**Mathematical Modeling and Algorithm Application**
ISSN: 3006-0842 | Vol. 5, No. 2, 2025

# A Study on Time Series Data Forecasting Methods Using ARIMA and BP Neural Network Hybrid Modelling

**Wen'e Duan[1], Zijing Wang[2], Mengqi Ren[2,*]**

[1]School of International Trade and Economics, University of International Business and Economics, Beijing, China

[2]China School of Banking and Finance, University of International Business and Economics, Beijing, China

[*] **Corresponding author**: Mengqi Ren(Email: 202202054@uibe.edu.cn)

**Abstract:** This study constructs a multi-model prediction framework integrating ARIMA and BP neural networks, aiming to reveal the dynamic relationships between variables and achieve accurate trend forecasting. First, multiple linear regression is used to preliminarily analyse the association mechanism between influencing factors and target variables. EWM-TOPSIS is then employed to allocate indicator weights and evaluate performance, providing a basis for feature selection in subsequent model construction. Furthermore, the ARIMA model is used to pre-process time series data and construct the model, enabling trend prediction and verification of model residual characteristics. Additionally, a three-layer neural network structure is constructed using the BP neural network model, with neuron configurations and activation functions set for the input layer, hidden layer, and output layer to enable complex system prediction driven by non-linear features. This framework achieves end-to-end modelling from variable association analysis to dynamic trend forecasting through multi-level data processing and model optimisation, providing a technical paradigm that combines interpretability and predictive accuracy for relevant fields.

**Keywords:** Multiple Linear Regression, EWM-TOPSIS Evaluation Model, ARIMA Model, BP Neural Network Model.

## 1. Introduction

In the study of data-driven complex systems, accurately capturing the dynamic relationships between variables and enabling trend prediction is of critical importance. As cross-domain data interactions become increasingly frequent, traditional single-model approaches struggle to effectively handle the non-linear characteristics and time-series patterns of multi-dimensional data [1].

This study focuses on predictive techniques based on multi-model collaboration, proposing a hybrid modelling scheme that integrates ARIMA and BP neural networks to address data scenarios with time-series characteristics and complex interdependencies [2]. First, a multiple linear regression model is used to quantify the marginal effects of influencing factors on the target variable, providing an interpretable linear correlation benchmark for subsequent models. Second, the EWM-TOPSIS model is employed to calculate indicator weights and process evaluation data, enabling the selection of important input features and differentiated weighting [3]. Furthermore, ARIMA is used to uncover latent patterns in time series data [4], combined with the non-linear fitting capabilities of BP neural networks to simulate dynamic trends in complex systems [5]. The study employs a multi-level model design and data processing workflow to overcome the limitations of traditional methods in variable association analysis and multi-modal data prediction, establishing a technical paradigm that combines statistical interpretability with complex modelling capabilities to support data-driven decision optimisation and risk management.

## 2. Analysis of the distribution and influencing factors of cybercrime

### 2.1. Multiple linear regression model

(1) Model establishment

This paper chose the economic amount of loss $Y$ as a quantitative indicator to finely measure the magnitude of the damage caused by cybercrime to the country. In order to explore the relationship between cybercrime and country characteristics, the fitting function is designed to take into account various factors including Internet penetration, GDP per capita, education level, and cybersecurity investment, which correspond to the following $X_1$to $X_4$ in turn.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \quad (1)$$

Thus, this paper successfully built a global cybercrime distribution model.

(2) Testing and evaluation

A detailed analysis of the model statistics reveals that the F-value is 30.534 with a significance level of P of $0.000^{***}$, which is less than 0.05. The model's $R$-squared value is 0.804, indicating that the independent variables can explain 80.4% of the variation in the dependent variable, thus suggesting a high degree of model fit. Furthermore, a detailed examination of the Durbin-Watson test statistic (2.320) within the range of 0 to 4 clearly demonstrates that the residuals are highly independent, thereby enhancing our confidence in the model's explanatory power. The regression equation this paper obtained is

$$y=35.588+0.604x_1-0.255x_2-0.291x_3-0.369x_4 \quad (2)$$

Where $y$ represents economic and financial losses which is measured in USD, $x_1$ is GDP per capita, $x_2$ is internet penetration rate, $x_3$ is education level, $x_4$ is cybersecurity investment.

Cybercrime is notably prevalent in Southeast Asia and Africa. These regions often share characteristics such as low educational levels, minimal cybersecurity investments, and limited international cooperation opportunities. Our empirical analysis confirmed these trends, highlighting the need for targeted interventions. To combat cybercrime effectively, these countries should focus on enhancing education,

increasing cybersecurity investments, and fostering international collaboration.

## 2.2. Global cybercrime patterns

The above analyses have provided us with a preliminary understanding of the basic distributional characteristics of cybercrime on a global scale. In order to gain a deeper understanding of the dynamics of cybercrime, this paper need to further explore the differences in the success and failure rates of cybercrime, as well as in the rates of reporting and prosecution, among different countries and regions.

(1) Success and prevention of cybercrimes

The cybercrime success rate is a measure of the rate at which a cybercrime offence is committed and achieves its intended purpose, and includes types of ransom attacks, data breaches, DDoS attacks, etc. This paper collected success and failure cases and calculated the success rate of cybercrime in different countries. Further, this paper compared the characteristics of the distribution of high and low success rates between countries by Pearson correlation analysis to obtain Table 1.

**Table 1.** Pearson correlation results

| | GDP/capita | CI | PPLR | IPR |
|---|---|---|---|---|
| Success Rate | 0.223 | -0.010 | 0.388* | 0.367* |
| Note: * represent the significance levels of 10% | | | | |

The success rate of cybercrime is positively correlated with GDP per capita and internet penetration rates, while it is inversely correlated with commercial investment in cybersecurity and privacy protection laws.

(2) Reporting and persecution of cybercrimes

This paper used a similar methodology to analyse the reporting and prosecution rates of cybercrime. By comparing countries with high versus low reporting rates, this paper finds that there are substantial differences in the priority given to cybercrime and in the privacy protection policies for victims in those countries. Similarly, by comparing countries with high versus low prosecution rates, this paper identifies insufficient law enforcement resources and challenges in transnational judicial cooperation as key factors affecting prosecution rates.

(3) Pattern recognition

Apply Data Mining Techniques, this paper was able to identify patterns of cybercrime across the globe. First, this paper collected historical data on prosecution rates, Internet penetration rates, and cybercrime rates from four typical countries, the United States, China, France, and Australia. Specifically, this paper used Vector Auto-Regression (VAR) to predict the temporal dynamics of cybercrime in order to assess the impact of policy implementation on crime trends [6]. Below this paper shows the construction of the model using the United States as an example.

The historical data for the United States is inherently not smooth, but the data after first-order differencing is smooth. It satisfies the same-order single-integer premise. Thus, a VAR model can be constructed.

This paper utilized SPSS for automatic order determination, and ultimately constructed the VAR model with a first-order lag:

$$CF = 0.858 - 0.152 * L.PR - 0.292 * L.IPR - 0.626 * L.CF \quad (3)$$

Where $CF$ refers to the first-order difference of crime frequency. $L.PR, L.IPR, and\ L.CF$ represent the first-order lag of prosecution rate, Internet penetration rate, and the first-order difference of crime frequency, respectively.

After constructing the VAR model, the stability of the model can be assessed through the AR root diagram. The results show that all eigenvalues lie within the unit circle indicating that the model is stable. This paper applied a similar approach to construct VAR models and make predictions for the internet crime rates of three other countries. Figure 1 provides an intuitive representation of the expected model results.

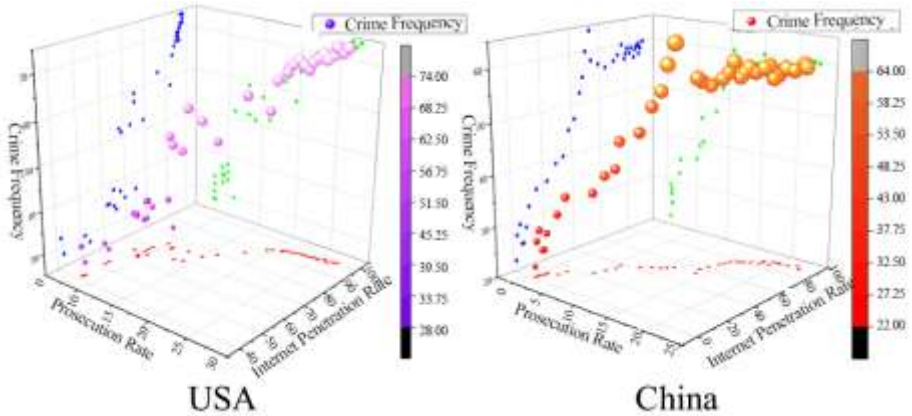


**Figure 1.** Model prediction results (using the United States and China as examples)

From the y-z axis mapping relationship, it can be seen that the growth rate of crime rate is slow and numerically less than 40% until the Internet penetration rate is 50%, and when the Internet penetration rate is higher than 50%, the growth rate of crime rate increases, and the crime rate even reaches more than 70%. However, as shown by the x-z mapping relationship, the rate of prosecution grows as time increases, and the growth rate of Internet crime in all four countries slows down at some point in time. This paper investigated the policy changes at the point in time in these countries and found that they all introduced effective cybercrime control policies at that point.

# 3. Cybersecurity policy effectiveness assessment

## 3.1. Evaluation model based on EWM-TOPSIS

In order to construct a system for evaluating policy effectiveness, this paper selects law, technology, and international cooperation as representative indicators.

After constructing the indicator system and collecting the data, this paper used the entropy-based weighting method (EWM) to build a model to calculate the weights of different indicators. Then this paper multiplied the evaluation indicator data with the weights to get new data. Finally, the new data are used to conduct a TOPSIS method study.

This paper started with the data inversion process. prosecution rate, etc. is inverse data in the indicators constructed above. This paper reversed the data for the reverse indicators in the following way. Where $x'_{ij}$ is the normalised value and $x_{ij}$ is the original value of the ith indicator for the $j$-th country.

$$x'_{ij} = \frac{\max_{ij} - x_{ij}}{\max_{ij} - \min_{ij}} \tag{4}$$

Next, normalisation was done for the data with the aim of solving the quantile problem. After doing the intervalisation, let the data be limited to an interval. Further, this paper calculated the information entropy $e_j$ of the indicator using the following formula.

$$e_j = -\frac{1}{\ln n}\sum_{j=1}^{n}\left(x_{ij} * \ln x_{ij}\right), \left(0 \le e_j \le 1\right) \tag{5}$$

Thus, this paper can derive the coefficient of variation $g_j$ for each indicator with the indicator weights $w_j$.

$$g_j = 1 - e_j, w_j = \frac{g_j}{\sum_{j=1}^{m} g_j} \tag{6}$$

This paper obtained the weighting matrix of the indicators at different levels.

Next, all scores were normalised using our TOPSIS model.

$$z_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}}, j = (1, 2, \ldots, n) \tag{7}$$

Determine the Positive Ideal Solution (PIS) and the Negative Ideal Solution (NIS). In the decision matrix $V$, a higher value of $V_{ij}$ indicates a more favorable alternative. Subsequently, calculate the Euclidean distance of each alternative from the PIS ( $S_{j+}$ ) and the NIS ( $S_{j-}$ ).

$$\text{PIS: } V^+ = \left(V_1^+, V_2^+, \ldots, V_m^+\right) = \left\{\max V_{ij} \mid j = 1, 2, \ldots, m\right\} \tag{8}$$

$$\text{NIS: } V^+ = \left(V_1^-, V_2^-, \ldots, V_m^-\right) = \left\{\min V_{ij} \mid j = 1, 2, \ldots, m\right\} \tag{9}$$

$$S_j^+ = \sqrt{\sum_{j=1}^{m}\left(V_j^+ - V_{ij}\right)^2} \tag{10}$$

$$S_j^- = \sqrt{\sum_{j=1}^{m}\left(V_j^- - V_{ij}\right)^2} \tag{11}$$

Finally, the relative proximity of each country was calculated as the country's policy score.

$$C_j = \frac{S_j}{S_j^+ + S_j^-} \tag{12}$$

Hierarchical analysis and entropy weights are used to derive policy effectiveness rankings for 128 countries.

## 3.2. Data visualization

After collecting data on secondary indicators for 128 countries from 2010 to 2022, the weights were determined based on the entropy weighting method above and ranked using hierarchical analysis. Three representative countries - the United States (the most effective in cybersecurity policy), China (an Asian country with more effective cybersecurity) and Benin (less effective cybersecurity policies) are visualised as shown below.

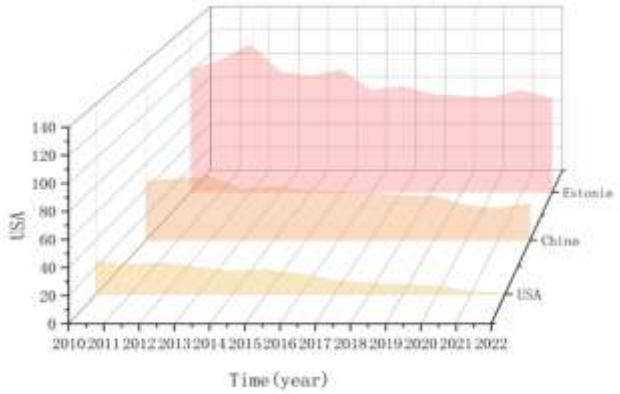


**Figure 2.** Policy score ranking changes for three typical countries

As shown in Figure 2, the United States has basically ranked within the top 20 globally for the last 10 years, demonstrating the strong effectiveness of its cybersecurity policies, while China's ranking has been more stable over the

last 10 years, basically around the 30th. In contrast, Benin's policy effectiveness ranking has improved over time, but still fluctuates around the 80th, as Benin is more vulnerable to cyberattacks, especially in terms of its legal system and technical security capabilities. Some countries in Africa, on the other hand, will have even lower global rankings, usually outside the 100th percentile.

# 4. Time series forecasting of policy effectiveness

## 4.1. The establishment of ARIMA model

Next, this paper will construct an ARIMA model to assess the effectiveness of policies over the next three years. To facilitate the demonstration of the modeling process, this paper will use the United States as an example.

Before applying the model, this paper first preprocess the policy effectiveness score data for the United States from 2011-2022. The Augmented Dickey-Fuller (ADF) test and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test are used to verify that the present data are smooth. Then, combined with the AIC information criterion, a number of potential alternative models were constructed and compared to select, and finally the optimal model was found to be AR(2):

$$C_t = 0.084 + 1.377 * C_{t-1} - 0.559 * C_{t-2} \tag{13}$$

From the results of the Q statistic, the $p$-value is greater than 0.1, then the original hypothesis cannot be rejected at a significance level of 0.1. The residuals of the model are white noise and the model basically meets the requirements.

## 4.2. Case analysis of policy effectiveness forecasting

Based on the AR(2) constructed above, this paper predicted the values of policy effectiveness for the United States for the years 2023-2025. Using a similar approach, this paper predicted for 128 countries and selected a typical three countries for which the prediction results are displayed as in Figure 3.

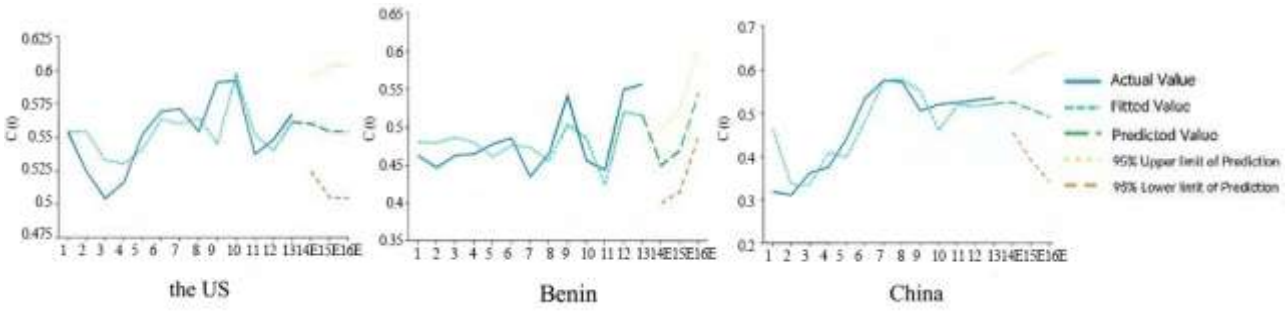


**Figure 3.** Prediction Results of Three Typical Countries

The first picture presents an assessment and forecast of future policy effectiveness in the United States. This paper found that policy effectiveness in the United States is at a high level, and policy intensity is expected to remain stable over the next three years. The second picture shows the movement of policy effectiveness in Africa, where values are low and volatile. This paper expected the volatility of policy effectiveness to be high in the future as well, with poorer traits. The third picture illustrates the movement of policy effectiveness in China. As this paper can see from the graph, the growth rate of China's policy effectiveness has increased significantly since 2012, when the Chinese government introduced a bill on strengthening information protection, and has remained at a high level since then. This movement is strongly related to the credibility of the Chinese government and the strength of its policies. Therefore, this paper predict that China's policy effectiveness maintains a trend of steady growth.

# 5. Predictions of future cybercrime rate based on BP neural network

## 5.1. The establishment of BP neural network model

This paper obtained the actual and predicted values of the policy scores for each country using integrated assessment and time series model. To explore the correlation between policy effectiveness scores and Internet crime rates, this paper established a BP neural network model.

The neural network structure has three layers. The input layer consists of three neurons corresponding to policy scores, GDP per capita, and school enrolment; the hidden layer contains ten neurons which are nonlinearly transformed using a Sigmoid activation function. The role of the hidden layer is to extract the complex relationships of the input features to provide richer information for the output layer. The output layer comprises of one neuron and outputs the predicted crime rate using the Sigmoid activation function, with the values of the output layer restricted to be between 0 and 1 for easy interpretation and comparison. The final training data was evaluated well and the test data was evaluated well, which means that there was no overfitting and the model was usable.

## 5.2. Case studies of model results

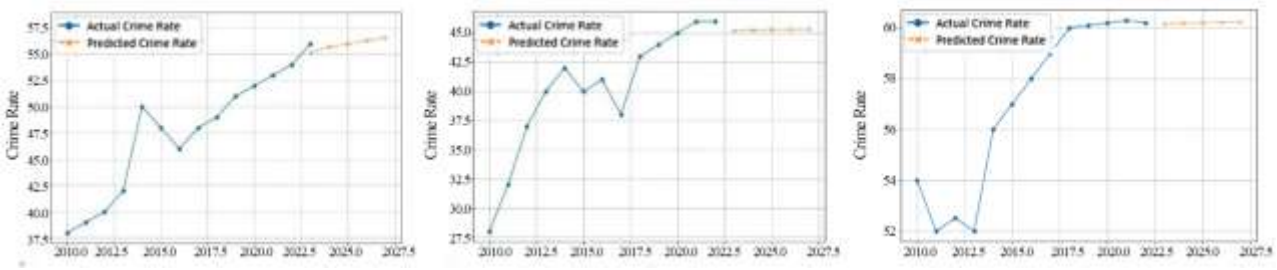


**Figure 4.** Prediction Results

As shown in Figure 4, this paper found that Internet crime rates in the United States are expected to remain high in the future. This result is due to the fact that technology in the United States continues to increase and criminals have found new means of committing crimes, while existing policies have not adapted to these changes. Benin's Internet crime rate is expected to remain low, and although its policies are less effective, the Internet crime rate in this country will remain stable because the low level of Internet penetration.

Clearly, China is a successful case. The Chinese government introduced an important law to protect cybersecurity in 2014, which led to a rapid decline in internet crime. In recent years, the government has been working to increase the intensity of its policies and seek international co-operation. This paper can expect the rate of Internet crime to decrease in the future, although it will still rise, and the last graph confirms our speculation.

Overall, this paper can make a conclusion that an increase in the effectiveness of policies and laws has a negative impact on Internet crime rates.

# 6. Conclusions

This study integrates multiple models, including ARIMA and BP neural networks, to construct a predictive framework suitable for complex data scenarios, achieving end-to-end modelling from variable association analysis to dynamic trend inference. First, multiple linear regression is employed to analyse the associative logic between influencing factors and target variables. Based on EWM-TOPSIS, indicator weight calculations and performance evaluations are conducted to provide feature selection support for subsequent model inputs. Subsequently, the ARIMA model is used to preprocess and model time series data, enabling trend prediction and verification of model residual characteristics. Additionally, a three-layer architecture comprising an input layer, hidden layer, and output layer is constructed using the BP neural network, with neurons and activation functions configured to predict complex systems driven by non-linear features.

The multi-model prediction framework constructed in this study provides a technical approach that combines methodological innovation and practical value for related fields. In the future, the model's dynamic update capabilities for real-time data can be further expanded, or cutting-edge algorithms such as attention mechanisms and graph neural networks can be introduced to enhance the modelling efficiency of high-dimensional heterogeneous data and promote the application of prediction technology in more complex systems.

# References

[1] Feng Xue, Zhang Jinsuo, Zou Shaohui. A Review of Energy Demand Forecasting Models Based on Multi-source Information Fusion [J]. Statistics and Decision Making, 2016, (23): 29-32.

[2] Cao Ying, Chen Xu, Zhang Yuebo, et al. Port throughput prediction based on ARIMA-BP neural network combination [J]. Logistics Technology, 2023, 42 (12): 84-91.

[3] Meng Jing, Xia Pei. Evaluation of Investment Decisions for New Energy Power Generation Projects Based on the EWM-TOPSIS Method [J]. Science and Industry, 2023, 23 (22): 142-147.

[4] Zhou Kun, Xu Yunfei, Qi Haowei. A Short-Term Dynamic Dispatch Model for New Energy Power Generation Based on an Improved ARIMA Model [J]. Computers and Information Technology, 2024, 32 (01): 56-61.

[5] Zhai Shengchang, Han Xiaohong, Wang Li, et al. A SARIMA-GRU Crime Prediction Model Based on a Nonlinear Combination of BP Neural Networks [J]. Journal of Taiyuan University of Technology, 2023, 54 (03): 525-533.

[6] Guo, Chushan. Application of VAR Model Based on Pearson Correlation for Optimisation of Coal-to-Water Ratio in Coal-Fired Power Generation [J]. Science and Technology Innovation, 2024, (03): 223-228.