Information Systems and Economics (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/infse.2024.050225
ISSN 2523-6407 Vol. 5 Num. 2

Research on wild illegal trade based on regression analysis and ARIMA prediction model

Yutong Li[1,*]

[1]*College of Rail Transportation, Suzhou University, Suzhou, China*
[*]*Corresponding author*

Keywords: Illegal wildlife trade, linear regression, ARIMA, predicting models

Abstract: The illegal trade in wildlife is estimated to be worth $26.5 billion annually, making it the fourth largest illegal trade in the world. The main content of this paper is to build a data-driven model to reduce illegal wildlife trade. Searching the governments of all the countries that decided to collect the subject for the five-year project, the search for the United States was particularly prominent, choosing the United States as the main place where the illicit trade took place. Further searches were conducted to collect relevant indicators, classify them, and categorize the main indicators of power, resources, and benefits into categories. Then, three main indicators and four dependent variables were selected to construct partial least squares regression analysis, and the relationship between each dependent variable and the three indicators was analyzed. To describe the difference before and after the 5-year intervention, a linear regression model was used directly. Therefore, other indicators not previously selected were analyzed, and new indicators were further developed based on the description of the three main indicators in the Statistical Yearbook of the United States, including the debt of other sectors to the domestic economy as a new additional driver and energy use as a resource. Start with a direct forecast using regular data to get a look at the next 10 years with and without real-time projects. Using the rules of the five-year plan again, the annual data for the next five years is interpolated to get a data set that can be used to predict the next ten years. The effects of different prediction models were weighted. To sum up, this paper builds a prediction model of illegal wildlife trade based on data collected from the Internet and public databases.

1. Introduction

The illegal trade in wildlife not only threatens biodiversity, but also has serious negative environmental, social and economic impacts. According to reports by organizations such as the International Union for Conservation of Nature (IUCN) and the World Wildlife Fund (WWF), the illegal wildlife trade includes the unlicensed hunting, transportation, and sale of wildlife and its products. Its scope is vast, covering trade in everything from ivory to rhino horn and tiger skins to rare plants and wood. The illegal wildlife trade is a global illegal activity that generates billions of dollars in annual revenue and is considered one of the world's largest black markets after the trade in drugs, people and arms. Across the globe, governments and international organizations have taken a number of measures to combat the illegal wildlife trade. In addition, many ngos are working

to raise public awareness of the issue of illegal wildlife trade, push for policy changes, and protect wildlife on the ground. Technologies such as big data analytics are also being innovated to monitor illegal trade activities and track illegal wildlife circulation routes. Despite some measures taken, the illegal wildlife trade remains a serious problem. To effectively combat illegal wildlife trade, global cooperation is needed, including strengthening the implementation of laws and policies, raising public awareness, promoting the application of science and technology in wildlife conservation, and strengthening cooperation in cross-border law enforcement operations. In conclusion, the illegal wildlife trade is a complex global problem that requires the joint efforts of the international community, governments, non-governmental organizations, scientific institutions and the public to adopt integrated strategies to address it. Through enhanced international cooperation, the use of modern technology and increased public participation, this criminal activity can be more effectively combated globally and the precious wildlife resources of the planet can be protected.

This paper reviews some important literature on wild illegal trade in recent years. Literature [1] proposes a linear regression model for the analysis of survival time, which may provide a valuable method for studying the life cycle of illegal transactions. Literature [2] compared the application of ARIMA and LSTM models in the prediction of cholesterol and glucose, and the application of these models may provide new ideas for predicting the trend and influencing factors of illegal trade. In addition, literature [3] introduces a hybrid method combining wavelet transform, ARIMA and LSTM models for the prediction of stock price index futures, which may have potential application prospects in the prediction of illegal trade. Finally, literature [4] explores the importance of wildlife farming in balancing economic and conservation interests, which may be instructive in designing illegal trade interventions. Taken together, these literatures provide rich theoretical and methodological support for the study of wild illegal trade based on regression analysis and ARIMA prediction model.

2. Model building and solving

2.1. Data preprocessing and analysis

To identify research subjects for the five-year project, governments were searched using the keyword "illegal wildlife." To highlight the results more clearly and visually represent the frequency of occurrence, a word cloud map was drawn, as shown in Figure 1. It can be seen from the results that the number of searches in the United States is the most, so this paper chooses the United States as the main customer. In order to conduct in-depth research on the United States and its related indicators, periodical search websites such as CNKI, PubMed, Google Scholar, etc. were used to search for illegal wildlife, and then the United States was used as a keyword for a secondary search. Finally, the 16 indexes that occur most frequently in the United States are taken as the index evaluation system of this problem, namely the legal rights index (0= weak, 12= strong). The proportion of all tax items whose duties are expressed to the highest international rate (%); Binding tax rates; Tax revenue (as a percentage of gross national product); Customs duties and other import duties; Other taxes; Imports of ores and metals (% of commodity imports); Net energy imports (percentage of energy use); Nurses and midwives (per 1,000 people); Human capital Index (value range 0-1); Total public expenditure on education, total (% of GDP); Gross domestic savings (current local currency); Population growth (annual percentage); Coverage of the public credit system (percentage of adults); High-tech exports (% of manufacturing exports); High-tech exports (current US $); Computer, communications and other services (% of imports of business services); Gini coefficient; Inventory change (constant LCU).

The Kolmogorov-Smirnov test is a non-parametric statistical test method that is used to test whether the data set obeys a certain distribution. The most commonly used test is to test whether the

data set obeys a normal distribution. The basic principle is to compare the cumulative distribution function of the data set with the theoretical distribution function, and determine whether the data set conforms to the theoretical distribution by calculating the maximum difference between the two. If the maximum gap is less than a certain critical value, the data set is considered to obey the theoretical distribution. The single-sample K-S test is used to test whether the observed distribution of a data is a known theoretical distribution. When the difference between the two is small, it is inferred that the sample was taken from a known theoretical distribution.

Figure 1: Data Collection

We assume, H: X obeys certain one-dimensional continuous distribution F. The testing statistical value:

$$Z = n \max_i(|F_{n-1}(x_{i-1}) - F(x_i)|, |F_n(x_i) - F(x_i)|) \tag{1}$$

If the proved to be real, Z convergences to Kolmogorov-Smirnov distribution according to distribution, namely:

$$Z \to \sup |B(F(x))| \tag{2}$$

The result of the KS test is usually a p-value. If the p-value is less than the significance level (generally 0.05), the null hypothesis is rejected, that is, the two samples are considered to come from different distributions.

To identify the distribution of the different parts of the collected data set, a q-q graph was generated using Python, as shown in Figure 2-3. In order to avoid the situation that the actual data does not conform to the normal distribution model, the box diagram is introduced to directly describe the discrete distribution of the data, and provides a standard for identifying outliers, that is, the values that are larger than the upper set of the block diagram or smaller than the lower bound are outliers, as shown in Figure 4.

Figure 2: Non-consistent with normal distribution

Figure 3: Consistent with normal distribution

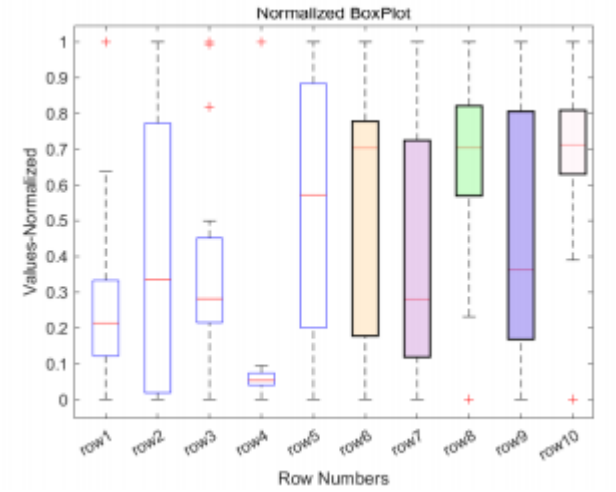

Figure 4: Box-graph

For this missing value, if not used directly, it will certainly have some impact on the result. For nodes where the missing data contains several missing value outliers, Newton interpolation is used to supplement them. Interpolation filling method is used to process the missing values in the data set.

If let n: the number of the missing values; x_i: the ith missing value, the formulae of Newton interpolation to supplement it is:

$$\delta_n(x) = f(x_0) + \cdots + \prod_{j=0}^{n}(x - x_n) f[x_0, x_1, \ldots, x_n] \tag{3}$$

$$f[x_0, x_1, \ldots, x_n] := \frac{f[x_0, x_1, \ldots, x_{n-1}] - f[x_0, x_1, \ldots, x_n]}{x - x_n} \tag{4}$$

δ_n is a polynomial of degree n.

This algorithm has been used to deal with the data gathered for this question. Figure 5 described wildlife trade goods rejection or illegal status prediction.

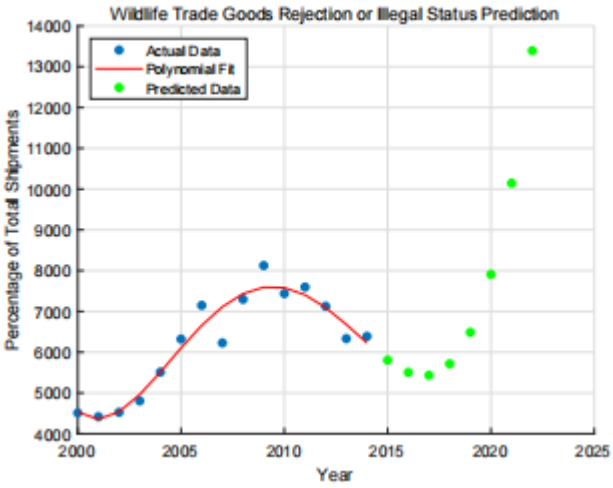

Figure 5: Wildlife Trade Goods Rejection or illegal Status Prediction

2.2. Data Dimensionality Reduction

Before deciding the method or techniques for reducing the dimension of data, KMO test and Bartlett Sphericity Test are conducted before. The greater the connection between indicators, the better the dimensionality reduction effect. Therefore, we use the KMO test and Bartlett The sphericity test verifies the indicators in advance and determines the relationship between indicators. Here we mainly use principal component analysis method to reduce the dimensionality of multi-dimensional indicators. For products that have not been inspected by KMO and Bartlett sphericity test, we use the t-SNE method to reduce the multi-dimensional nonlinear indicators into a two-dimensional sequence to achieve the purpose of dimensionality reduction.

$$KMO = \frac{\sum_{i \neq j} r_{ij}}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} \alpha_{ij}^2} \tag{5}$$

Table 1: Results of KMO test and Bartlett sphericity test

Statistical Variables	Illegal Animal Trade	power	Resources	Interests
KMO	0.658	0.326	0.591	0.524
chi-square(approx.)	80.347	7.688	91.963	102.529
df	6	6	10	15
P	0.000	0.524	0.000	0.000

Table 1 includes the results of KMO test and Bartlett Sphericity test. KMO(short for Kaiser Mayer-Olkin)tests the partial-relativity between different variables, which is a dimensionless quantity ranging between 0 and 1.A higher KMO value indicates that the variables has less partial relativity and factor analysis could be put into use. Bartlett's test of sphericity is used to test whether the overall correlation coefficient between variables is significantly different from zero. In other words, it tests whether the correlation matrix of the original variables is the identity matrix. If the correlation matrix is an identity matrix, that is, all variables are independent, then there are no common factors. As for the parameters, it will not be fit for analyzing Interests and Resources.

Illegal animal trade: the overall correlation coefficient between variables is significantly non zero, and it is suitable for factor analysis analyze.

Power: Overall correlation coefficient between variables is not significant and is not suitable for factor analysis.

Resources: At the 0.1% significance level, the overall correlation coefficient between variables is significantly non-zero, which is very suitable for conducting factor analysis.

Interest: At the 0.1% significance level, the overall correlation coefficient between variables is significantly non-zero, which is very suitable for conducting factor analysis.

To sum up, we use the principal component analysis method to reduce the dimensionality of indicators other than rights.

Figure.6 shows a visualized result of indicators. The percentage of wildlife trade goods that are rejected or deemed illegal as a percentage of total shipments: This variable has a high positive correlation with principal component 1 (0.723) and a certain positive correlation with principal component 2 (0.684). This may mean that illegal wildlife trade is associated with both the underlying factors represented by principal component 1 and principal component 2.

Estimated number of illegally killed elephants per year: This variable has a high negative correlation with principal component 1 (-0.971) and a weaker correlation with principal component 2(0.181). This may indicate that the number of illegally killed elephants is negatively related to an underlying factor represented by principal component 1 (0.925) and a very weak correlation with principal component 2 (-0.07). This shows that the number of world wildlife seizures has a strong positive association with the underlying factor represented by principal component 1. Number of poaching incidents in Africa: This variable has a high positive correlation (0.891) with principal component 1 and a certain negative correlation (-0.285) with principal component 2. The degree of commonality is 0.875, indicating that the variation of this variable is mainly explained by principal component 1, but principal component 2 also has a certain influence. This may mean that the number of poaching incidents in Africa has a strong positive correlation with the latent factors represented by principal component 1, and a certain negative correlation with the latent factors represented by principal component 2.

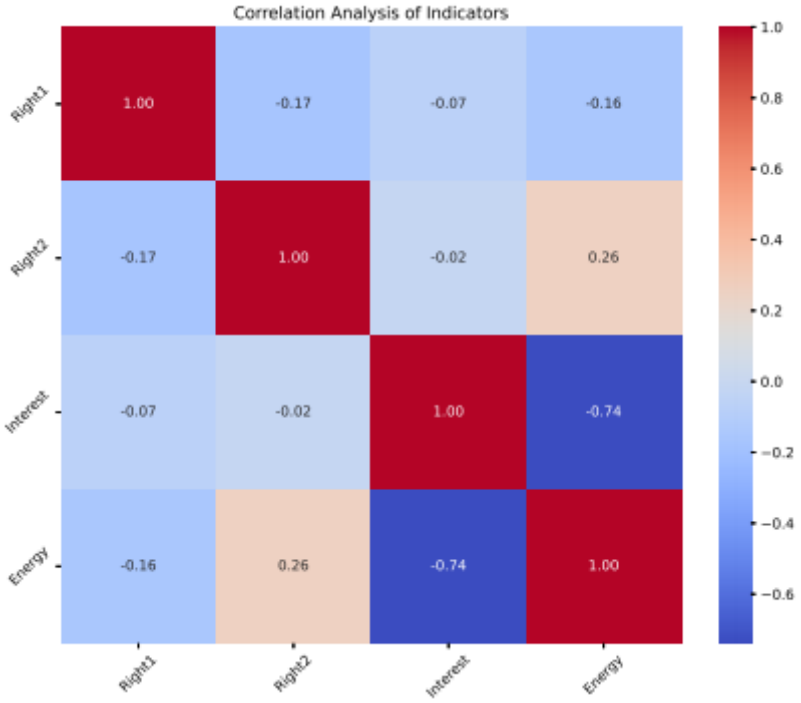

Figure 6: A Visualized Result of Indicators

From Figure 7, we can see that the relationship between the four indicators is not good enough. Therefore, T-SNE is used for nonlinear dimensionality reduction, and the results are shown as Figure 8. The overall time series data is shown in Figure 9.

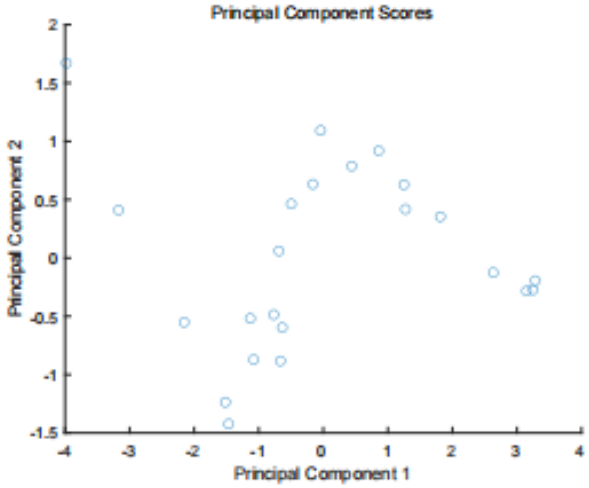

Figure 7: Overall Principle Component Analysis

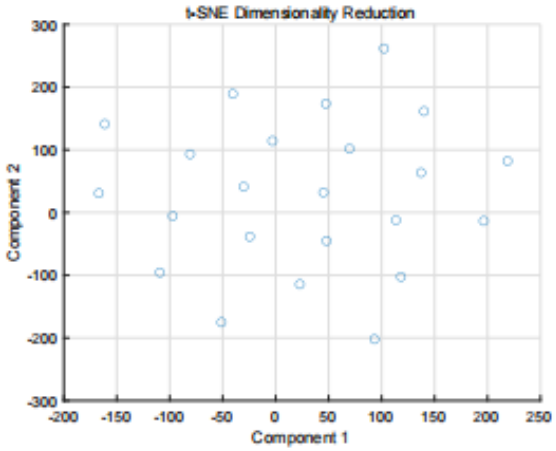

Figure 8: t-SNE Dimensionality Reduction Results

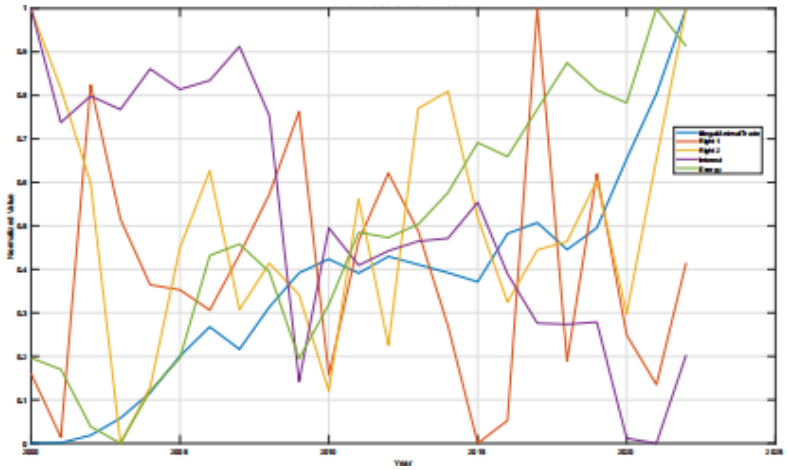

Figure 9: The Overall Time Series Data

2.3. Multiple linear regression model

Through the heat map (Figure 10), we can see that these indicators have a good relationship. Therefore, a multiple linear regression model of four indicators and illegal animal trade was established. The multiple linear regression model has the form resembling to:

$$Y = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4 + c \tag{6}$$

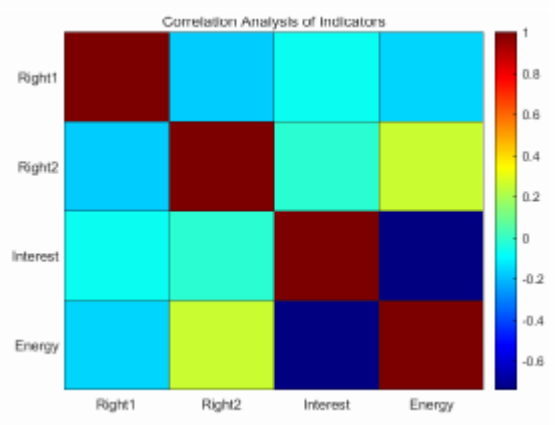

Figure 10: Heat Map Describing the Correlations of Indicators

From which Y represents the amount of illegal animal trade, right 1, right 2, interest and energy are X1 ~ X4 in sequence. c exists for a random error. When we hold other variables constant, interest has the largest negative impact on illegal animal trade, while energy has the largest positive impact on illegal animal trade. Specifically, for every unit increase in interest, the illegal animal trade volume decreases by 0.516 units on average; and for every unit increase in energy, the illegal animal trade volume increases by 0.604 units on average. The impact of Rights 1 and 2 is relatively small, but also positive. The visualizations below figure 11 show the relationship between each of the four indicators

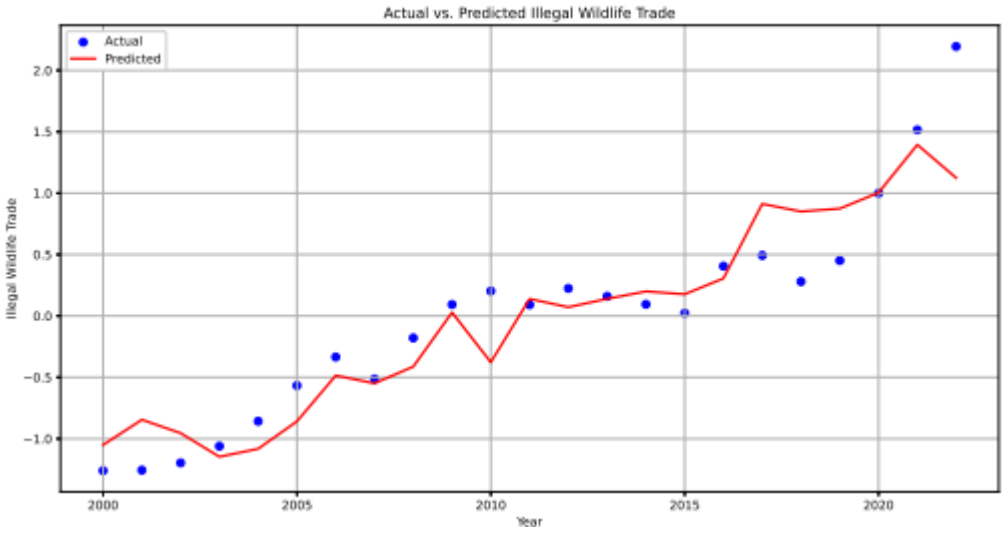

Figure 11: multiple linear regression model

(Right1, Right2, Interest, Energy) and time, through simple linear regression models. Each plot represents one of the indicators over time:

Use a linear regression model to predict the four indicators in the next five years. Bringing the predicted values into the multiple linear regression model of the relational model, single linear regression model results are as Figure 12 shown; the specific results are as Figure13 shown.

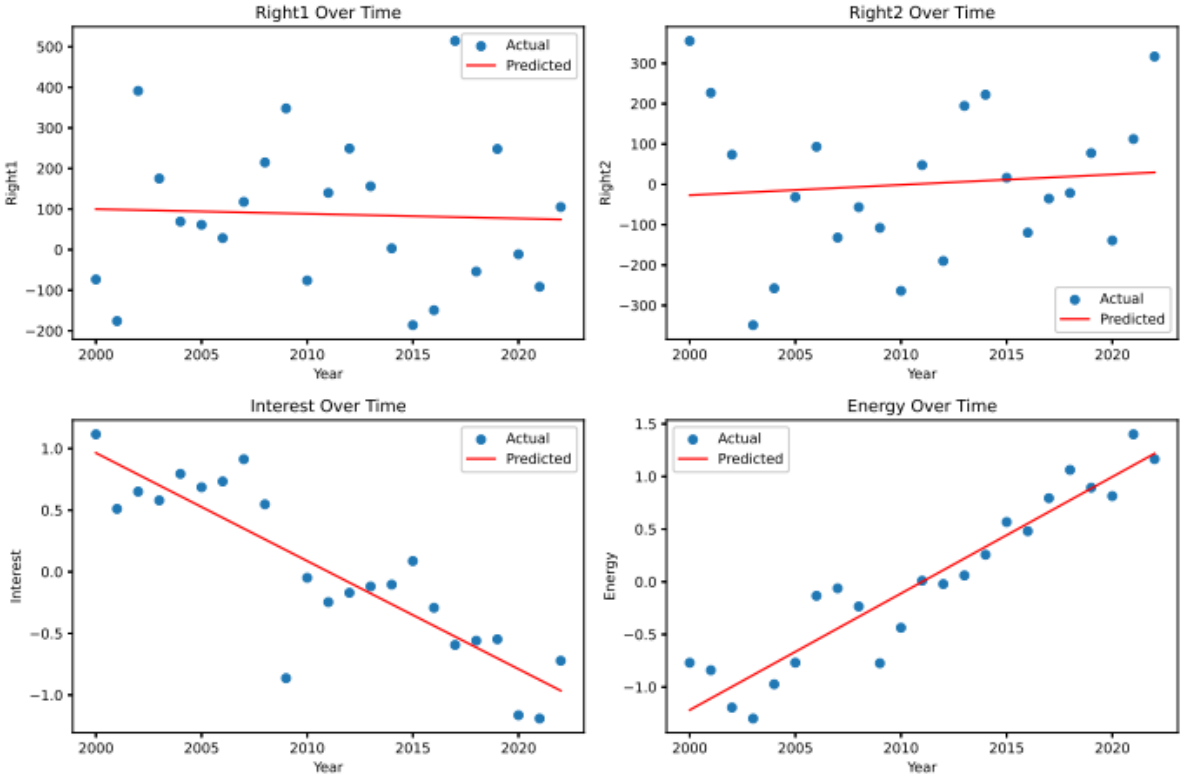

Figure 12: single linear regression model

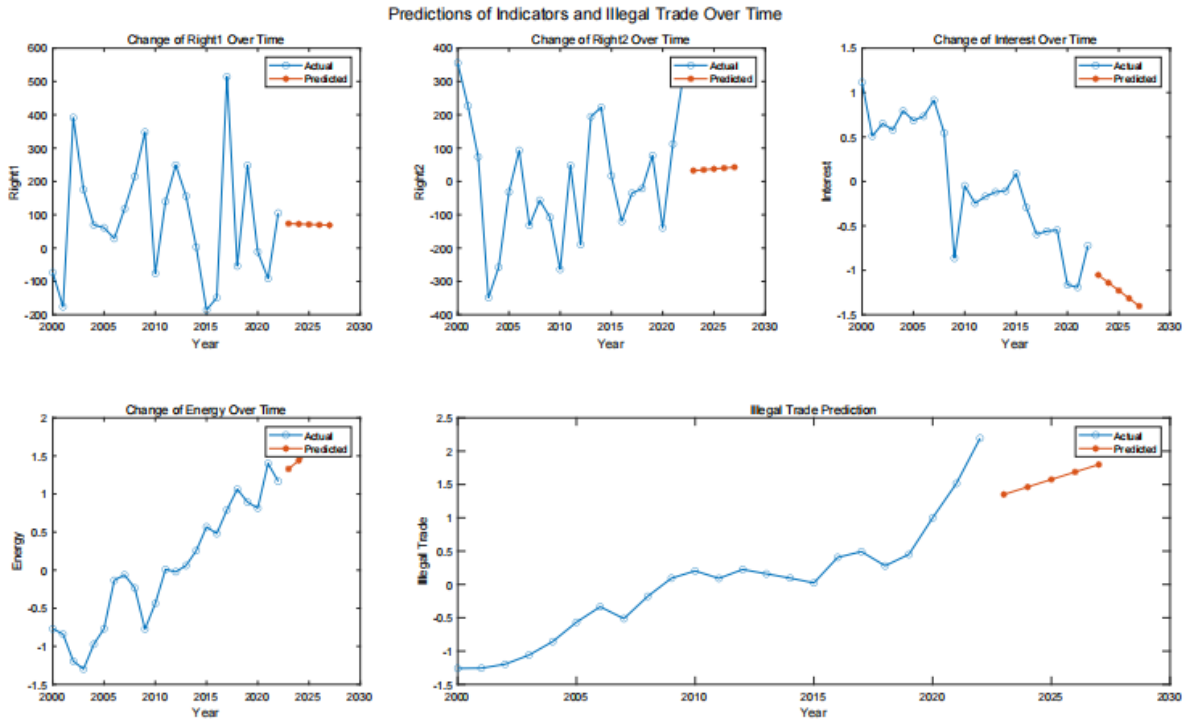

Figure 13: Further Predictions: Multiple Linear Regression Model

A line chart of illegal animal trading indicators with and without project intervention for comparative analysis is drew (Figure14). The solid blue line (marked with a circle) represents the condition with program intervention, and the dashed red line (marked with a cross) represents the condition without intervention. It can be clearly seen from the figure 15 that the predicted value with intervention is significantly higher than the predicted value without intervention, especially starting from 2023, the gap between the two becomes more significant. This shows that the project intervention has a significant positive impact on improving illegal animal trade indicators.

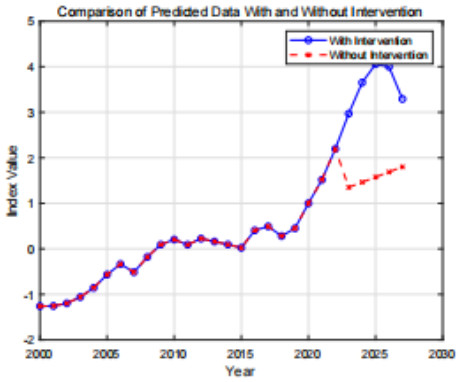

Figure 14: Comparison of Predicted Data with and without Intervention

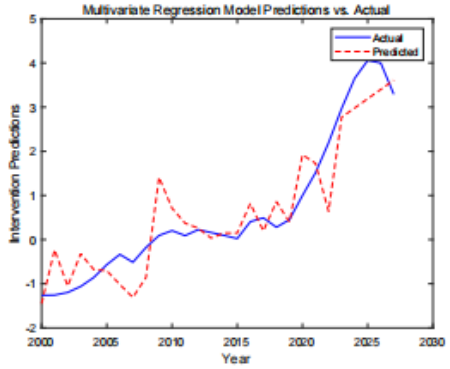

Figure 15: Multi-variable Regression Model: Predictions vs Actual

2.4. ARIMA prediction model

Use the interpolated results in 2027 to forecast the data for the next ten years from 2028 to 2037.

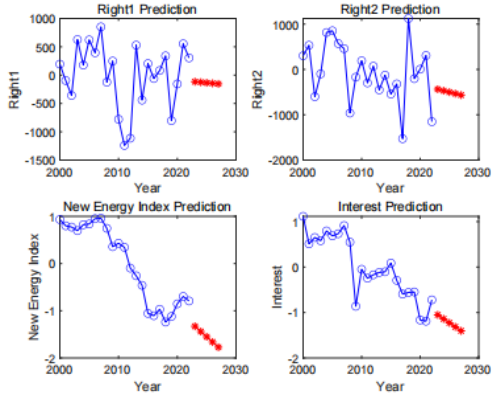

Figure 16: Predictions: Right1, Right2, New Energy, and Interest (to 2030s)

First, a linear regression model is constructed to supplement the data set from 2022 to 2027.Results are included in Figure 16.

A common method to test the stationarity of a time series is the ADF unit root test. The unit root test refers to testing whether there is a unit root in the time series. If a unit root exists in the time series, it is a non-stationary time series. The general formula for testing is:

$$\Delta y_t := y_t - y_{t-1} = \alpha + \beta_t + \delta y_{t-1} + \sum_{i=1}^{p} \xi_i \Delta y_{t-1} + \varepsilon_t \tag{7}$$

t represents time, α, βt, δ are all parameters. p is the number of order left behind, {ε} is white noise sequence. The differentiated data is then used for unit root testing using E views method. Calculating statistical parameters, the conclusion is drawn by Table2: The results of this sequence test show that

Table 2: Results for ADF Testing

Differential order	t	P	AIC	1%critical value	5%critical value	10%critical value
0	-0.869	0.764	1.308	-3.869	-3.087	-2.69
1	-2.914	0.044	2.689	-3.974	-3.052	-2.636
2	-3.469	0.008	-1.788	-3.694	-3.085	-2.682

based on the variables, when the difference is 1 order, the significance P value is 0.044, showing significance on the level, rejecting the null hypothesis, and confirming the sequence is stationary. Noting that {Xt, t ∈ Z} is a non-stationary series.If there exists d ∈ Z+,s.t. ΔdXt = Wt and Wt is an ARMA(p,q) sequence,then Xt is ARMA(p,d,q) sequence.It is clear that:

$$\phi(B)\Delta^d X_i = \theta(B)\varepsilon_t \tag{8}$$

If ΔdXt is a non-stationary series but its mean μ, 0, then ΔdXt − μ is a stationary series with mean zero.It obeys:

$$\phi(B)\left(\Delta^d X_i - \mu\right) = \theta(B)\varepsilon_t, t > d \tag{9}$$

In this situation Xt is defined as ARIMA(p,d,q) sequence.

Firstly,we calculate the sample auto-correlation function and sample partial correlation function for the sample of Xt. If it is censored or tailed, it means it has obeyed the ARMA model. If at least one of the auto-correlation function and the partial correlation function are not censored or tailed, it means that Xt is non-stationary, and calculate its first difference.

Consider ARIMA(p,d,q) sequence {Xt , t ∈ Z}.

When d = 1:

$$X_k(m) = X_k(m-1) + W_k(m) = X_k + \sum_{j=1}^{m} W_k(j) \tag{10}$$

When d = 2:

$$X_k(m) = X_k + m\left(X_k - X_{k-1}\right) + \sum_{j=1}^{m}(m+1-j)W_k(j) \tag{11}$$

Hence, we set up ARIMA (0,1,1) model and gained the results as Figure 17 shows.

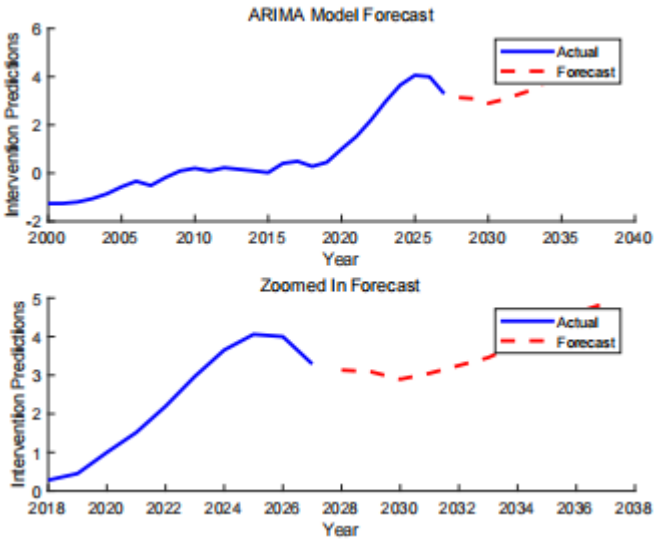

Figure 17: ARIMA Model Forecast

Yi is the series prediction result obtained by using m kinds of prediction methods (each method passes its own test) for the same problem. y′ i j represents the simulated value of the forecast method on the original data.

$$J = \sum_{i=1}^{m} x_m \hat{y}_m \tag{12}$$

$$\sum_{i=1}^{j} x_i = 0, x_i \geq 0 \tag{13}$$

The i-th predicting method and j-th simulation data based on historical data gets a difference:

$$e_{ij} = \hat{y}_{ij} - y_j \tag{14}$$

$$E_j := \begin{pmatrix} e_{1j}^2 & e_{1j}e_{2j} & \cdots & e_{1j}e_{mj} \\ e_{1j}e_{2j} & e_{2j}^2 & \cdots & e_{2j}e_{mj} \\ \vdots & \vdots & \vdots & \vdots \\ e_{1j}e_{mj} & e_{2j}e_{mj} & \cdots & e_{mj}^2 \end{pmatrix} \tag{15}$$

The objective function(the minimized error):

$$\min L := \sum_{j=1}^{n} x^T E_j x \tag{16}$$

The constrains: sum of the weights =1. This kind of mathematical programming will cause a huge amount of computing, multi-goal particle swarm optimization method is adapted. A brief introduction of it can be seen in. Figure18 shows the results of different predicting results.

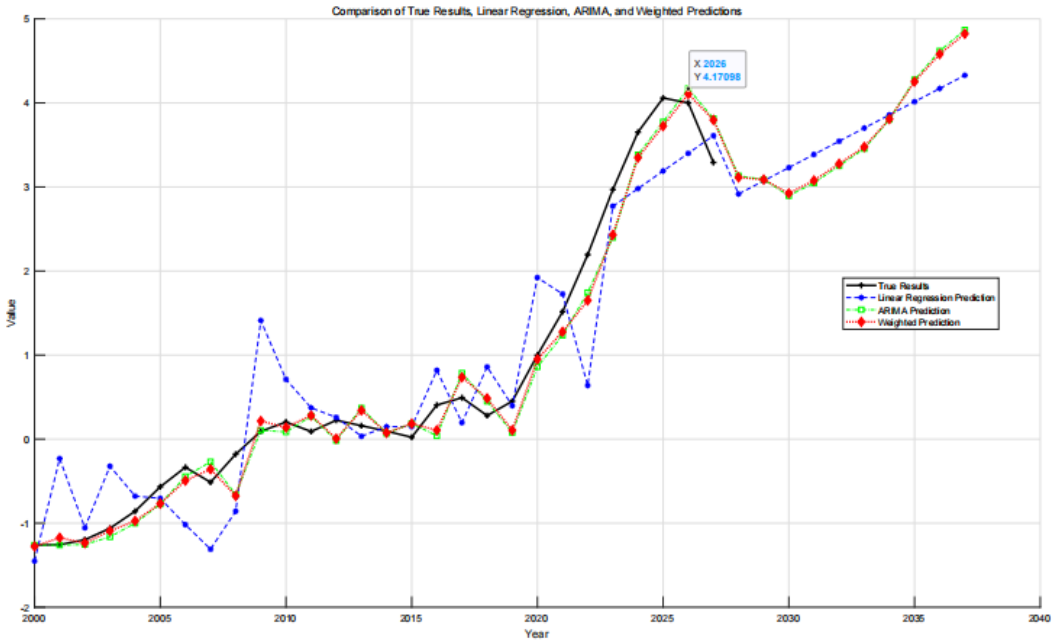

Figure 18: Comparision to Different Predicting Methods

3. Conclusions

The main research content of this paper is to build a data-driven model to reduce the illegal wildlife trade, which provides a comprehensive analysis of the global illegal trade problem through multiple data sources and indicators, thereby providing a deep understanding of the complexity and multi-dimensional nature of the problem. It provides decision-makers with a clearer perspective through advanced data analysis techniques, making their decisions more rational. By creating predictive models, the model is able to predict future trends in illicit trade and the potential impact of interventions, helping policymakers and conservation agencies optimize resource allocation and action plans. Its design allows forecasts to be updated based on new data and information, improving the ability to adapt to changing patterns of illicit trading. However, the model also faces some challenges, including problems with data reliability and integrity, the complexity of the mathematical model, and the uncertainty of the predictions themselves. In addition, the development and maintenance of models requires significant resources, can be biased, and putting effective intervention methods into practice may depend on political, economic, and social factors that are difficult to quantify or carefully consider.

References

[1] Aalen O O. A linear regression model for the analysis of life times [J]. Statistics in medicine, 1989, 8(8): 907-925.
[2] Krishnamoorthy U, Karthika V, Mathumitha M K, et al. Learned prediction of cholesterol and glucose using ARIMA and LSTM models–A comparison[J]. Results in Control and Optimization, 2024, 14: 100362.
[3] Zhang J, Liu H, Bai W, et al. A hybrid approach of wavelet transform, ARIMA and LSTM model for the share price index futures forecasting[J]. The North American Journal of Economics and Finance, 2024, 69: 102022.
[4] Meeks D, Morton O, Edwards D P. Wildlife farming: Balancing economic and conservation interests in the face of illegal wildlife trade[J]. People and Nature, 2024.