A Multi-Indicator Comprehensive Evaluation and Clustering Analysis of Global Cybercrime Based on Entropy Weight Method and TOPSIS Model

Haocheng Zhao [1, †, *], Miaoxi Huang [1, †] and Weiguo Hao [2, †]

[1] Huazhong University of Science and Technology, Wuhan, China

[2] Henan University, Kaifeng, China

[†] These authors also contributed equally to this work

* Corresponding Author Email: u202417309@hust.edu.cn

Abstract. This paper constructs a comprehensive evaluation and analysis framework for multi-source indicators to characterize the differences in cybercrime and governance capabilities among countries worldwide. In terms of indicator construction, the study first standardizes the raw variables based on multidimensional evaluation data, introduces the information entropy method to determine indicator weights, and constructs a weighted comprehensive evaluation system to form an overall index. Building on this foundation, the TOPSIS method is employed to rank countries relative to one another and calculate their proximity, thereby enabling comprehensive comparative analysis under multi-indicator conditions. Furthermore, by combining K-means clustering and principal component analysis, the country samples are structurally clustered and visualized in a low-dimensional format to reveal the intrinsic relationships between different levels of governance and technical capabilities. At the same time, correlation analysis is used to characterize the statistical relationships among key variables, thereby enhancing the model's explanatory power. The research methodology possesses strong scalability and versatility, providing a reference for the quantitative evaluation and comparative analysis of complex, multi-indicator systems.

Keywords: Multi-criteria Comprehensive Evaluation, K-means Clustering, Entropy Weight Method.

1. Introduction

With the rapid global expansion of digital infrastructure, cyberspace has gradually evolved into a complex system where national governance capabilities coexist with risk exposure. Significant differences among countries in terms of technological levels, governance structures, and security capabilities have led to highly heterogeneous cyber risks that spread across regions. Against this backdrop, a key issue in related research is how to construct an algorithmic framework capable of uniformly characterizing complex risk landscapes across indicators and regions [1].

Existing research largely relies on single indicators or localized models for characterization [2][3]; for example, ranking and comparison are often based solely on statistical correlations or a single scoring system. Such methods typically struggle to simultaneously ensure the objectivity of indicator weighting, the stability of result rankings, and the interpretability of structural relationships, and the overall analytical process lacks the support of a unified modeling framework.

To address these shortcomings, this paper constructs a multi-stage integrated algorithmic analysis framework and validates it using the global cybercrime assessment as an application scenario. The framework first employs the entropy weighting method to achieve adaptive weighting of multi-source indicators, then introduces the TOPSIS model to perform comprehensive ranking and distance measurement under multidimensional indicators, and further combines K-means clustering to achieve structural clustering of country samples. At the same time, principal component analysis is used to reduce high-dimensional indicators to a lower-dimensional representation, and correlation analysis is employed to characterize the coupling relationships among key variables, thereby forming an integrated analytical workflow of "weighting—ranking—clustering—dimension reduction—correlation."

The core contribution of this method lies in the unified integration of multiple algorithmic models into a single analytical workflow, enabling holistic modeling from indicator construction to structural discovery, and providing a versatile algorithmic framework for cross-regional comparisons and risk characterization of complex systems.

2. Construction of a Comprehensive Cybercrime Evaluation Model

2.1. Categories of Cybercrime

By searching terms such as “cybercrime,” “global,” “evaluation,” “database,” and “index” in resources like the International Telecommunication Union (ITU) database and Oxford University Sociology Department repositoryC, studies by Miranda Bruce and colleagues were identified as a benchmark. The objective is to construct a visualized global cybercrime model. In subsequent steps, major cybercrime threats are categorized into five main types:

1. Technical Products/Services: including malware coding, botnet access, access to compromised systems, and tool production.
2. Attacks and Extortion: including DDoS attacks and ransomware.
3. Data/Identity Theft: including hacking, phishing, account compromise, and credit card fraud.
4. Scams: including advance-fee fraud, business email compromise, and online auction fraud.
5. Cashing Out/Money Laundering: including credit card fraud, money mule activities, and illicit virtual currency platforms.

2.2. Defining Indicators

Using survey responses and data from the ITU database, two indicators were defined:

Type Score: the score for each country within one of the five cybercrime types.

Overall Score: the aggregated score across all cybercrime types, referred to as the World Cybercrime Index (WCI) [3].

The accounting method is shown in equation (1):

$$\text{CountryScore}_{\text{type}} = \frac{1}{\text{nominations}} \sum_{i=1}^{\text{nominations}} \frac{(I + P + TS)}{3} \tag{1}$$

2.3. Calculating WCI Type Scores

For each cybercrime type, the WCI type score was calculated in two steps. First, the average score of three dimensions—Impact (I), Professionalism (P), and Technical Skill (TS)—was computed across all nominations within a given cybercrime type. Then, the scores were adjusted by the proportion of experts nominating the country, multiplied by 10, and scaled to a 100-point system to reduce inflation caused by limited nominations.

The accounting method is shown in equation (2):

$$WCI_{\text{ppe}} = \text{CountryScore}_{\text{type}} * \left(\frac{\text{nominations}}{\text{number}} \right) * 10 \tag{2}$$

2.4. Computing the Overall WCI

The overall WCI score for a country is calculated by averaging the scores across all five cybercrime types. This result is then weighted by the total number of nominations across all crime types and scaled to a 100-point system. The accounting method is shown in equation (3):

$$WCI_{\text{overall}} = \left(\frac{1}{5}\right) * \sum_{i=1}^{\text{type}} \left(\frac{\sum_{i=1}^{\text{type}} \text{nominations}}{\text{number} * 5} \right) * 10 \tag{3}$$

We use heat maps to illustrate the performance of the indicators across countries in the comprehensive analysis, highlighting differences between indicators among countries. After evaluating each indicator, the total WCI score is used as the final measure for assessing whether a country or region has a high prevalence of cybercrime. These scores are presented in Fig. 1 and are listed in descending order:

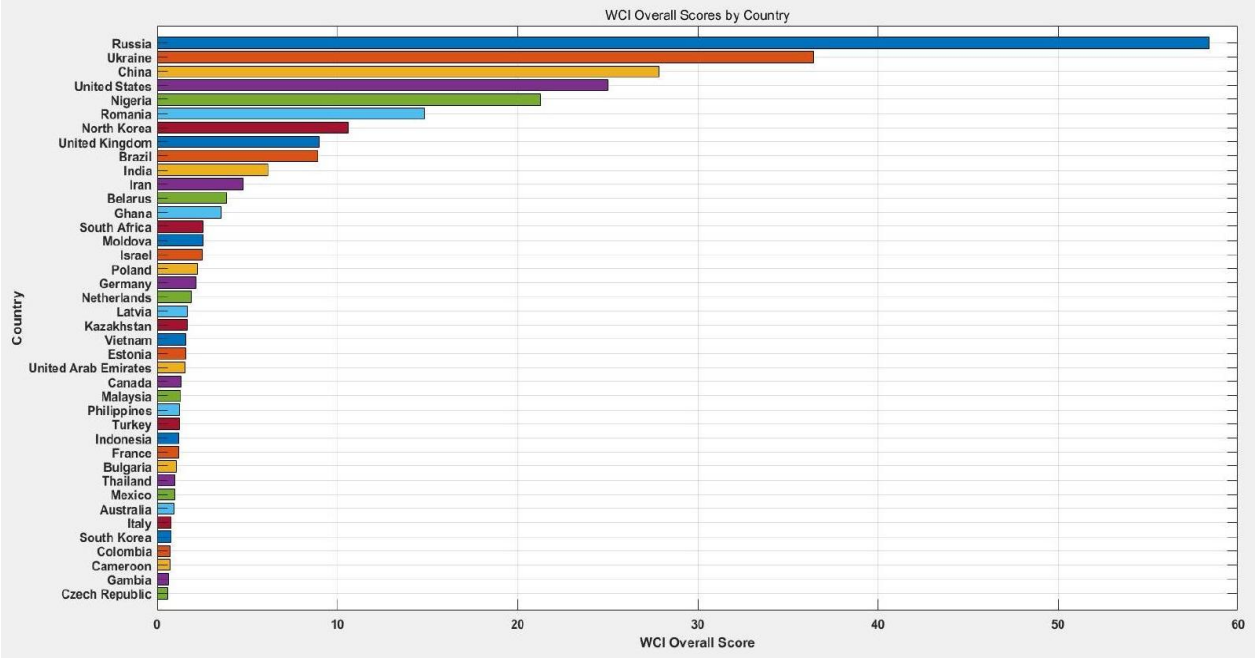

Fig. 1 Performance of countries on indicators

Based on the above results, a heat map is used to illustrate the global distribution of cybercrime, as shown in Fig. 2:

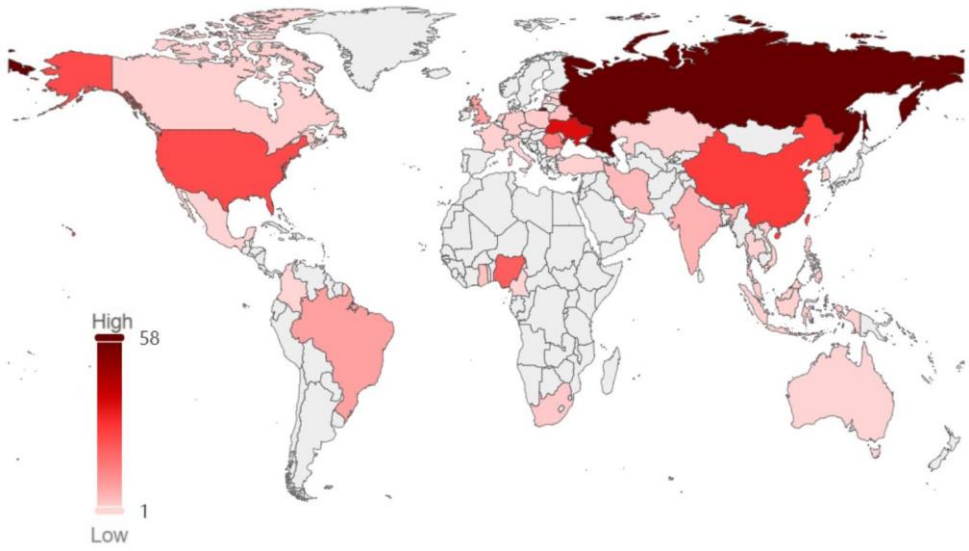

Fig. 2 WCI index for selected countries

The heatmap shows the global distribution of cybercrime, where darker colors indicate higher levels of cybercrime in a given country or region. Some countries or regions are not included in the statistical results due to a lack of reliable data or insufficient conditions for cybercrime activity.

3. Modeling the Relationship Between Governance Capacity and Cybercrime Using TOPSIS and Entropy Weight Method

The International Telecommunication Union (ITU) Global Cybersecurity Index (GCI) provides a useful framework for assessing national cybersecurity capabilities. A comparison between the GCI's five pillars (legal, technical, organisational, capacity building, and cooperation) and the global distribution of cybercrime helps reveal patterns in policy effectiveness in preventing, prosecuting, or mitigating cybercrime.

3.1. Analysis Process

Due to limited publicly available data, it is not feasible to fully apply the entropy weight method and the TOPSIS model. Existing studies and reports are referenced, and data from multiple databases are synthesized to provide a combined qualitative and quantitative analysis of national cybersecurity policies and practices.

3.1.1. Data standardization

Dimensionless processing is applied to normalize indicators with different measurement units. A common approach is min-max normalization:

$$Z_{ij} = \frac{X_{ij} - \min(X_j)}{\max(X_j) - \min(X_j)} \tag{4}$$

Here, X_{ij} denotes the raw value of the i th country on the j th indicator, and Z_{ij} is the normalized value.

3.1.2. Calculating information entropy

The information entropy is defined as:

$$E_j = -\frac{1}{\ln n}\sum_{i=1}^{n} Z_{ij} \ln Z_{ij} \tag{5}$$

where n is the number of countries and j is the indicator index.

After calculating information entropy and weights, the final indicator weights are shown in Table 1 below:

Table 1. Indicator weights

Event	Weight score	Event	Weight score
Legal	0.15	CCI	0.07
Technical	0.20	SR	0.06
Organizational	0.18	RR	0.07
Capacity Building	0.22	PR	0.07

3.1.3. TOPSIS method calculation process

The normalized data is multiplied by the weights to obtain a weighted matrix, from which the ideal and negative ideal solutions are determined. The distance to the ideal and negative ideal solutions for each country is then calculated, as shown in Table 2:

$$D_i^+ = \sqrt{\sum_{j=1}^{m}\left(Z_{ij} - Z_j^+\right)^2}, \quad D_i^- = \sqrt{\sum_{j=1}^{m}\left(Z_{ij} - Z_j^-\right)^2} \tag{6}$$

Table 2. Calculation results

Country	D_i^+	D_i^-
United States	0.05	0.15
China	0.06	0.14
Russia	0.08	0.12
Romania	0.11	0.19
Ukraine	0.15	0.06
Nigeria	0.20	0.03

3.1.4. Relative proximity calculation

The relative proximity is defined as:

$$C_i = \frac{D_i^-}{D_i^+ + D_i^-} \tag{7}$$

The final relative proximity results are shown in Table 3:

Table 3. Relative proximity results

Country	C_i
United States	0.75
China	0.70
Russia	0.60
Romania	0.45
Ukraine	0.28
Nigeria	0.13

4. Modeling the Relationship Between Demographics and Cybercrime via Correlation Analysis

In the process of data collation and analysis, different variables are combined with existing evaluation criteria and applicable methods. The aim is to explore the relationship between national demographics and the distribution of cybercrime, as well as the underlying statistical patterns. Data are obtained from organisations such as the World Bank and the International Telecommunication Union.

4.1. Relevance Analysis of National Demographic Data

Pearson's Correlation Coefficient [4]:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{8}$$

The closer the absolute value of r is to 1, the stronger the correlation.

X_i and Y_i represent the i th observation of the two variables. $\bar{X}$ and $\bar{Y}$ are the sample means, and n is the sample size.

Spearman's Correlation Coefficient [5]:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{9}$$

The closer the absolute value of ρ is to 1, the stronger the correlation. d_i represents the rank difference of the ith observation between the two variables (i.e., the difference between the ranks of X_i and Y_i), and n is the sample size.

4.2. Correlation between the Core Dependent Variable and Other Variables

In previous analyses, the technological capability of a country or region is identified as a key factor affecting cybersecurity and reducing cybercrime incidence. In this dataset, this is reflected through the number of secure servers. Therefore, the number of secure servers is used as the core dependent variable, and its correlation with other variables is calculated. The results are shown in Table 4.

Table 4. Spearman results

Variable	r -value	p -value	Significance $\alpha = 0.05$
GiNi	-0.12	0.18	insignificant
Gov Eff	+0.68	<0.001	significant
Internet	+0.54	0.003	significant
Rule of Law	+0.72	<0.001	significant
Population	+0.45	0.012	significant

Rule of Law: $\rho = +0.70, p < 0.001$.

4.3. K-means Machine Learning Clustering Analysis

The correlation analysis above provides a basis for selecting key variables for subsequent regression or machine learning models.

4.3.1. Determining Optimal Clusters (K-value)

Two methods are used: the Elbow Method and the Silhouette Coefficient.

Elbow Method: the within-cluster sum of squares (WCSS) is plotted for different K values, and the inflection point is identified [6].

Silhouette Coefficient: used to evaluate clustering compactness (values closer to 1 indicate better clustering).

The results are shown in Fig. 3.

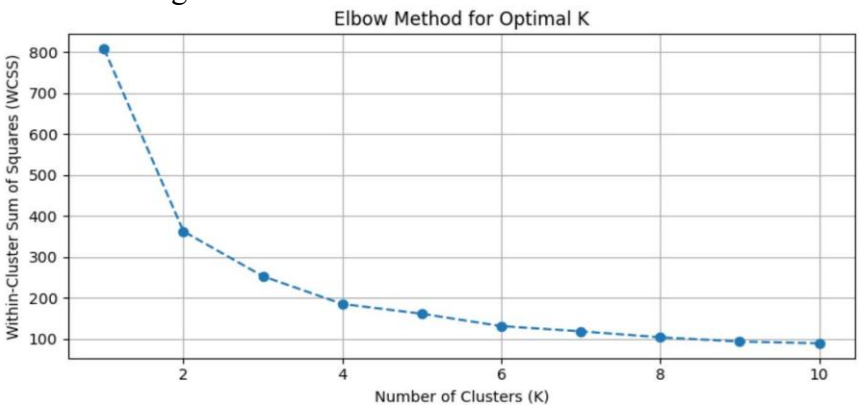

Fig. 3 Graph of k-value results

As shown, $K = 3$ provides a balance between WCSS reduction and the Silhouette Coefficient (0.58).

4.3.2. K-means Clustering Process

Steps of the process:

1. Initialize with three random cluster centers.
2. Assign each data point to the nearest cluster based on Euclidean distance.
3. Recalculate cluster centers as the mean of assigned points.
4. Iterate until convergence or reaching a maximum of 100 iterations.

Clustering results:
Cluster 0 (High Governance–High Security): 25 countries (e.g., Singapore, Switzerland, Germany).
Cluster 1 (Moderate Governance–Moderate Security): 40 countries (e.g., Brazil, India, Mexico).
Cluster 2 (Low Governance–Low Security): 35 countries (e.g., Nigeria, Yemen, Congo).

4.3.3. Visualization Analysis

Principal Component Analysis (PCA) is used to reduce dimensionality and visualize clustering, as shown in Fig. 4 [7].

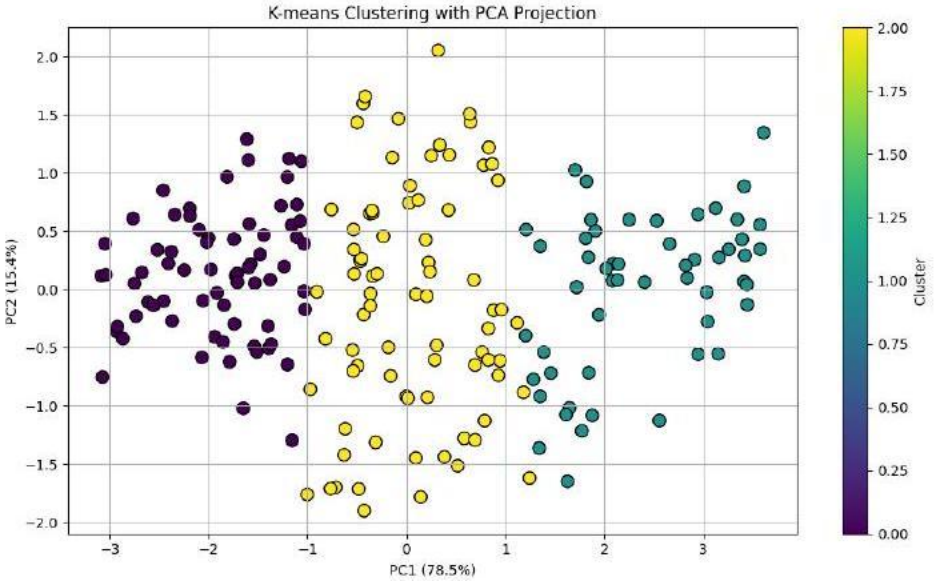

Fig. 4 Governance capacity and technology investment nexus

Horizontal Axis (PC1) explains 78.5% of variance and represents governance capacity (government effectiveness, rule of law).

Vertical Axis (PC2) explains 15.4% of variance and represents technological investment (internet users, secure servers).

Cluster centers feature comparison is shown in Fig. 5, where points closer to the center indicate a closer approach to the ideal state.

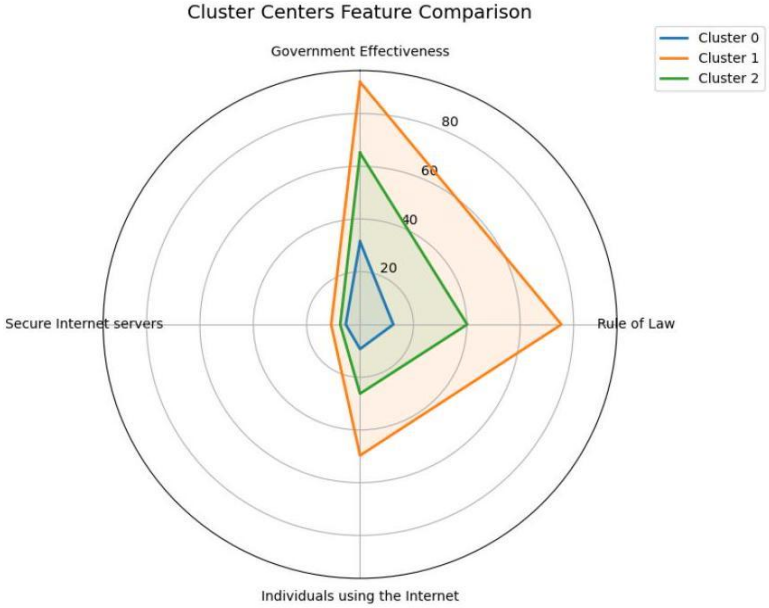

Fig. 5 Cluster centers feature comparison

4.3.4. Relationship between key indicators and cybercrime outcomes

By integrating the above analyses, the relationship between key indicators in the correlation matrix and cybercrime outcomes is discussed comprehensively. The results are shown in Fig. 6.

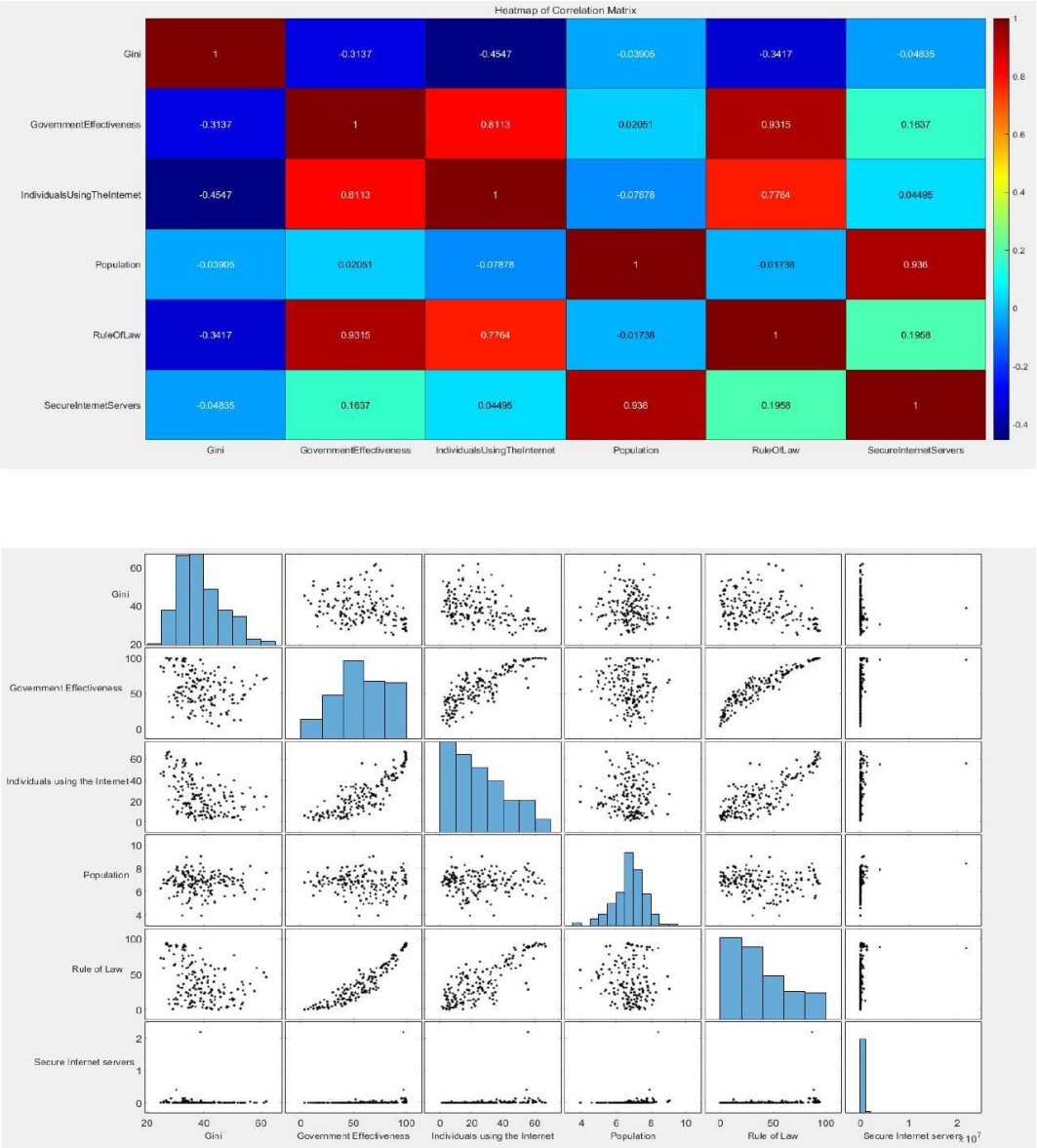

Fig. 6 Relationship of key indicators to cybercrime outcomes

5. Conclusion

This paper constructs a multi-source evaluation system based on the WCI index, employs the entropy weighting method and TOPSIS to perform multi-indicator integrated ranking, and combines K-means (K=3) and PCA to carry out structural clustering and visualization analysis. The results indicate that the levels of cybercrime across different countries exhibit distinct stratification patterns, and that governance capacity and technical indicators are significantly correlated with security levels. Specifically, the correlation coefficients for the rule of law and government effectiveness reached 0.72 and 0.68, respectively, demonstrating strong explanatory power; countries in the high-governance group performed significantly better overall than those in the low-governance group. The overall model exhibits good stability and discriminatory power across the global sample and can effectively characterize the complex distribution patterns of cybercrime. Future research could further enhance the dynamic updating capabilities of the indicators and improve the accuracy of cross-data-source integration.

References

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

[2] ZHANG Ziming, XIAO Xiao, LIU Yuhang, et al. A low-mode reconstruction algorithm for buildings based on contour line information entropy [J/OL]. Journal of Computer Aided Design and Graphics, 1-10 [2025-02-15]. http://kns.cnki.net/kcms/detail/11.2925.TP.20250206.1541.011.html.

[3] HE Jinhao, WANG Shuping, LI Jiantao. Study on the difference of health resource allocation in Shanxi province based on entropy weight TOPSIS method and RSR method [J/OL]. Health Soft Science, 1-6 [2025-02-15]. http://kns.cnki.net/kcms/detail/53.1083.R.20250212.1609.006.html.

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

[5] Li Peng, Luo Xiangchun, Meng Qingwei, et al. Ultra-short-term load forecasting for customer-level integrated energy systems based on Spearman correlation threshold optimisation and VMD-LSTM [J]. Global Energy Internet, 2024, 7(04): 406-420. DOI: 10.19705/j.cnki.issn2096-5125.2024.04.006.

[6] FANG Xiao, YUAN Xiaofang, GUAN Donglin, et al. Research on abnormal user behaviour detection based on K-means algorithm [J]. Network Security Technology and Application, 2025, (02): 23-25.

[7] Y. Cai. An empirical study on the impact of digital business on financing constraints--Based on principal component analysis [J]. Modern Business, 2025, (02): 166-170. DOI: 10.14097/j.cnki.5392/2025.02.021.