, which may face greater challenges.

2.2. Global Distribution of Cybercrime and Corresponding Influencing Factors

Since global-scale clustering analysis must be conducted on comprehensive datasets related to cybersecurity governance, multiple experimental approaches were initially tested, including

ISSN: 2664-9640

hierarchical clustering algorithms and other clustering techniques. After evaluating the performance and applicability of these methods, K-means clustering was selected as the primary technique for further analysis due to its computational efficiency and suitability for partitioning large datasets [3].

In this context, the K-means algorithm is applied to categorize countries based on their measures for combating cybercrime, specifically across five key dimensions. The objective of this clustering process is to divide the countries into four risk-based groups (i.e., K=4), reflecting their relative effectiveness and priorities in cybercrime governance.

Formally, given a dataset with n observations, where each observation represents a dimensional real vector of indicators, K-means clustering partitions the observations into K clusters (S1, S2, …, SK) in such a way that the within-cluster sum of squares (WCSS) is minimized. The optimization objective can be expressed as:

$$\arg\min \sum_{i=1}^{k} \sum_{x \in S_i} \left\| x - \mu_i \right\|^2 = \arg\min \sum_{i=1}^{k} |S_i| Var S_i \tag{1}$$

where μ_i is the mean of points in S_i.According to the global average value of each of the five indicators. This is equivalent to minimizing the pairwise squared deviations of points in the same cluster:

$$\arg\min \sum_{i=1}^{k} \frac{1}{2|S_i|} \sum_{x,y \in S_i} \left\| x - y \right\|^2 \tag{2}$$

Equivalence can be deduced from identity:

$$\sum_{x \in S_i} \left\| x - \mu_i \right\|^2 = \sum_{x \neq y \notin S_i} (x - \mu_i)(\mu_i - y) \tag{3}$$

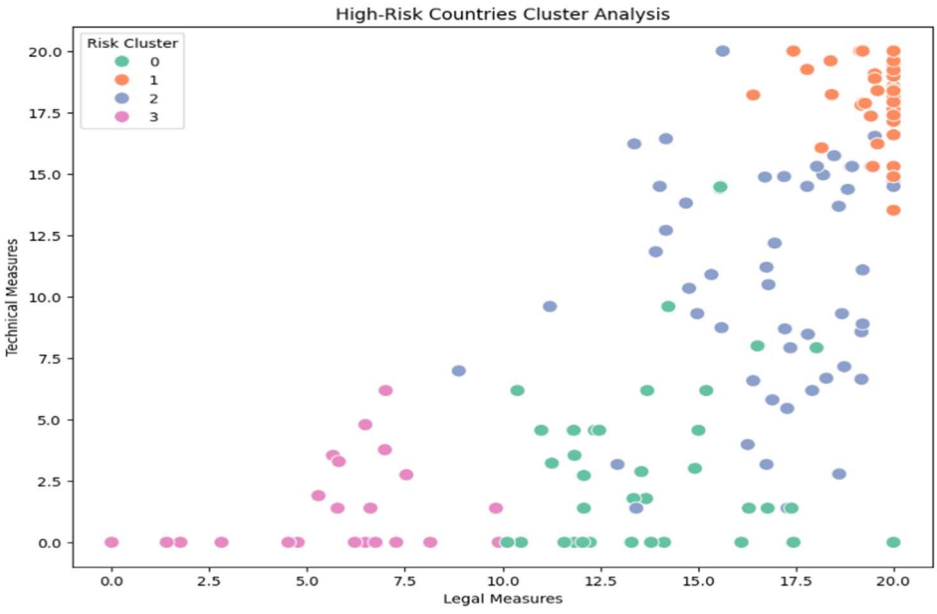

Figure 2. Cluster analysis of high-risk countries

ISSN: 2664-9640

Through this equivalence, the clustering process inherently maximizes the variance between points in different clusters, thereby ensuring distinct separation among the groups. The resulting classification of countries into four clusters provides critical insights into their risk levels and priorities in addressing cybercrime, offering a basis for targeted policy recommendations to improve global cybersecurity measures.

Since the total variance is constant, this is also equivalent to maximising the squared deviation between the midpoints of the different clusters, and the analysis is shown in Figure 2.

Using K-mean clustering, we classify countries into four risk clusters. Some of the results are shown in Table 1:

Table 1. Risk Cluster classification

Country	Risk Cluster
Angola	0
Benin	1
Botswana	2
Burkina Faso	2
Burundi	3

Risk Cluster 0: The governance of cyber - crimes in these countries may be relatively weak, or they lack effective measures.

Risk Cluster 1 - 3: As the scores increase, the governance capabilities of countries in terms of law, technology, organization, etc., gradually strengthen. The differences in clustering provide valuable information for subsequent in - depth analysis.

Judging from the clustering results, some African countries (such as Angola) are in higher - risk groups, while some developed countries are in lower - risk groups.

2.3. Model Solving

Pearson Correlation Analysis

The cybersecurity scores of each country in the five main pillars will be used as X and Y sets for correlation analysis [4].

(1) Substitute the evaluation scores of the five major pillars into the following formula to calculate the Pearson correlation coefficient.

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}. \quad (4)$$

(2) The correlation coefficient r<0.05 obtained indicates that there is a correlation between X and Y. The sign and degree of the correlation coefficient are analysed and obtained as shown in Figure 3:

ISSN: 2664-9640

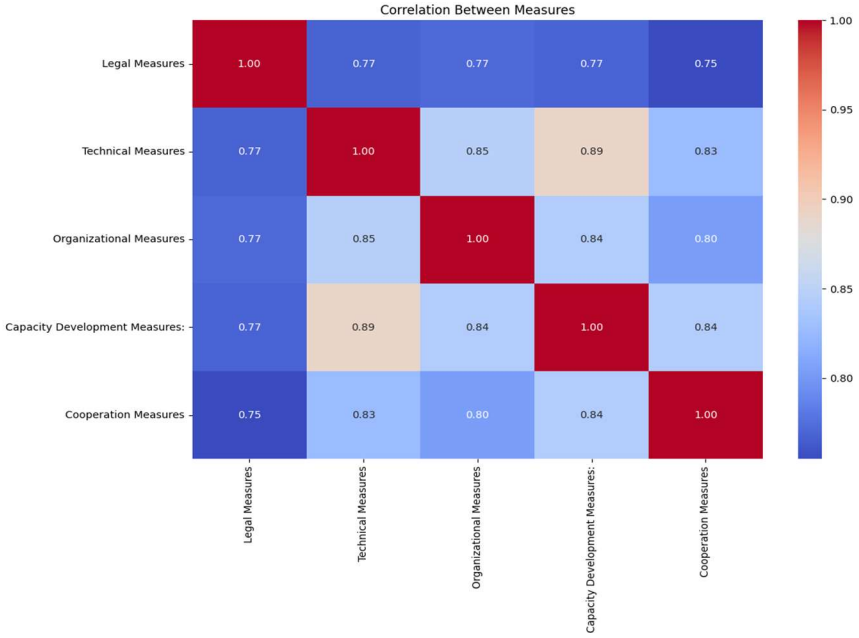

Figure 3. Heatmap of scoring results

The correlation matrix reveals the relationships between different measures:

Legal Measures vs Technical Measures: The correlation score is 0.77, indicating a strong positive correlation between legal measures and technical means. This implies that legal and technical aspects often complement each other, with countries that have strong legal protections also tending to invest more in technical measures.

Technical Measures vs Organizational Measures: There is a high positive correlation (0.85), suggesting that the enhancement of technical measures is often accompanied by improvements in organizational management.

Cooperation Measures vs Other Measures: All measures (legal, technical, technical, organizational, capacity development) show a certain degree of positive correlation with cooperation measures, particularly with capacity development measures (0.84), which have a high correlation with cooperation measures.

Regional averages are shown in Figure 4.

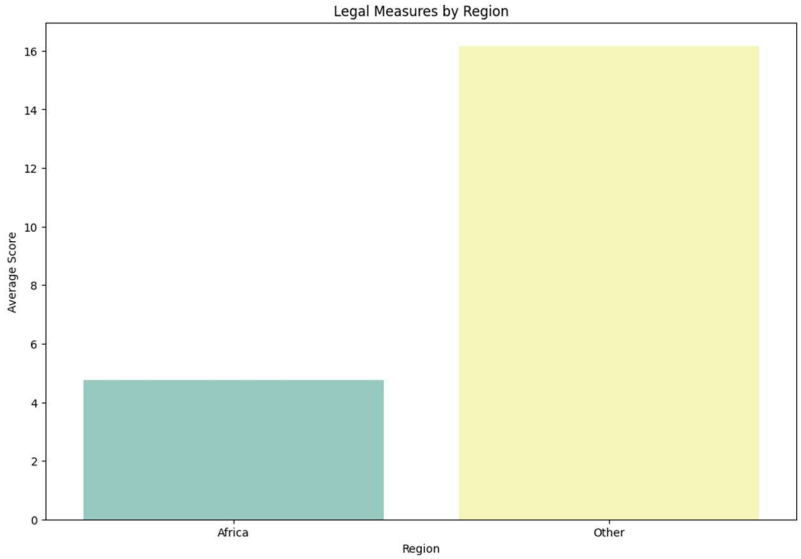

Figure 4. Average score by region

ISSN: 2664-9640

Countries were grouped by region (e.g., regions such as Africa) and the average score for each region was calculated, as shown in Table 2:

Table 2. Score for each measure

Region	Legal Measures	Technical Measures	Organizational Measures	Capacity Development Measures:	Cooperation Measures
Africa	4.76	0.00	0.00	0.00	0.00
Other	16.16	11.47	13.52	11.48	13.55

Africa: The scores of African countries are generally low, especially in terms of legal, technical, organizational, and capacity development measures. The average values of almost all indicators are close to 0, indicating that African countries are very weak in dealing with cyber - crimes. Specifically, African countries generally lack legal frameworks, technical support, and organizational management capabilities. This may make cyber - crimes more likely to occur in these regions and difficult to effectively prevent.

Other regions (such as developed countries): In contrast, the scores of other regions are significantly higher. Especially in terms of legal and technical measures, the average scores are close to or reach 20, indicating that these countries have stronger measures and systems in cyber - crime governance.

Through the analysis, it is found that most developing countries have relatively weak capabilities in curbing cyber - crimes, while most developed countries have more comprehensive countermeasures against cyber - crimes. By combining the cyber - crime rates in Africa and developed countries, it can be concluded that there is a negative correlation between the incidence of cyber - crimes and the scores of countermeasures. The more comprehensive the countermeasures, the lower the cyber - crime rate is usually.

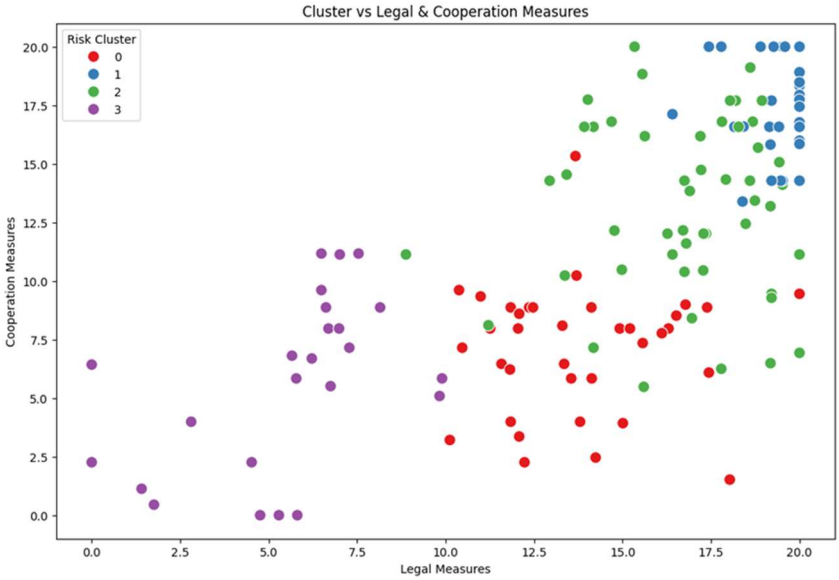

Figure 5. Cluster vs Legal & Cooperation Measures

Figure 5 shows that several key patterns can be identified:

Significant Regional Disparities: African countries generally score significantly lower across key dimensions, while other regions, particularly developed countries, tend to achieve higher scores. The lack of legal frameworks, technical measures, organizational capacity, and capacity

ISSN: 2664-9640

development initiatives make regions like Africa more vulnerable to becoming hotspots for cybercrime.

Positive Correlation Among Measures: There is a strong positive correlation between various measures, indicating that countries often adopt comprehensive strategies to combat cybercrime. Improvements in one area, such as legal frameworks, are frequently accompanied by advancements in other areas like technical and organizational measures.

Clustering Results Highlight Risk Groups: The clustering analysis enables the categorization of countries into risk groups, providing targeted insights. For high-risk countries, such as those in Africa, specific policy recommendations can be proposed to strengthen their legal, technical, and organizational capabilities. By enhancing these areas, the incidence of cybercrime can be effectively reduced.

3. Effective Models of Different Policies to Curb Cybercrime

Through modeling, we hope to obtain the impact of policy scores and pattern scores on the cyber - crime rate. Since the regression equation obtained by the least - squares method has good interpretability. We can understand the degree of influence of independent variables on the dependent variable through the regression coefficients, make inferences and predictions. Therefore, we choose the least - squares method in linear regression [5].

For binary linear regression, let the dependent variable be y, and the two independent variables be x_1 and x_2 respectively. We have n groups of sample observations $(x_{i1}, x_{i2}, y_i), i = 1,2,\cdots,n$, and its regression model can be expressed as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \tag{5}$$

Among them, β_0 is the intercept term, β_1 and β_2 are the regression coefficients of the independent variables x_1 and x_2 respectively, and ϵ_i is the random error term. It is usually assumed that $\epsilon_i \sim N(0, \sigma^2)$.Select the incidence rate of cyber - crimes as the dependent variable, the policy score as the independent variable x_1, and the pattern score as the independent variable x_2. According to the data survey, Policy usually has a greater impact on cyber - crimes. Therefore, we hope that Policy has a higher weight. So we set β_1=- 0.544, β_2= - 0.36. The formula is obtained as:

y = 0.044 − 0.544* Policy score −0.36 ∗ Pattern Score

The goal of the least - squares method is to find the estimated values β_0, β_1 and β_2 of the parameters $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$, such that the sum of squared residuals.

$$Q(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2 \tag{6}$$

The output of the regression model will display the regression coefficients of each variable (pattern score and policy score). According to these coefficients, we can determine which factors have a greater impact on the cyber - crime rate and which policies or patterns are particularly effective in reducing cyber – crimes [6].

The results of the above regression model are displayed in Figure 6:

ISSN: 2664-9640

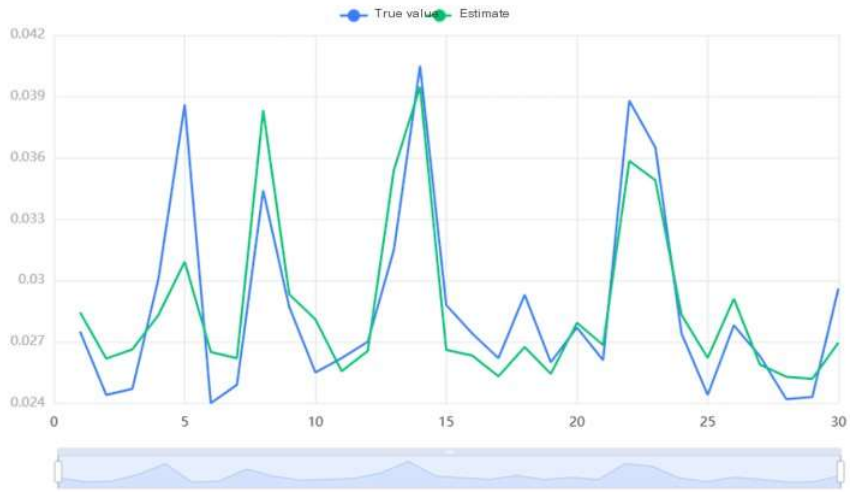

Figure 6. The results of the regression model show that

The coefficient of Pattern Score is -0.544, which suggests that higher pattern scores are associated with lower cybercrime rates.

The coefficient of Policy Score is -0.36, indicating that higher policy scores are also correlated with a reduction in cybercrime rates.

The absolute value of the coefficient for Pattern Score is larger than that of Policy Score, implying that Pattern Score has a stronger influence on the cybercrime rate than Policy Score.

These findings suggest that both Pattern Score and Policy Score have a significant impact on the cybercrime rate when analyzing countries/regions, and they can serve as crucial factors in the formulation of cybersecurity policies.

4. Conclusion

This study analyses the state of global cybercrime governance in multiple dimensions, and the results of the research are of great significance in the field of cybersecurity protection for intelligent transportation systems.

In terms of global cybercrime governance, there are obvious differences among countries in the dimensions of law, technology, organisation, capacity development and cooperation measures. The governance capacity of countries in some regions such as Africa is weak, while developed countries are relatively strong. Risk clusters are divided through K-means clustering analysis, which clearly presents the differences in governance levels of different countries and provides a basis for the targeted formulation of strategies. Meanwhile, Pearson correlation analysis reveals a strong positive relationship between governance measures, indicating that comprehensive promotion of various measures is crucial to effectively curbing cybercrime.

The results of the regression model show that both the policy score and the pattern score have a significant effect on the cybercrime rate, and the pattern score has a stronger effect. This implies that when formulating cybersecurity policies, multiple factors should be considered comprehensively and focus on aspects that contribute to the improvement of mode score.

Analogously to ITS, as the cybersecurity challenges it faces are closely related to cybercrime governance, experience can be gained from global research results on cybercrime governance. In the process of ITS construction, it is necessary to improve relevant laws and regulations, strengthen technical protection, optimise organisational management, strengthen international cooperation, and build an all-round network security protection system. Only in this way can we guarantee the data security and stable operation of ITS, promote the healthy

ISSN: 2664-9640

development of the ITS industry, and enable people to fully enjoy the convenience brought by ITS.

References

[1] Liu Tingting. Application and development thinking of intelligent transport system in urban transport management[J]. Auto weekly,2025,(03):65-67.

[2] Chen Haiyan. Study on the New Path of Responding to Cybercrime in the Context of Digital Era[J]. Legal Expo,2025,(03):25-27.

[3] CAI Wei, SUN Guangyu, YANG Fei, et al. Adaptive k-means clustering GDCW-AKM algorithm based on grid and domain centre-of-mass weights[J/OL]. Oil and Gas Storage and Transportation,1-11[2025-02-18].http://kns.cnki.net/kcms/detail/13.1093.TE.20250217.1007.002.html.

[4] Feng Xingjin. Exploration of usual grade evaluation method based on Pearson correlation coefficient[J]. Gansu Education Research,2025,(02):114-117.

[5] WANG Wei, DING Cong, LIU Shifan, et al. A step-type landslide warning surface model based on moving least squares[J]. Journal of Hohai University(Natural Science Edition),2025,53(01):95-102+120.

[6] ZHOU Xiao,LIN Zhi,LI Junxia. An empirical study on senior vocational students' willingness to participate in “dual-creation” based on fixed-order regression model[J]. Public Relations World, 2025,(03):171-173.