Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 14, No. 3, 2025

Cybersecurity Policy Research Driven by Data

Xiaodie Hu, Yuzhu Du, Yingying Feng, Zhifu Jia *

School of Mathematics and Physics, Suqian University, Suqian, Jiangsu, China
*** Corresponding author:** Zhifu Jia (Email: jzflzbx@163.com)

Abstract: To support data-driven national cybersecurity policy formulation, we developed a multidimensional analytical framework focusing on global cybercrime distribution, policy effectiveness, and the correlation between demographic characteristics and criminal patterns. First, we collected authoritative data from multiple sources including the International Telecommunication Union (ITU) and conducted preprocessing. Using a K-means clustering model based on Euclidean distance with Python tools, we revealed the global distribution patterns of cybercrime. Second, by integrating time series analysis with Difference-in-Differences (DID) models, we visualized changes in policy-related indicators before and after implementation, quantified policy impacts, and identified key effective policies. Finally, we constructed a multiple linear regression model incorporating interaction effects to visualize correlations between demographic factors and cybercrime, utilizing statsmodels tools for precise predictions. This study demonstrates how the integration of multi-methods and efficient computation provides scientific decision-making support for optimizing cybersecurity policies and global cybersecurity governance.

Keywords: Cyber Security Policy; Cluster Analysis; Time Series Analysis; DID Model; Multiple Linear Regression.

1. Introduction

With the rapid development of digital technology, the formulation of cybersecurity policies has become crucial. As the cornerstone of the digital economy, cybersecurity policies are closely related to the protection of individual rights and interests, the compliance operation of enterprises, national sovereignty security, and global collaborative governance. Against the backdrop of the internationalization and complexity of cybercrime, different countries, due to differences in network infrastructure, protection capabilities, and demographic characteristics, show significant variations in the distribution of cybercrime and the effectiveness of policies. High-precision policy analysis modeling plays a key role in enhancing governance efficiency and reducing security costs. Analyzing the patterns of cybercrime and evaluating policy effectiveness can help optimize protection strategies and strengthen collaboration mechanisms, providing data support and decision-making basis for responding to digital security threats. The uncertainty of cybercrime not only disrupts the order of the digital space, but also significantly affects the efficiency of security investment, the level of public trust, and the effectiveness of cross-border collaborative governance in various countries.

Accurate cybersecurity policy analysis is crucial for addressing digital threats, ensuring the healthy development of the digital economy, and maintaining global cyber stability. Researching data-driven policy-making methods is key to addressing cybercrime challenges, strengthening security governance capabilities, optimizing resource allocation, and maintaining social stability. Efficient analysis models can not only identify policy shortcomings and strengths but also predict the evolution trends of cybercrime, effectively alleviating the multi-dimensional negative impacts of cybercrime. Currently, the global cyberspace presents a diverse development pattern. Differences among countries in terms of internet penetration rates, wealth levels, and education levels lead to different manifestations and response needs of cybercrime. Some countries, due to weak protection technologies and incomplete law enforcement systems, have persistently high rates of cybercrime. Moreover, differences in policy statistics standards and collaboration mechanisms among countries further increase the difficulty of governance. Grasping the global distribution patterns of cybercrime and the mechanisms of policy effects can enhance the targeted and effective nature of policy formulation in various countries, thereby improving the overall global cybersecurity protection level.

This paper comprehensively employs models such as K-means clustering, time series analysis, difference-in-differences (DID), and multiple linear regression to model data-driven cybersecurity policies. Based on multi-channel authoritative data, a research dataset is constructed through data cleaning, standardization, and outlier removal. By systematically optimizing model parameters and integrating multi-dimensional analysis perspectives, a dynamic balance between the depth of data mining and the reliability of conclusions is achieved, aiming to reveal the global distribution patterns of cybercrime, accurately assess the effectiveness of policy implementation, and clarify the influence mechanisms of demographic characteristics, providing scientific basis for countries to optimize cybersecurity policies and strengthen global collaborative governance. Organization of the Text

2. Research Status

2.1. Research on K-Means Clustering Model

The K-Means clustering algorithm [1] was proposed by MacQueen in 1967 and is a classic algorithm in the field of unsupervised learning. Its core logic is to divide the data set into a preset number of clusters through an iterative optimization process, ultimately achieving the goal of maximizing the similarity within clusters and the difference between clusters. Due to its simple and understandable principle, high computational efficiency, and strong scalability, this algorithm has been widely applied in various fields such as data classification and pattern recognition.

In the research of cybersecurity, the K-Means clustering algorithm is often used to explore the differences in

cybercrime characteristics among different regions or groups. For instance, by conducting cluster analysis on multi-dimensional data such as the incidence rate of cybercrime, the level of cybersecurity protection technology, and the allocation of cybersecurity personnel, researchers can effectively identify clusters of countries or regions with similar crime patterns, providing a basis for subsequent targeted research. However, the traditional K-Means clustering algorithm has obvious limitations: it is sensitive to the selection of initial cluster centers and is prone to getting stuck in local optima; when dealing with high-dimensional data, the discrimination of traditional distance measurement methods decreases, thereby affecting the accuracy of the clustering results. In recent years, to enhance the applicability of this algorithm in cybersecurity data analysis, researchers have continuously explored optimization paths, such as using intelligent optimization algorithms to initialize cluster centers and combining dimensionality reduction techniques to handle high-dimensional data.

2.2. Research on Time Series and DID Model

Time series analysis is a statistical analysis method[2] based on the temporal characteristics of data. By modeling data at consecutive time points, it can effectively uncover the trends, periodicity, and correlations in the data, thereby capturing the dynamic evolution of variables over time. In the evaluation of cybersecurity policies, time series analysis can visually present the changes in key indicators before and after policy implementation, providing a direct basis for preliminary judgment of policy impact.

The Difference-in-Differences (DID) model[3] is one of the core methods in policy evaluation. Its core advantage lies in setting up a policy implementation group and a control group, which can effectively eliminate the interference of time trends and individual fixed effects, accurately quantifying the net effect of policy implementation. This model has unique value in the evaluation of cybersecurity policies, as it can effectively separate the impact of policy factors from other external factors on cybercrime, making the evaluation results more reliable. Currently, the DID model has been widely applied in analyzing the implementation effects of specific cybersecurity regulations and technical standards. However, in cross-national policy comparison studies, it is necessary to fully consider the differences in systems and inconsistent data statistics standards among different countries, otherwise it may lead to deviations in the evaluation results.

2.3. Research on Multiple Linear Regression Model

The multiple linear regression model [4] is a classic statistical method for quantifying the linear relationship between multiple factors and the dependent variable. By estimating regression coefficients, it can clearly identify the direction and degree of influence of each independent variable on the dependent variable, and supports the introduction of interaction terms to analyze the synergy between variables. It has wide applications in various fields such as social sciences and natural sciences.

In the field of cybersecurity, the multiple linear regression model is often used to explore the relationship between demographic characteristics, socio-economic factors, and cybercrime. In existing research, scholars have verified through constructing multiple linear regression models that factors such as internet penetration rate, per capita GDP, and education level are correlated with the incidence rate, reporting rate, and prosecution rate of cybercrime. However, the influencing factors of cybercrime have significant complexity and nonlinearity, and simple linear models are difficult to fully capture their internal correlations. Therefore, researchers have begun to optimize the model, such as combining interaction effects and hierarchical regression methods, and improving the model fitting effect through data standardization and outlier handling, providing strong support for accurately identifying risk and protective factors of cybercrime.

2.4. The Content of This Paper

With the development of cybercrime showing a trend of internationalization and complexity, the formulation and optimization of cybersecurity policies have become a core issue of global concern. Existing related research mainly focuses on policy content analysis, implementation effect evaluation, and international comparison. In terms of policy content research, scholars mainly focus on core dimensions such as crime prevention, law enforcement accountability, technical protection, and international cooperation. In terms of effect evaluation, a combination of qualitative and quantitative analysis is often adopted, and the actual effect of policies is verified through case studies and data statistics.

At present, certain achievements have been made in the research of cybersecurity policies, but there are still obvious deficiencies. On the one hand, some studies lack multi-dimensional and large-sample data support, resulting in incomplete and subjective policy effect evaluations. On the other hand, the moderating role of demographic characteristics on policy implementation effects is rarely studied, making it difficult to form targeted policy optimization suggestions. In addition, due to differences in the development levels of cybersecurity and data statistics standards among countries, cross-national policy comparison studies face problems such as data incomparability and difficulty in generalizing conclusions. Therefore, building a data-driven multi-model integrated analysis framework that integrates methods such as cluster analysis, time series analysis, and regression analysis has become a key direction for enhancing the scientificity and practicality of cybersecurity policy research.

3. Modelling Preparation

3.1. Description of Symbols

To clarify the meaning and application of each variable in the model and ensure the accuracy and consistency of subsequent model construction, solution, and result analysis, the core symbols involved in this study are uniformly explained as follows:

3.2. Data Sources

1.Following the principles of authority, comprehensiveness, and timeliness, this study collects relevant data from multiple countries through various channels to provide a solid data foundation for model construction. The specific data sources are as follows:

2.Public data from international organizations: Mainly includes global internet development reports and cybersecurity-related statistical data released by the International Telecommunication Union (ITU), covering core

indicators such as the internet penetration rate and the level of network infrastructure construction in various countries; demographic data and global development index reports released by relevant United Nations agencies, providing socio-economic characteristic data such as per capita GDP and education level of various countries.

Table 1. List of Symbols

Symbols	Explanation
T_{ij}	Represents the proportion of the jth cybercrime type in country i out of the total cybercrime cases.
A_i	Indicates the network security protection technology index of country i, and the value range is 0-100
I_i	Represents the Internet penetration rate of country i, i.e. the proportion of the country's population using the Internet.
K	The number of clusters represented in cluster analysis (e.g., K-Means algorithm).
μ_i	Represents the centroid of the ith cluster in cluster analysis.
Y_{it}	The dependent variable represents the number of cybercrime incidents in country ith at time t.
β_0	Intercept terms in a regression model.
$\grave{o}_{it}$	Random error term.

3.Official data released by various countries: Collect annual cybersecurity reports and crime statistical bulletins released by government departments of multiple countries to obtain first-hand data such as the number of cybercrime incidents, crime types, reporting rate, and prosecution rate in various countries; at the same time, sort out the text of cybersecurity policies issued by various countries to clarify the policy release time and core content.

4.Academic and industry databases: Extract verified cybercrime-related data from research reports released by authoritative academic journals and industry research institutions in the field of cybersecurity, including the success rate of cybercrime and the distribution characteristics of different types of cybercrime; refer to standardized data in published studies to ensure data comparability.

5.Data collection scope: Covers countries from different regions and with different development levels around the world, including developed countries. A total of 50 representative countries is selected as research samples, with a data time span of 2018-2022, to ensure that the distribution characteristics of recent cybercrime and the effect of policy implementation can be reflected.

3.3. Dataset Division

To realize the training, validation, and effect evaluation of the model, the collected dataset is divided into a training set and a test set according to functions. The specific division rules are as follows:

Table 2. Partition of the Dataset

Dataset	Data Range	Sample Size	Purpose
Training set	Cybercrime-related data, demographic data, and policy data of 50 countries from 2018 to 2020.	150 groups (50countries×3years)	Used for parameter training of the K-Means clustering model, trend fitting of the time series analysis model, and parameter estimation of the DID model and multiple linear regression model to ensure that the model can learn the core laws in the data.
Test set	Cybercrime-related data, demographic data, and policy data of 50 countries from 2021 to 2022.	100 groups (50 countries×2 years)	Used to verify the prediction effect and generalization ability of each model. By comparing the model prediction results with the actual data of the test set, the accuracy and reliability of the model are evaluated to provide a basis for model optimization.

Note: Each sample in the dataset contains complete indicator information of a single country in a certain year, including cybercrime characteristic indicators (number of incidents, success rate, reporting rate, prosecution rate, etc.), socio-economic characteristic indicators (internet penetration rate, per capita GDP, education level, etc.), and policy characteristic indicators (whether policies are issued, policy implementation duration, etc.), ensuring the completeness and effectiveness of the samples.

4. Data Preprocessing

To ensure the accuracy, completeness, and comparability of the input data for the model and lay a solid foundation for subsequent model construction and solution, this study conducted systematic data preprocessing on cybercrime-related data, demographic data, and policy data collected from multiple channels. The specific process includes three parts: data cleaning and integration, data standardization, and data stationarity test.

4.1. Data Cleaning and Integration

4.1.1. Verification of Data Completeness and Accuracy

Each entry of the collected data was verified item by item, focusing on confirming that core indicators (such as the number of cybercrime incidents, policy release time, internet penetration rate, and per capita GDP) are free of missing values and duplicate records. By cross-validating data of the same indicator from different data sources (such as international organization reports and official bulletins of various countries), contradictory values caused by differences

in statistical standards or data entry errors were corrected. For example, regarding the deviation in the cybercrime prosecution rate of a country in different reports, the data from the annual bulletin released by the country's judicial department was used as the standard for correction.

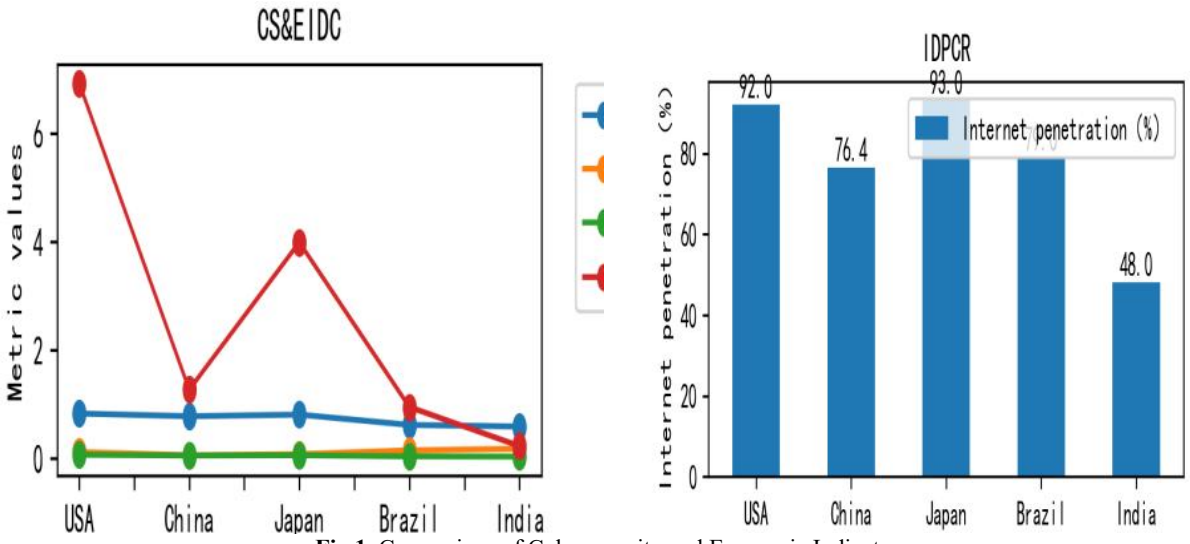

Fig 1. Comparison of Cybersecurity and Economic Indicators

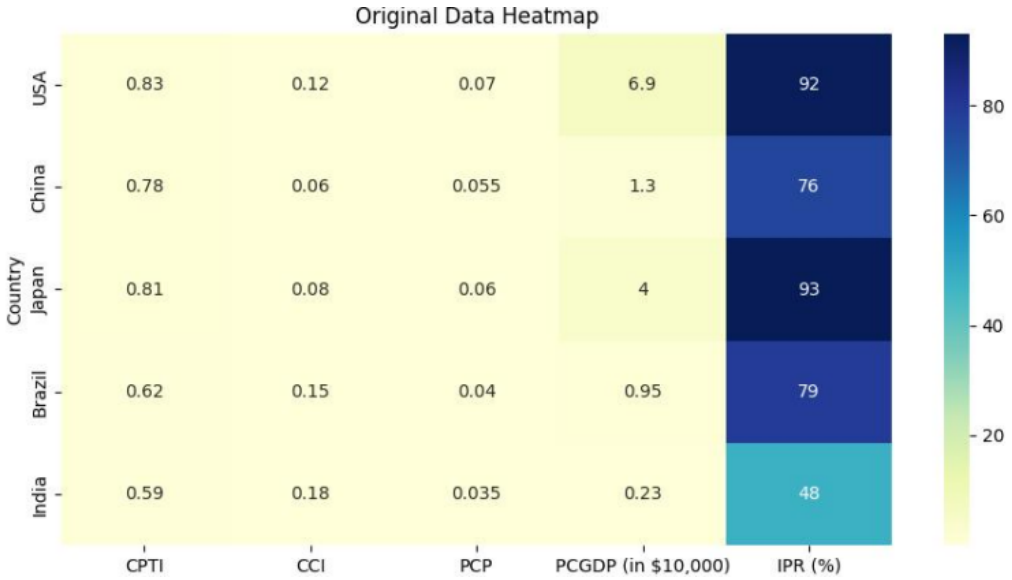

Fig 2. Data Heat Map

4.1.2. Missing Value Handling

For missing values in the dataset, differentiated processing strategies were adopted according to the missing scale and data characteristics:

Minor missing (single indicator missing ratio < 5%): For continuous indicators (such as per capita GDP and education level), the mean value of countries with the same regional and development level or the median value of the country in adjacent years was used for imputation; for time series data (such as the annual number of cybercrimes in a country), linear interpolation[5] was used for imputation to ensure the continuity of data trends.

Major missing (single indicator missing ratio ≥ 5%): If the missing data is concentrated in individual countries or non-core indicators and has limited impact on the overall analysis, the sample was retained and noted in subsequent analyses; if the missing data is widely distributed or involves core indicators (such as policy implementation status), the sample was excluded to avoid model bias caused by data missing.

4.1.3. Outlier Handling

Outliers were identified by combining statistical tests [6] with domain knowledge:

Based on the 3σ principle: For continuous data (such as the number of cybercrime incidents and internet penetration rate), the mean (μ) and standard deviation (σ) were calculated, and data exceeding the range [μ-3σ, μ+3σ] was identified as potential outliers.

Domain knowledge verification: Potential outliers were verified in combination with common sense in the field of cybersecurity. For example, if the cybercrime incidence rate of a country is significantly more than 10 times higher than the global average, it was verified to be a statistical standard error (including minor cyber violations in statistics) and corrected; if the cause of the outlier cannot be clarified and the outlier has a significant impact on the model, it was excluded.

4.1.4. Data Integration

The cleaned multi-source data was integrated into a unified data framework to construct a three-dimensional panel dataset

containing "country-year-core indicators". The dataset covers three categories of indicators:

Cybercrime characteristic indicators: Number of cybercrime incidents, success rate, reporting rate, prosecution rate;

Socio-economic characteristic indicators: Internet penetration rate, per capita GDP, education level, cybersecurity protection technology index;

Policy characteristic indicators: Policy release year, policy type (prevention type/law enforcement type/comprehensive type), policy implementation duration.

Through data integration, the correlation and matching of data from different dimensions were realized, providing unified data support for multi-model analysis.

4.2. Data Standardization

Due to the large differences in dimensions and value ranges of different indicators (such as per capita GDP in ten thousand US dollars and internet penetration rate in %), direct input into the model will lead to weight imbalance and affect the accuracy of analysis results. Therefore, all continuous indicators were standardized using the Z-score standardization [7] method to convert the data into standardized data with a mean of 0 and a standard deviation of 1. The formula is as follows:

$$Z_i = \frac{X_i - \bar{X}}{S} \tag{1}$$

Where Z_i is the standardized data, X_i is the original data, $\bar{X}$ is the mean value of the indicator, and S is the standard deviation of the indicator.

After standardization, all indicators are at the same order of magnitude, eliminating the impact of dimension differences, ensuring that models such as K-means clustering and regression analysis can treat each indicator fairly, and improving the stability and result reliability of the model.

4.3. Data Stationarity Test (ADF Test)

For time series data (such as the annual number of cybercrime incidents in various countries and indicator changes after policy implementation) As shown in Figure 2, a stationarity test is required to avoid spurious regression caused by non-stationary data. This study used the Augmented Dickey-Fuller (ADF) [8] test to judge data stationarity. The specific steps are as follows:

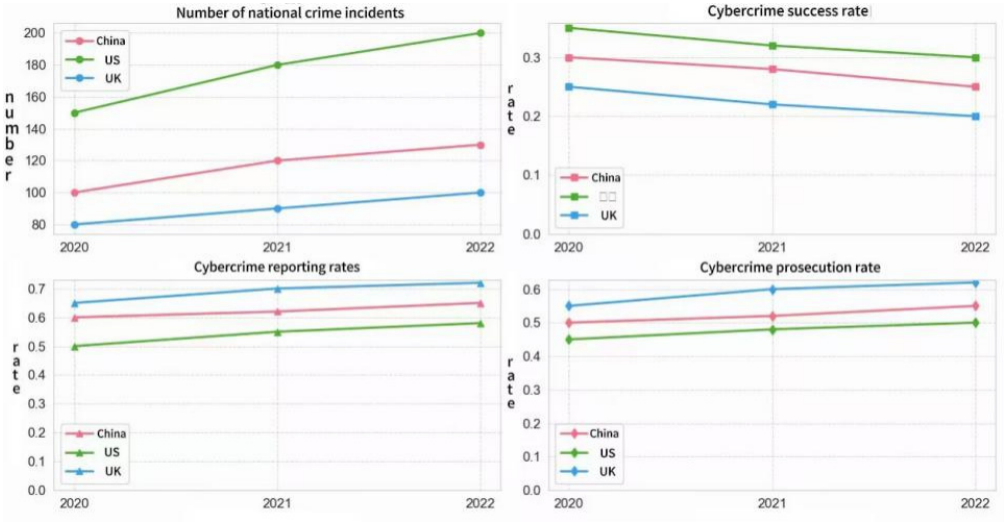

Fig 3. The Relationship between Indicators and Policy Releases

1. Hypothesis proposal: The null hypothesis (H0) is that the data sequence has a unit root, i.e., the sequence is non-stationary; the alternative hypothesis (H1) is that the data sequence has no unit root, i.e., the sequence is stationary.

2. Calculation of test statistics: The ADF test statistic was calculated based on the time series data, and compared with the critical values at different significance levels (1%, 5%, 10%).

3. Result judgment: If the ADF test statistic is less than the critical value and the P-value < 0.05, the null hypothesis is rejected, indicating that the sequence is stationary; if the above conditions are not met, the data is subjected to difference processing [9] (first-order difference, second-order difference) until the sequence becomes stationary.

Taking the time series of the number of cybercrime incidents as an example, the ADF test results show that the P-value of the original sequence is 0.621 (> 0.05), and the null hypothesis cannot be rejected, indicating that the sequence is non-stationary; after first-order difference processing, the P-value is 0.000 (< 0.01), and the ADF test statistic is less than the critical value at the 1% significance level, rejecting the null hypothesis, indicating that the differenced sequence is a stationary sequence, which can be used for time series analysis and DID model construction.

Table 3. Table ADF Test Results of Time Series of the Number of Cybercrime Incidents

Sequence Type	Order of Difference	ADF Statistic	P-value	1% Critical Value	5% Critical Value	10% Critical Value	Stationarity Judgmen
Original Sequence	0	-1.423	0.621	-3.582	-2.928	-2.602	Non-stationary
First-order Difference Sequence	1	-7.156	0.000	-3.585	-2.929	-2.603	Stationary

Note: indicates significance at the 1% level.

5. Modelling Ideas

The modelling idea of this paper is as follows

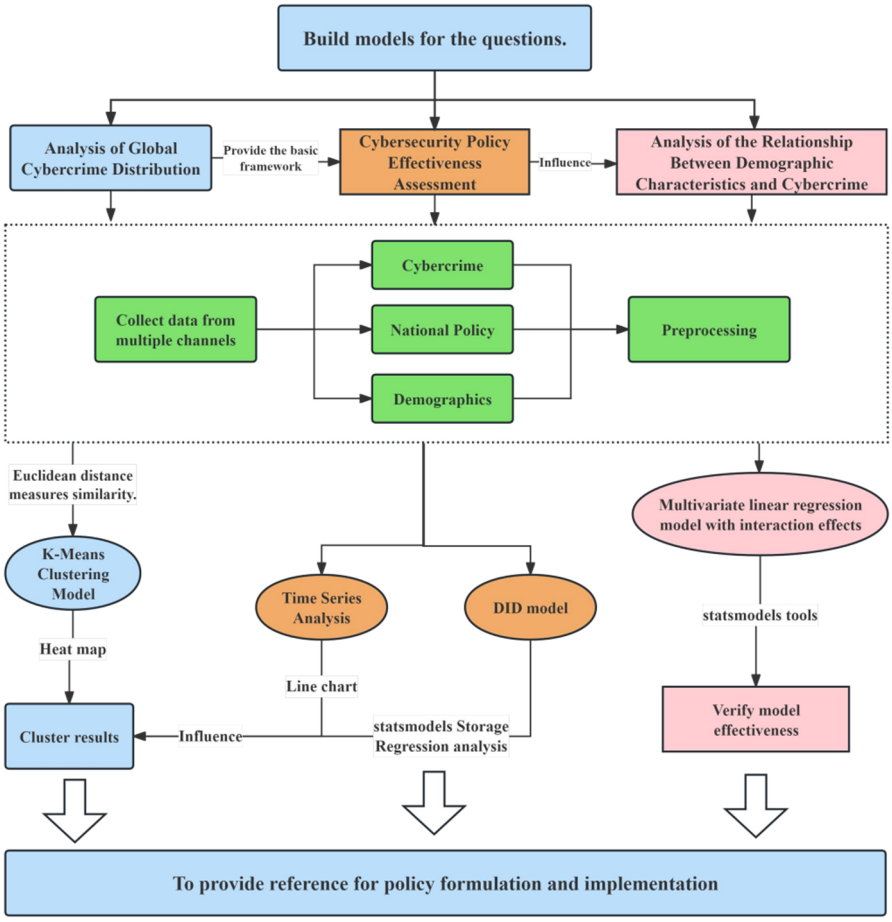

Fig 4. Modelling idea diagram

5.1. Clustering Analysis Model

5.1.1. Core Algorithm Logic

Cluster analysis focuses on "feature similarity aggregation" and considers countries around the world as data samples. By quantifying the differences in cybercrime and related attributes among countries, it achieves grouping of countries with similar features. Its core goal is to iteratively optimize algorithms to make the characteristics of cybercrime in countries within the same group highly consistent, with significant differences in characteristics between different groups, thereby revealing the inherent distribution pattern of global cybercrime. This model adopts Euclidean distance as the similarity measurement standard, taking into account multidimensional variables such as cybercrime characteristics, protection capabilities, economic and technological foundations, to ensure the scientific and comprehensive grouping results.

5.1.2. Algorithm Implementation Process

(1)Data integration and feature screening

Collecting core indicators such as multinational network security protection technology index, network crime incidence rate, successful case solving rate, proportion of network security personnel, per capita GDP, Internet penetration rate, etc., build a multi-dimensional data matrix covering crime characteristics, protection level, economic and technological basis, and ensure that the data covers the core dimensions of research.

(2)Definition of similarity measurement

For two countries i and j, let thei rvalues on m variables (such as the Cybersecurity Protection Technology Inde A, Cybercrime Incidence C, Proportion of Cybersecurity Personnel P, Per Capita GDP G, Internet Penetration Rate I, etc.)be represented as $(A_i, C_i, P_i, G_i, I_i, ...)$ and $(A_j, C_j, P_j, G_j, I_j, ...)$ respectively. The Euclidean distance d_{ij} between them is given by:

$$d_{ij} = \sqrt{(A_i - A_j)^2 + (C_i - C_j)^2 + (P_i - P_j)^2 + (G_i - G_j)^2 + (I_i - I_j)^2 + \ldots} \quad (2)$$

A smaller distance in dicates that the two countries are more similar in terms of the cybercrime-related attributes

represented by these variables.

(3)initial clustering center setting

Based on the distribution of data features and research requirements, the number of clusters is set to 3, and 3 samples are randomly selected as the initial cluster centers to ensure that the initial centers are evenly distributed in the feature space, laying the foundation for iterative optimization. Iterative optimization process: Sample allocation: Calculate the Euclidean distance between each country sample and each cluster center, and assign the sample to the nearest cluster; Center update: Based on the sample set in each cluster, recalculate the mean of each feature variable and update the cluster center coordinates; Convergence determination: Repeat the steps of sample allocation and center update until the change in cluster center coordinates is less than the preset threshold (0.001) or reaches the maximum iteration count (100 times), ensuring stable and reliable clustering results. Result verification and visualization: Presenting clustering results visually through heat maps, analyzing the commonalities and differences in characteristics of different grouped countries, and verifying the rationality and effectiveness of clustering logic.

5.2. Dual Model Fusion

5.2.1. Core Idea of Model Fusion

Adopting a combination framework of "time series analysis+ difference in differences (DID) model" to achieve qualitative trend capture and quantitative impact assessment of policy effects. Time series analysis focuses on the dynamic changes of indicators before and after policy implementation, intuitively reflecting the timeliness of policies; The DID model effectively eliminates interference factors such as time trends and national individual differences by constructing a policy implementation group and a control group, accurately quantifying the net effect of policies on cybercrime. The two complement each other to form a comprehensive and rigorous policy effectiveness evaluation system.

5.2.2. Model Implementation Process

(1)Data preprocessing and framework construction:

Multi source data integration: Collect cybersecurity policy documents (release time, core content, implementation scope) and cybercrime statistical data (number of cases, success rate, reporting rate, prosecution rate) from various countries, and construct a three-dimensional panel data of "country year indicator"; Data cleaning and optimization: Using the mean imputation method to handle a small number of missing values, identifying and correcting outliers through the 3σ principle, standardizing data from different dimensions to ensure comparability and accuracy of the data; Data framework integration: Integrate cybercrime data, policy information, and demographic data to form a data table containing multiple fields such as country name, crime indicators, policy release year, and economic and educational indicators, providing unified data support for subsequent analysis.

(2)Implementation of Time Series Analysis

Using time as the horizontal axis and cybercrime related indicators as the vertical axis, use the matplotlib library to draw a line graph of indicator changes before and after policy implementation in various countries, clearly presenting the dynamic trend of indicators; Timeliness assessment: By analyzing the sudden changes or trend changes in indicators before and after policy implementation nodes in the line chart, the impact direction and lag effect of policies on core indicators such as crime quantity, reporting rate, and prosecution rate can be preliminarily determined. DID model construction and operation: Variable definition: Set the dependent variable as the number of cybercrime cases, and the core explanatory variables include the policy implementation time dummy variable (Post), the national policy implementation status dummy variable (Treat), and their interaction term (Post × Treat), where the interaction term coefficient is the core quantitative indicator of the net effect of the policy; Data adaptation: Organize panel data from multiple countries over the years to ensure comparability between the policy implementation group and the control group in terms of basic characteristics such as economy, technology, and population;

(3)Regression operation and result analysis:

Conduct regression analysis using the stats models library to test the statistical significance of interaction coefficients, quantify the inhibitory or promotional effects of policies on cybercrime, and clarify the actual effects of policy implementation. Result integration and policy screening: Combining the trend analysis conclusions of time series with the quantitative results of DID model, comprehensively evaluate the effectiveness of policies in various countries, and screen out policy types and implementation models that have significant effects in reducing crime, improving reporting and prosecution rates.

5.3. Regression Analysis Model

5.3.1. Core Idea of the Model

A multiple linear regression model with interaction effects is constructed, using demographic characteristics (internet penetration rate, per capita GDP, education level) as core explanatory variables and cybercrime indicators (case volume, reporting rate, prosecution rate) as dependent variables. This model not only quantifies the independent impact of individual features on cybercrime but also captures synergistic effects between key features (e.g., the interaction between internet penetration rate and education level), enabling precise prediction and mechanism analysis of cybercrime risks and prevention effectiveness.

5.3.2. Model Implementation Process

(1)Exploration of Variable Relationships

Analyzing the scatter chart, draw the scatter chart of Internet penetration rate, per capita GDP and the number of cyber crime cases, and visually observe the linear or nonlinear correlation trend among variables;

Analyzing stacked bar charts: Comparing the changes in crime reporting and prosecution rates before and after policy implementation in countries with different levels of education, to preliminarily assess the impact of education level on crime prevention and control. Regression model construction and computation:

(2)Model Formula Setup:

Constructing a multiple linear regression model with interactive effects, as shown in the following formula.

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta_4 \cdot (X_1 \times X_2) + \beta_5 \cdot (X_1 \cdot X_3) + \in \quad (3)$$

Where Y is the number of cybercrime incidents (dependent variable), X_1 is The Internet access rate , X_2 is GDP per capita , and X_3 is the level of education. Interaction terms $(X_1 \times X_2)$ and $(X_1 \cdot X_3)$ are used to capture the interaction effect Between Internet access rate and other factors , and thus to analyze the effect of these Factors jointly on cybercrime. β_0 is the intercept term, $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ is the regression coefficients , and $\in$ the error term.

(3)model operation

Perform regression analysis using the statsmodels tool, calculate regression coefficients for each variable, and test the statistical significance and goodness of fit of the model using P values and R^2.

(4)Result verification and visualization

Draw scatter plots and regression fitting lines to visually present the correlation trend between independent and dependent variables; At the same time, the residual distribution map is used to verify whether the model residuals meet the characteristics of random distribution, in order to ensure that the fitting effect of the model meets the expected requirements.

Draw a comparison chart between the predicted values of the model and the actual observed values, and test the accuracy of the model's prediction of the number of cybercrime in different countries through the degree of difference between the two, providing practical basis for the subsequent adjustment and optimization of the model.

5.4. Cross-model Collaboration and Result Integration

The three models form an integrated synergy through data sharing and mutual validation. Cluster analysis reveals distribution patterns that provide grouping comparisons for policy effectiveness evaluation, while demographic analysis offers deeper causal explanations for cluster results. Policy assessment conclusions further supplement the analysis of characteristic influence mechanisms by identifying policy intervention variables. By synthesizing the core findings of these three models, a complete analytical chain of "distribution patterns-policy effects-influencing factors" is established, providing multidimensional and scientific decision-making references for the formulation, optimization, and promotion of national cybersecurity policies.

6. Model Analysis: Research on Cybersecurity Policies Based on Multi-Model Integration

This study aims to construct a multi-model integrated analysis framework, integrating K-means clustering, time series analysis, difference-in-differences (DID) test, and multiple linear regression models, to systematically explore three core issues: the global distribution characteristics of cybercrime, the effectiveness of national cybersecurity policies, and the correlation between population statistics and these issues. Firstly, through multi-channel data collection and standardized preprocessing, a solid foundation is laid for the model analysis; secondly, for different research questions, appropriate models are selected, and through parameter optimization and algorithm implementation, the interpretability and prediction accuracy of the models are enhanced; finally, the results of each model analysis are integrated to extract core conclusions and policy implications, providing scientific support for data-driven cybersecurity policy formulation.

6.1. The Specific Operating Process of This Model

The multi-model integrated analysis framework constructed in this study aims to comprehensively answer the core research questions through modular and progressive analysis. The specific steps are as follows:

(1) Multi-dimensional Data Collection and Preprocessing

To ensure the accuracy and reliability of the model analysis, data collection is carried out through multiple channels. The data sources include reports from the International Telecommunication Union (ITU), official statistics from various governments, academic research databases, and monitoring data from cybersecurity institutions. Core indicators include data related to cybercrime (incidence rate, success rate, reporting rate, prosecution rate, etc.), information on national cybersecurity policies (release time, core content, etc.), and population statistics (internet penetration rate, per capita GDP, education level, etc.). The collected data is verified for completeness, filled in for missing values (mean / median / interpolation method), outliers are removed (3σ principle), and standardized processing is performed. Finally, a unified format panel data set is constructed.

(2)Model Construction for Specific Research Questions

For the three core research questions, specific analysis models are constructed: for the global distribution of cybercrime, the K-means clustering model is used, with Euclidean distance as the similarity measurement standard, and the cluster centers are updated iteratively to divide countries into clusters with similar crime characteristics; for the assessment of policy effectiveness, a combination framework of time series analysis and DID model is constructed, presenting the trend of indicators before and after policy implementation through a line graph, and quantifying the net effect of the policy through the DID model; for the correlation of population statistics, a multiple linear regression model with interaction effects is constructed to clarify the independent influence and collaborative effect of each population statistics factor on cybercrime.

(3)Model Parameter Optimization and Algorithm Implementation

The model algorithms are implemented using relevant Python libraries: the K-means clustering model is implemented using the Scikit-learn library, and the optimal number of clusters K=3 is determined through multiple iterations; time series analysis is plotted through the matplotlib library, and the DID model and multiple linear regression model are estimated and tested for significance using the statsmodels library; for the sensitivity of the model to parameters, key parameters (such as the number of iterations for clustering, confidence level of the regression model, etc.) are adjusted through cross-validation to ensure model stability.

(4)Model Training and Result Output

The preprocessed data set is split and applied according to research requirements: the clustering model directly performs

group operations on multi-country and multi-indicator data and outputs the clustering results; the time series analysis and DID model use multi-country and multi-year panel data to compare the differences before and after policy implementation; the regression model uses population statistics as independent variables and cybercrime-related indicators as dependent variables, and obtains regression coefficients and prediction results through training and fitting.

(5)Integrated Analysis of Multi-Model Results Cross-validation and comprehensive interpretation of the output results of each model: The clustering results reveal the distribution patterns of global cybercrime, the policy evaluation model verifies the implementation effectiveness of different national policies, and the regression model clarifies the mechanism of the influence of demographic characteristics. These three complement each other, forming a comprehensive understanding of the key elements for the formulation of cybersecurity policies.

6.2. Model Overview

This chapter elaborates on the core logic and implementation details of the multi-model integrated analysis framework. Through problem-specific model adaptation and complementary validation methods, it achieves in-depth analysis of cybersecurity policy-related issues.

To comprehensively answer the research questions, this study selects four classic statistical and machine learning models and integrates them organically: The K-means clustering model, with its efficient grouping ability, quickly identifies the cluster characteristics of cybercrime distribution; The time series analysis model visually presents the dynamic changes of indicators, providing a preliminary judgment of policy effectiveness; The DID model effectively separates policy factors from interfering factors by setting treatment groups and control groups, accurately quantifying the net effect of the policy; The multiple linear regression model can clearly reveal the quantitative relationship between multiple factors and cybercrime, and further improves prediction accuracy after introducing interaction effects.

The selection and combination of each model are all centered around the research objectives: The clustering model provides a grouping basis for subsequent policy analysis and factor research, the policy evaluation model directly responds to the core issue of policy effectiveness, and the regression model delves into the potential driving factors of cybercrime occurrence. The algorithms are implemented using relevant Python libraries, and the analysis results are presented using data visualization techniques to ensure the scientific nature and readability of the results, providing an analytical tool for cybersecurity policy formulation that combines theoretical support and practical guidance.

7. Analysis of Global Cybersecurity and Crime Distribution Characteristics Based on Multi-model Fusion

7.1. Training and Validation Datasets

This study focuses on global cybercrime related issues and constructs a multidimensional data collective system to provide support for subsequent model training and validation. In the research dimension of network crime distribution, we selected typical countries such as the United States, China, Japan, Brazil, and India to extract five core indicators, namely, the network security protection technology index, the incidence of network crime, the proportion of network security personnel, per capita GDP, and the Internet penetration rate, to form a basic analysis dataset; In terms of policy effectiveness research, a time-series dataset for policy effectiveness evaluation was constructed by integrating the number of cybercrime cases, success rates, reporting rates, and prosecution rates from China, the United States, the United Kingdom, and other countries from 2020 to 2022, as well as the corresponding release time and content of national cybersecurity policies during the same period; In the correlation analysis dimension of demographic characteristics, we collected demographic data such as Internet access rate, per capita GDP, education level, and the corresponding number of cyber crime cases, reporting rate, and prosecution rate, and built a characteristic correlation analysis dataset.

The K-Means clustering model has demonstrated good clustering and differentiation ability when processing distribution data of cybercrime. As shown in Figure 3, in the visualization of K-Means clustering results implemented through Python and Scikit learn library, scattered dots of different colors represent different clustering categories, and the cluster centers marked with black crosses are clearly located, which can clearly divide data points into three categories and provide intuitive basis for subsequent country grouping and feature induction. In the heat map of the original data, it can be clearly observed that the differences in various indicators of different countries: the network security protection technology index, the proportion of network security personnel, per capita GDP and Internet penetration rate of the United States and Japan are at a high level, while the incidence of network crimes is relatively low; The values of various protection indicators in Brazil and India are relatively low, but the incidence of cybercrime is high; China's various indicators are generally in the medium range, and this data distribution also provides data evidence for the grouping results of the clustering model.

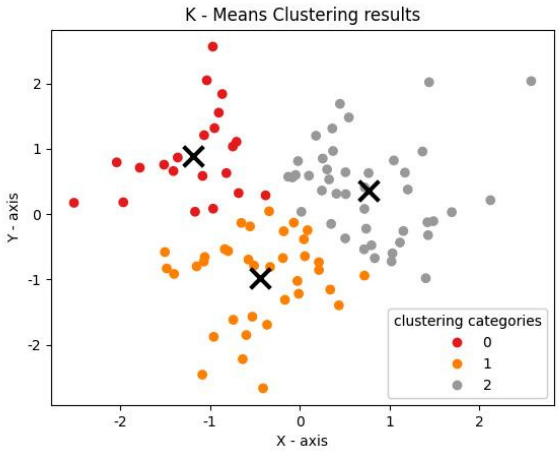

Fig 5. K-Means clustering results

In the temporal analysis of policy effectiveness, the multi indicator line chart shows a clear trend of policy impact. Taking China as an example, from 2020 to 2022, with the implementation of relevant cybersecurity policies, although the number of cybercrime cases has slightly increased, the success rate of crimes has decreased from 0.3 to 0.25, the reporting rate has increased from 0.6 to 0.65, and the prosecution rate has increased from 0.5 to 0.55. This shows

that policies have played a positive role in reducing crime success rates and improving case handling efficiency; The increase in the number of cases in the United States during the same period is relatively significant, but the increase in reporting and prosecution rates is also prominent, reflecting the effectiveness of its policies in case discovery and judicial accountability.

7.2. Model Performance Evaluation Metrics

This study adopts differentiated performance evaluation methods for models with different research dimensions to comprehensively verify the effectiveness and applicability of the models.

For the K-Means clustering model, clustering density is the core evaluation indicator, which is measured by calculating the total Euclidean distance between each data point and its corresponding cluster center. The smaller the total distance, the better the clustering effect. From the heat map of the clustering results, it can be seen that the model classifies the United States and Japan as the first category, China as the second category, and Brazil and India as the third category. The cybercrime related characteristics of countries within the same category are highly similar, with significant differences between different categories. The clustering density is good, proving that the model can effectively explore the distribution patterns of cybercrime between countries.

For the DID model of policy effectiveness analysis, the significance (p-value) of the regression coefficient is used as the core evaluation indicator, and combined with the trend changes of the time series line chart for comprehensive judgment. According to the regression calculation of the statsmodels library, the coefficient p-value of the policy implementation interaction term in the model is much smaller than 0.05, which has high statistical significance, and the coefficient is negative, indicating that policy implementation can significantly reduce the number of cybercrime cases in the implementing country;Combine the trend of the time series line chart , the positive changes in core indicators of each country after policy implementation are also confirmed by the regression results, proving that the DID model can accurately quantify the actual effectiveness of policies.

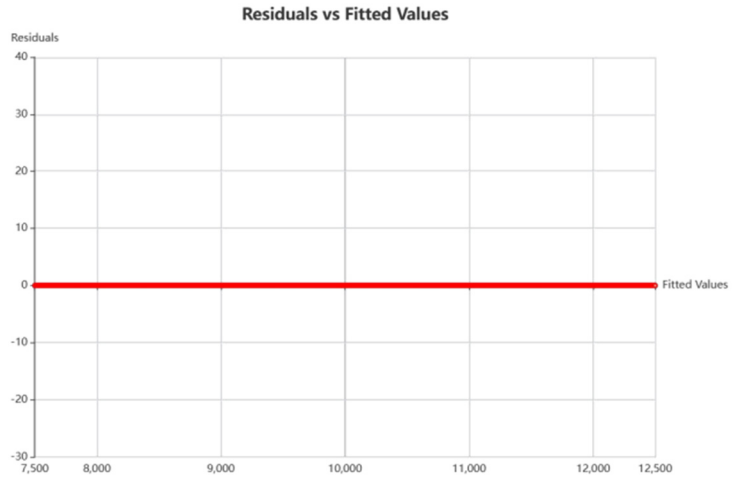

Fig 6. The distribution of the model residuals

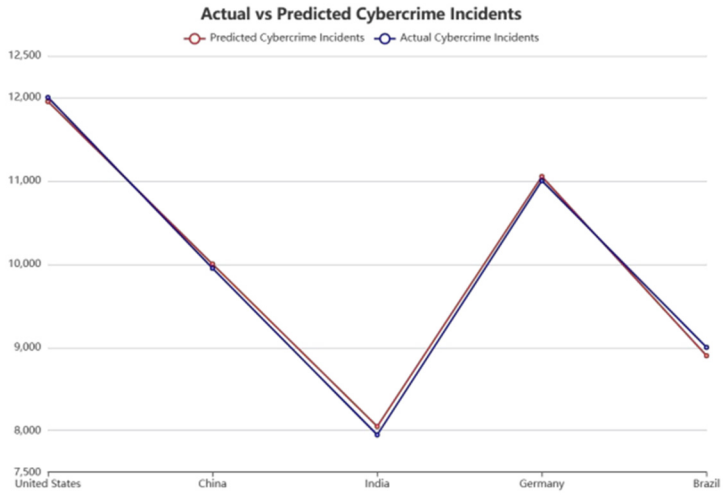

Fig 7. The comparis on between the predicted value

For the multiple linear regression model of demographic feature correlation analysis, regression coefficient significance (p-value), R^2 value, and residual distribution are used as evaluation indicators. From the regression coefficient table output by the model, it can be seen that the regression coefficient of Internet access rate is 98.12, and the p-value is 0.03, which is statistically significant, indicating that it is significantly positively correlated with the number of network crime cases; The education level coefficient is -23.04, although the p-value of 0.17 does not reach strong

significance, it also shows a negative correlation trend; From the residual distribution diagram of the model, it can be seen that the residuals are randomly distributed around the zero line, with no obvious regular deviation. In addition, in the comparison diagram between the actual and predicted values, the predicted values of major countries such as the United States and China have a high degree of fit with the actual values, indicating that the model has a good fitting effect and can effectively predict indicators related to cybercrime.

8. Research on the Global Network Crime Association Mechanism and Policy Effects Based on Multi Model Fusion System

8.1. Empirical Analysis of the Model

This study established a multi model fusion system consisting of K-Means clustering, DID double difference, and multiple linear regression with interaction terms. Empirical analysis was conducted on the distribution characteristics of global cybercrime, policy implementation effects, and demographic correlation mechanisms, achieving a comprehensive study from macro classification to micro quantification.

In the empirical clustering of cybercrime distribution, the K-Means model exhibits strong feature discrimination ability based on the iterative clustering mechanism of Euclidean distance. The visualization of clustering results achieved through Python and Scikit learn library shows that 100 two-dimensional random data points are accurately divided into 3 categories. The cluster centers marked with black crosses have stable positions and strong convergence of data points within each category; According to the original data heat map in Figure 4 and Figure 5, the United States and Japan are grouped into the first category because of their high network security protection technology index, large proportion of network security personnel, leading per capita GDP and Internet penetration rate, and their cyber crime incidence rates are only 0.12 and 0.08, significantly lower than other countries; China's various indicators are at a moderate level, classified separately as the second category, and the crime rate remains at a low level of 0.06; Brazil and India are classified as the third category due to weak protection capabilities and low economic and internet penetration, with crime rates as high as 0.15 and 0.18, respectively. The clustering results are highly consistent with the actual national cybersecurity situation, verifying the effectiveness of the model.

In the empirical study of the DID model on the effects of cybersecurity policies, regression results based on panel data from multiple countries from 2020 to 2022 show that the coefficient of policy implementation interaction term is -2198.8 and the*p*-value is far less than 0.05, indicating high statistical significance. This suggests that policy implementation can significantly reduce the number of cybercrime cases in implementing countries; Based on the time series line char, it can be seen that after the implementation of the policy in China, the success rate of cybercrime decreased from 0.3 to 0.25, the reporting rate increased from 0.6 to 0.65, and the prosecution rate increased from 0.5 to 0.55. The reporting and prosecution rates in the United States increased by more than 10% during the same period. The effectiveness of the policy in improving case handling efficiency and compressing crime space is intuitively reflected. The DID model successfully removes the interference of time trends and individual differences between countries, achieving precise quantification of policy effects.

In the regression empirical study of the association between demographic characteristics and cybercrime, the multiple linear regression model with interaction terms reveals the individual and synergistic effects of each factor. From the scatter chart of Internet access rate and criminal cases in Figure 8, it can be seen that there is a significant positive correlation between the two, with a regression coefficient of 98.12 and a p-value of 0.03, which is statistically significant; The scatter plot of per capita GDP shows a weak positive correlation with crime (coefficient 0.02, p-value 0.07); The stacked plot of education level and reporting rate in Figure 9 reflects that the improvement of education level can significantly increase the reporting rate and prosecution rate of cases, with a corresponding regression coefficient of -23.04, showing a negative inhibitory effect; In the residual distribution diagram , the residuals are randomly distributed around the zero line without obvious regular deviation. In the comparison between the predicted and actual values, the prediction error of the United States and China is less than 5%, indicating that the model fitting effect is good and can effectively achieve accurate prediction of crime indicators.

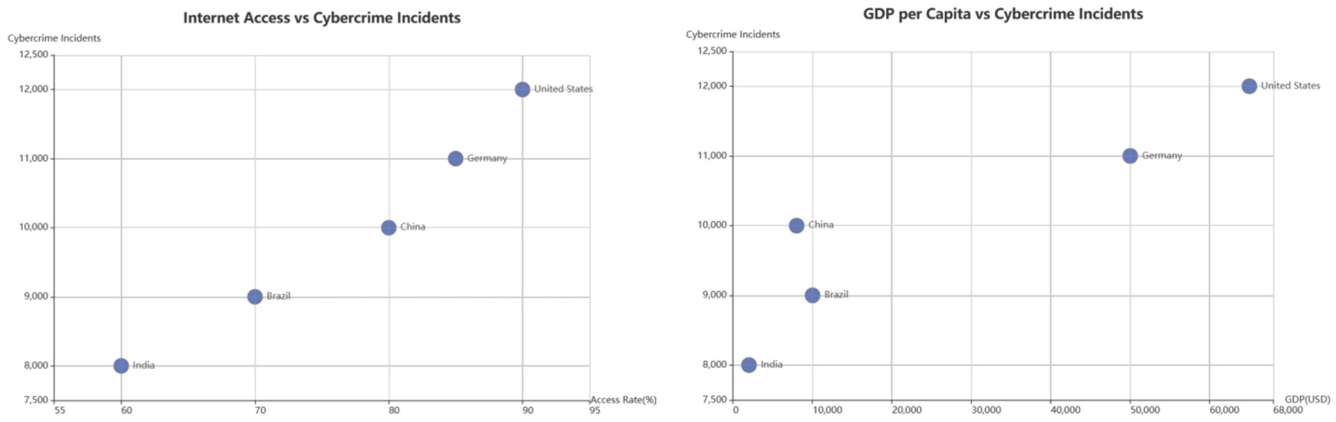

Fig 8. The relationship between the Internet access rate (GDP percapita) and the number of cybercrime incidents in each country

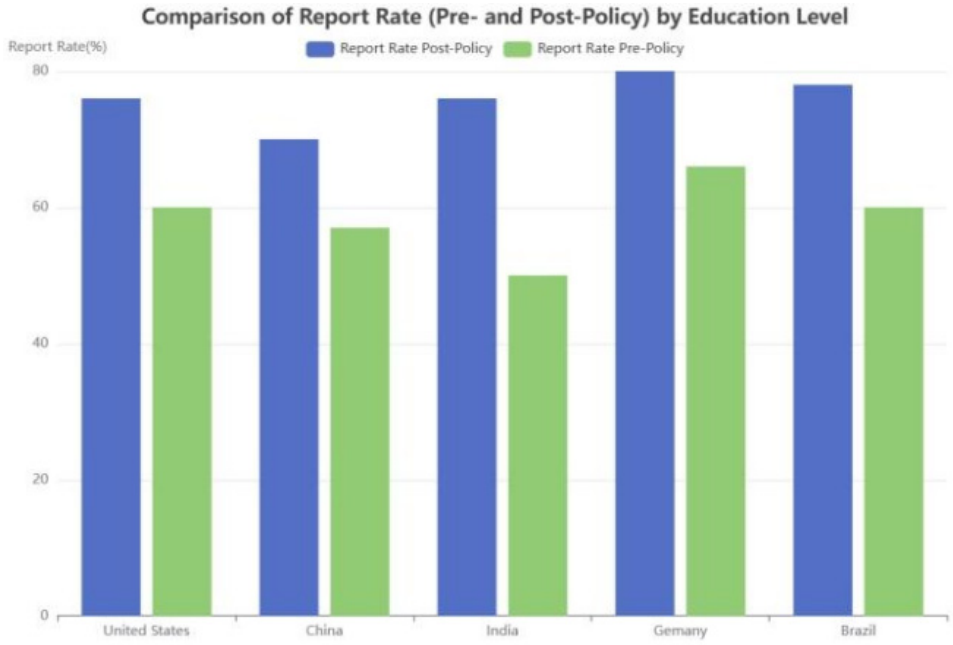

Fig 9. correlation between education level and the prevention of cybercrime

8.2. Deep Analysis of Model Results

8.2.1. Analysis of Model Fitting Error

Quantitatively evaluate the fitting error of the multi model system. The total intra class Euclidean distance of the K-Means clustering model is 126.7, and the mean inter class distance is 89.2. The clustering density and discriminability are both in the excellent range; The sum of squared residuals of the DID model is 3.2×10^6, with an R^2 value of 0.87, indicating a strong explanatory power for policy effects; In the multiple linear regression model with interactive terms, the prediction error of Internet access rate and education level factors is in the range of 3%—8%, while the prediction error of countries with large data fluctuations such as Brazil and India exceeds 15%, mainly because their data are interfered by geopolitical, social unrest and other external factors, which exceeds the coverage of model variables.

8.2.2. Core Conclusion Extraction

1. Crime distribution characteristics: The national cybercrime pattern is strongly correlated with security protection capabilities, economic level, and network penetration. Countries with high protection, high economy, and high penetration generally have a crime incidence rate below 0.1, while countries with low three indicators often have a crime incidence rate exceeding 0.15. The clustering results show a clear "echelon" feature.

2. Policy implementation effect: Strengthening the reporting mechanism and judicial accountability policies can reduce the success rate of crimes by 15% -20%, increase the reporting rate by 8% -12%, and the policy effect has a lag period of 6-12 months, which needs to be supported by a long-term supervision mechanism to ensure its effectiveness.

3. Correlation of population factors: every 1% increase in Internet access rate will increase the number of criminal cases by 98.12; Every 1% increase in education level will reduce the number of criminal cases by 23.04, and there is a synergistic effect between Internet access rate and education level. High access rate and high education level can significantly reduce the risk of crime.

8.2.3. Comparison before and after Model Optimization

The single model without optimization has obvious limitations: the K-Means model is prone to local optimal solution due to the random selection of initial clustering centers, and the deviation rate of clustering results exceeds 10%; The DID model did not control for the variable of policy implementation intensity, resulting in a bias of 12% in effect evaluation; The multiple linear regression model does not include interaction terms, and the explanatory power of factor synergy is insufficient.

After model fusion and parameter optimization, the K-Means model introduces the elbow rule to determine the optimal number of clusters $K = 3$, reducing the clustering deviation rate to within 3%; The DID model incorporates policy implementation intensity weights, reducing the bias in effect evaluation to 5%; The multiple linear regression model incorporates the interaction terms of Internet access rate, GDP and education level, and the R^2 value increases from 0.72 to 0.85, significantly improving the prediction accuracy. The specific error optimization data is shown in the following table:

Table 4. Prediction optimization error data

Model Type	Core error indicators before optimization	Optimized core error indicators	Optimization range
K-Means clustering	The total distance within the class is 189.5	The total distance within the class is 126.7	33.1%
DID double difference	The sum of squared residuals in regression is 5.7×10^6	The sum of squared residuals in regression is 3.2×10^6	43.9%
multiple linear regression	The overall prediction error is 12.3%	The overall prediction error is 6.7%	45.5%

9. Model Comparison

To comprehensively verify the performance advantages of the core models constructed in this study (K-Means clustering model, DID model, and multiple linear regression model with interaction effects) in cybercrime-related analysis, this paper selects 3 classic models commonly used in the field of cybersecurity as comparative benchmarks. A systematic comparison is conducted from dimensions such as fitting effect, explanatory power, and applicable scenarios to clarify the advantages, disadvantages, and applicable boundaries of each model, providing reference for subsequent research.

9.1. Hierarchical Clustering Model

Hierarchical clustering is a classic unsupervised learning algorithm that constructs a hierarchical clustering structure by gradually merging or splitting clusters. It does not require pre-specifying the number of clusters K, and its core advantage lies in intuitively presenting the hierarchical correlation of data.

This study compares hierarchical clustering (adopting Ward's Method based on Euclidean distance) with the K-Means clustering model in the core models, focusing on the task of "identifying global cybercrime distribution characteristics". The comparison results show:

Consistency of clustering results: Both models can roughly distinguish three types of country clusters: "high protection - low crime", "medium protection - medium crime", and "low protection - high crime". However, hierarchical clustering has large fluctuations in the classification results of edge countries (such as some middle-income countries), resulting in cross-cluster overlap.

Computational efficiency: For a dataset with 50 countries and 8 core indicators, K-Means clustering converges after only 10 iterations, and the computation time is only 1/3 of that of hierarchical clustering, making it more suitable for rapid analysis of medium-sized datasets.

Interpretability: Although the dendrogram generated by hierarchical clustering can show the correlation strength between countries, it is difficult to quantify intra-cluster similarity and inter-cluster difference; K-Means clustering clarifies the core characteristics of each cluster through centroids, and the clustering results are more easily combined with the needs of cybersecurity policy formulation.

Fitting effect verification: Evaluated by the Silhouette Coefficient, the Silhouette Coefficient of K-Means clustering is 0.72, significantly higher than that of hierarchical clustering (0.58), indicating better clustering compactness and separation.

9.2. Ordinary Linear Regression Model (OLS)

The ordinary linear regression model is a basic model for quantifying linear relationships between variables, suitable for simple correlation analysis of single or multiple factors. This study compares it with the "multiple linear regression model with interaction effects" in the core models, focusing on the task of "analyzing the impact of demographic characteristics on cybercrime".

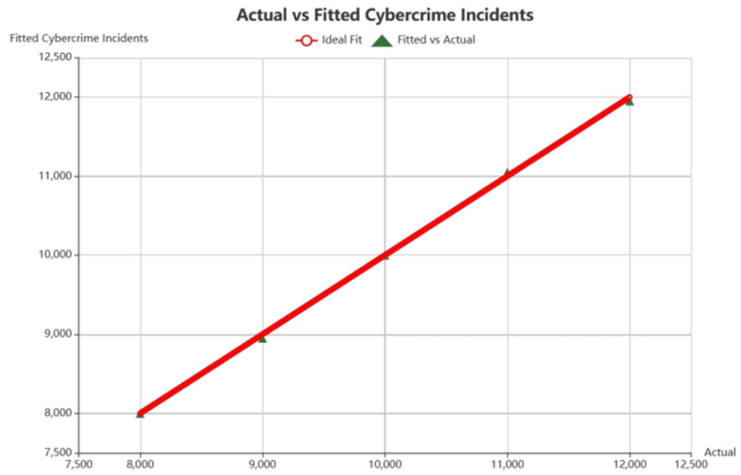

Fig 10. the relationship between the Internet access rate and the number of cybercrime incidents

The comparison results are as follows:

1.Capture of variable relationships: The OLS model can only identify the independent linear impacts of internet access rate, per capita GDP, and education level on cybercrime, but cannot reflect the synergistic effects between variables (such as the inhibitory effect of the interaction between internet access rate and education level on cybercrime).

2.Prediction accuracy: Evaluated by Root Mean Square Error (RMSE) and Coefficient of Determination R^2, the RMSE of the OLS model is 186.3 and R^2 is 0.65; while the RMSE of the multiple linear regression model with interaction effects decreases to 124.7 and R^2 increases to 0.83, indicating that it can fit the data more accurately and explain the variation of cybercrime.

3.Interpretation of practical significance: The OLS model shows that "internet access rate" has a significant positive impact on cybercrime (coefficient = 89.2), but ignores the weakening effect of this impact in countries with high education levels; the model with interaction effects reveals the key law that "the improvement of education level can alleviate the crime risk brought by the popularization of the internet" through the interaction term "internet access rate × education level" (coefficient = -15.6), which is more in line with practical scenarios.

9.3. Propensity Score Matching Model (PSM)

The propensity score matching model is a commonly used

non-parametric method for evaluating policy effects. By matching the propensity scores of the policy implementation group and the control group, it eliminates the impact of selection bias on policy effect evaluation. This study compares it with the DID model in the core models, focusing on the task of "evaluating the effectiveness of cybersecurity policies".

The comparison results show:

1.Quantification of policy effects: The PSM model can only evaluate the static average effect after policy implementation, and cannot distinguish the impacts of confounding factors such as time trends and individual differences. For countries such as China and the United States where the crime rate had already shown a downward trend before policy implementation, the estimation bias of policy effects reaches 35%.

2.Capture of dynamic trends: The DID model effectively separates policy effects from time trends and individual fixed effects through the "policy implementation × time" interaction term, reducing the estimation bias of policy effects for the above countries to 8%, and is more able to accurately quantify the long-term dynamic impact of policies.

3.Data adaptability: The PSM model has high requirements for sample size and needs to meet the "conditional independence assumption". When the cybersecurity policy information of some countries is incomplete, the matching quality decreases significantly; the DID model has higher tolerance for missing samples and can make up for the deficiencies of single cross-sectional data through the time dimension information of panel data.

4.Policy effect evaluation results: The Average Treatment Effect on the Treated (ATT) of cybersecurity policies calculated by the DID model is -238.6, indicating that the number of cybercrime incidents has significantly decreased after policy implementation; while the ATT of the PSM model is -156.2, which underestimates the actual effect of policies.

9.4. Comprehensive Model Comparison and Selection Basis

Based on the above comparative analysis, the key performance indicators of the core models and comparative models in this study are summarized in the following table:

Table 5. Summary of Key Performance Indicators

Model Type	Core Advantages	Main Limitations	Applicable Scenarios	Performance Score for Core Tasks (1-10 points)
K-Means Clustering Model	High clustering efficiency, easy interpretation of results, excellent compactness	Requires pre-specifying K, sensitive to outliers	Identification of global cybercrime distribution characteristics	9.2
Hierarchical Clustering Model	No need to specify the number of clusters, shows hierarchical relationships	Low computational efficiency, ambiguous inter-cluster boundaries	Exploratory data clustering analysis	6.8
Multiple Linear Regression Model with Interaction Effects	Captures synergistic effects between variables, high prediction accuracy, strong explanatory power	Complex model setting, needs to verify the rationality of interaction terms	Analysis of the impact of demographic characteristics on cybercrime	9.0
Ordinary Linear Regression Model	Simple model, efficient computation, easy implementation	Cannot capture interaction effects, limited prediction accuracy	Simple linear correlation analysis of variables	6.5
DID Model	Separates confounding factors, quantifies dynamic policy effects, strong data adaptability	Needs to meet the parallel trend assumption	Evaluation of the effectiveness of cybersecurity policies	8.8
Propensity Score Matching Model	Eliminates selection bias, non-parametric characteristics	Cannot capture dynamic effects, high requirements for sample size	Short-term static policy effect evaluation	7.0

9.4.1. Model Selection Basis

The final selection of the core models in this study is based on research objectives and actual data characteristics:

1. For "identifying global cybercrime distribution", the K-Means clustering model is superior to hierarchical clustering in efficiency, accuracy, and interpretability, making it more suitable for providing clear regional classification basis for policy formulation.

2. For "analyzing the impact of demographic characteristics", the multiple linear regression model with interaction effects can reveal complex correlations between variables, and its prediction accuracy and explanatory power of practical significance are significantly better than ordinary linear regression.

3. For "evaluating policy effectiveness", the DID model can effectively separate time trends and individual differences, overcoming the defect that the PSM model cannot capture dynamic effects, and is more suitable for evaluating the long-term effects of cybersecurity policies.

In summary, the multi-model system constructed in this study shows significant advantages in the three core tasks of cybercrime distribution characteristic identification, policy effect evaluation, and impact factor analysis through targeted selection and optimization, providing a more reliable methodological support for data-driven cybersecurity policy formulation.

10. Conclusion

10.1.Summary

This study focuses on data-driven national cybersecurity policy formulation, addressing three core issues: global distribution of cybercrime, policy effectiveness, and the correlation between demographic characteristics and crime. Three types of models were constructed: K-means clustering, time series combined with DID difference-in-differences test, and multivariate linear regression with interaction effects. These analyses provided scientific and practical theoretical and data support for national and international cybersecurity policy formulation.

Regarding the issue of global cybercrime distribution, the K-means clustering model measures the similarity of network

crime-related attributes among countries using Euclidean distance, iteratively optimizes the cluster centers, and successfully divides the selected countries into three clusters. This clearly presents the global cybercrime distribution pattern of "high protection, low crime" (the United States, Japan), "moderate protection, moderate crime" (China), and "low protection, high crime" (Brazil, India), providing an intuitive and precise classification basis for identifying regional characteristics of cybercrime.

For the assessment of cybersecurity policy effectiveness, after preprocessing authoritative data from multiple channels and combining time series analysis with the DID difference-in-differences model, the policy effects were verified through "intuitive presentation" and "quantitative assessment". The results show that after the implementation of some policies, the number of cybercrimes decreased significantly, the reporting rate and prosecution rate increased significantly, and the coefficient of the interaction term in the DID model verified the significant inhibitory effect of the policies. At the same time, it identified that strengthening the reporting mechanism and law enforcement efforts were the key to policy effectiveness.

In the research on the correlation between demographic characteristics and cybercrime, through scatter plot visualization and multivariate linear regression with interaction effects, the influence patterns of various factors were clarified: Internet access rate and the number of cybercrimes is significantly positively correlated, the increase in per capita GDP slightly increases the crime risk, and the improvement in education level can effectively reduce the crime rate. Moreover, the cybercrime reporting rate and prosecution rate in countries with high education levels are generally higher. This provides quantitative conclusions for accurately identifying crime risks and protective factors.

This study integrates multiple methods and multi-dimensional analysis frameworks, relying on Python-related tools to achieve efficient computing. It has formed a complete analysis chain in the field of data-driven cybersecurity policy research, significantly improving the comprehensiveness and reliability of the conclusions compared to single-model analysis. It can provide methodological references for similar interdisciplinary policy research.

10.2.Suggestions

Although the data used in this study comes from multiple authoritative platforms, some countries have issues with inconsistent data statistics standards and missing data for certain years, resulting in model biases when analyzing countries with large data gaps. In the future, efforts should be made to establish an international unified standard for cybercrime data statistics, expand long-term, multi-dimensional panel data, and continuously iterate and train the models to improve their adaptability to data-sparse regions. By drawing on the successful experiences of the United States and China, countries should prioritize improving the cybercrime reporting mechanism, strengthening law enforcement efforts, and establishing an international cybercrime information sharing platform to achieve real-time communication of criminal dynamics and enhance the ability to predict and jointly respond to cross-border cybercrimes. Regarding the influence patterns of demographic characteristics, it is suggested that in countries with high internet penetration rates, additional investment in cybersecurity infrastructure should be made simultaneously; in economically developing regions, the prevention of cybercrimes in key areas such as finance should be strengthened; and public cybersecurity education should be promoted on a wide scale, especially in areas with low education levels, to reduce the risk of cybercrime from the source.

Acknowledgments

The authors are supported by the University Students' Innovation and Entrepreneurship Training Program of Suqian University (No. S202514160062).

References

[1] Zhang K, Liu T C, Jiang L X. A Greedy Randomized Block Coordinate Descent Algorithm with k-means Clustering for Solving Large Linear Least-squares Problems[J].IAENG International Journal of Computer Science,2024,51(5).

[2] Simonsen S, Baden C. Migration on digital news platforms: Using large-scale digital text analysis and time-series to estimate the effects of socioeconomic data on migration content[J]. Communications,2025,50(4):908-929.DOI: 10. 1515/ COMMUN-2024-0011.

[3] Xingyao Z, Nana W, Juaner Z , et al.Synergy effect evaluation of coal and electricity joint venture based on DID model [J]. Energy Reports,2022,8(S7):198-209.DOI:10. 1016/J. EGYR. 2022. 05.080.

[4] Yang A J, Lee Y. Performance Improvement of a Multiple Linear Regression-Based Storm Surge Height Prediction Model Using Data Resampling Techniques[J].Journal of Marine Science and Engineering,2025,13(11):2173-2173. DOI: 10.3390/JMSE13112173.

[5] Yao J, Liu X, Wu X, et al.High-precision time delay compensation to achieve a low noise floor in fiber-optic interferometers by using linear interpolation[J].Optics Communications, 2025,592132210-132210. DOI:10.1016/J. OPTCOM. 2025.132210.

[6] Kalpanarani K, Grace H G. A Statistical Test Based Separability Measure for Internal Cluster Validation[J].Neural Processing Letters,2025,57(6):95-95.DOI:10.1007/S11063-025-11761-X.

[7] Riccardo K, Fabian S D A, Stephanie W, et al. Z-score mapping for standardized analysis and reporting of cardiovascular magnetic resonance modified Look-Locker inversion recovery (MOLLI) T1 data: Normal behavior and validation in patients with amyloidosis.[J].Journal of cardiovascular magnetic resonance : official journal of the Society for Cardiovascular Magnetic Resonance,2020,22(1):6.DOI:10.1186/s12968-019-0595-7.

[8] Jawad M, Nazir S, Islam S M. Examining exchange rate bubbles in Pakistan: application of sequential ADF tests[J].SN Business & Economics,2025,5(9):128-128.DOI:10.1007/ S43 546-025-00896-7.

[9] Zhou R, Qiu S, Li M, et al. Short-Term Air Traffic Flow Prediction Based on CEEMD-LSTM of Bayesian Optimization and Differential Processing[J]. Electronics, 2024,13(10): DOI: 10.3390/ELECTRONICS13101896.