Research of Conservation Factors of Wildlife Trade Based on Regression Model

Jiashu Han [1,*], Sihao Chang [2], Hao Wang [1]

1College of Mechanical and Electrical Engineering, Hohai University, Changzhou, China, 213000

2College of Information Science and Engineering, Hohai University, Changzhou, China, 213000

* Corresponding Author Email: 13867571914@163.com

Abstract. The problem of illegal wildlife trade has become an international problem of common concern around the world, and the existing research lacks inquiries on the underlying factors of this problem. In this paper, the ridge regression model, the ARIMA model and the multiple linear regression prediction model are constructed, and the latter two are weighted and optimized. The study finds that the increase of resource input and the increase of national interest play a positive role in combating illegal trade, while the volatility of power factors also has a significant impact on trading activities. The results of the optimized forecasts show that projects to combat wildlife trade have not only short-term effects, but also long-term positive impacts. After sensitivity analysis, the models established in this paper has a high robustness and can provide a reference for the world's anti-wildlife trade activities.

Keywords: Principal Component Analysis, Ridge Regression, Multiple Linear Regression, ARIMA Model, Particle Swarm Optimization.

1. Introduction

Illegal wildlife trade has become the fourth largest illegal trade in the world, second only to drugs, weapons, and human trafficking. It not only threatens biodiversity, but also has profound negative impacts on the environment, society, and economy. According to reports from organizations such as the International Union for Conservation of Nature (IUCN) [1], this illegal activity includes unauthorized hunting, transportation, and sale of various wildlife and their products, such as ivory, rhino horn, tiger skin, and rare plants. A series of measures such as global cooperation are needed to address this issue [2].

Current research has mainly focused on exploring the development and implementation of laws. For example, Josephine Amankwa's team [3] studied the role of legal factors in wildlife conservation and found that in order to increase compliance with forest laws, it is necessary to raise awareness and support for the law through educational activities; the Kwan Kit Yue team [4] conducted a questionnaire survey to understand people's exposure, understanding, and attitudes towards the Marine Species Wildlife Conservation Law (MWPL). Their research and conclusions lack research on underlying factors.

Based on the limitations of current research, this paper will construct a comprehensive evaluation model to systematically analyze multidimensional factors at the national level through methods such as ridge regression, ARIMA model construction, and multiple linear regression prediction. It is expected that the research in this paper will provide scientific basis for policy-making and contribute to global biodiversity conservation and sustainable economic development.

2. Models and Solutions

2.1. Data source

This paper used illegal animal trade as the search term, reviewed 63 articles, established indicators, and collected relevant data for preprocessing. Perform principal component analysis on this indicator data and calculate the comprehensive score for each country. To further construct the model, this

paper takes the United States as an example and searches for a large amount of data through open source websites. This paper collects data on the percentage of wildlife trade shipments rejected or found to be illegal as a percentage of total shipments, estimated number of elephant illegally killed per year, number of world wildlife seizures, and number of poverty incidents in Africa as indicators for evaluating the effectiveness of wildlife conservation [5]. In order to enrich the model and increase its completeness, this paper obtains 20 basic indicators.

In order to conduct a reasonable analysis of the paper, it is necessary to clean the data of each indicator. To determine whether the data follows a normal distribution, this paper uses the KS test method and analyzes it using Matlab programming. According to the KS test results, 14 sets of data follow a normal distribution, while 6 sets do not follow a normal distribution. For data that follows a normal distribution, this paper uses the 3σ method for judgment. For datasets that do not follow a normal distribution, the box plot method is used to determine outliers. After removing outliers, process the missing values and use Newton interpolation to fill in the missing values.

2.2. Data analysis

Firstly, perform dimensionality reduction on the data and use KMO test to determine whether principal component analysis can be performed. Based on the size results of KMO, which indicator can be selected for principal component analysis dimensionality reduction. Afterwards, this paper uses T-SNE method for nonlinear dimensionality reduction, and Matlab programming can be used to obtain the final dimensionality reduction results.

2.3. Establishment of Ridge Regression Model

According to the previous analysis, it can be seen that the four indicators have a good correlation with time. In order to further explore, first conduct a correlation analysis between the dependent variables, and then establish a ridge regression model to explore the linear relationship between illegal wildlife trade (dependent variable) and these four indicators (independent variables) [6]. Illegal animal trade is the dependent variable (Y), with power 1, power 2, interest, and resource as independent variables ($X1, X2, X3, X4$).

This paper constructs the following correlation model:

It is necessary to transform an ill posed problem into a well posed problem by adding a small constant value to the diagonal elements of the matrix, and then taking its inverse to obtain the coefficients:

$$\beta^{\wedge}_{ridge} = (X^T X + \lambda I_n)^{-1} X^T Y \tag{1}$$

Among them, represents the independent variable matrix, Y represents the dependent variable matrix, the superscript T is the transpose, and I_n is the identity matrix.

Cost function $J_{\beta}(\beta)$ adds a penalty to the coefficient values based on RSS, which becomes:

$$J_{\beta}(\beta) = RSS + \lambda \sum_{j=0}^{n} B_j^2 + \lambda \|\beta\|^2 \tag{2}$$

The matrix form is:

$$J_{\beta}(\beta) = \sum_{i=1}^{p} (y_i - X_i \beta)^2 + \lambda \sum_{j=0}^{n} B_j^2 = \sum_{i=1}^{p} (y_i - X_i \beta)^2 + \lambda \|\beta\|^2 \tag{3}$$

Firstly, the K value is determined through the ridge plot, and the model formula is obtained. Then, a univariate linear regression model is established for each of the four indicators, and the univariate linear regression model is used to predict the next five years for the five indicators. Substitute the predicted values into the ridge regression model mentioned earlier. Afterwards, use the Illegal Trade data for three rounds of fitting and use the fitting function for prediction.

2.4. Construction and Solution of Multiple Linear Regression Model

Within the prediction framework based on indicator evaluation system, the primary task is to construct an accurate and comprehensive quantitative indicator system aimed at capturing the dynamic changes of the target system. Subsequently, preliminary predictions were made using the previously validated model. To enhance prediction accuracy, Newton interpolation method is introduced, which effectively processes non-uniform data points by constructing polynomial approximation functions, providing smooth transitions for short-term prediction. Meanwhile, the univariate linear regression model quantifies the linear relationship between independent variables such as time and indicator values, and predicts data for the next ten years from 2028 to 2037, serving as the data foundation for multiple linear regression.

2.5. Establishment of Seasonal ARIMA Prediction Model

ADF test: The commonly used method for testing the stationarity of time series is the ADF unit root test, and its general formula for testing is:

$$\Delta y_t = y_t - y_{t-1} = \alpha + \beta_t + \delta y_{t-1} + \sum_{i=1}^{p} \xi_i \Delta y_{t-1} + \varepsilon_t \tag{4}$$

In the formula, Δy_t is the first-order difference of y_t, and y_t, y_{t-1} are the values of the variables at moments *t* and *t-1*, respectively; $\alpha, \beta, \delta, \xi$ are all parameters. *t* is time; *p* is the lag order; ε_t is a white noise sequence. The results show that the series is a smooth time series when the difference is of order 1.

Model Selection: Let $\{X_T, t = 0, \pm 1, \pm 2, ...\}$ be a non-stationary sequence, If there exists a positive integer d such that $\Delta^d X_t = W_t$, $\{X_T, t = 0, \pm 1, \pm 2, ...\}$ is ARMA(*p,q*), X_T is ARMA(*p,d,q*). At this point X_T is satisfied:

$$\varphi(B)\Delta^d X_i = \theta(B)\varepsilon_t \tag{5}$$

If $\Delta^d X_t$ is a smooth sequence, but the mean value $\mu \neq 0$, then $\Delta^d X_t - \mu$ is a smooth zero-mean sequence, satisfying $\varphi(B)\Delta^d (X_i - \mu) = \theta(B)\varepsilon_t$, *t>d.* Now X_i is named as *ARIMA(p,d,q).* If μ is unknown, it can be estimated using the mean value $\bar{x}$ of $\Delta^d X_t$.

The system automatically searches for the optimal parameters based on the AIC information criterion, and the model results are the ARIMA model (0,1,0) test table. Based on the variable: with project intervention prediction, the analysis of the Q statistic results shows that Q6 does not show significance at the horizontal level, and the hypothesis that the residuals of the model are white noise sequences cannot be rejected. At the same time, the model's goodness of fit R^2 is 0.862, indicating excellent performance and meeting the requirements.

2.6. Establishment of Weighted Optimization Prediction Model

In order to improve the accuracy of the prediction model and better predict the effectiveness of the project. This paper selects two prediction models: ARIMA model and multiple linear regression prediction. Build a weighted comprehensive prediction model for the predicted results to improve accuracy.

Let the predicted outcome vector be $\hat{y}_i (i = 1,2,...,m)$, $\tilde{y}_{ij}$ is the simulated value of the original data $y_i (i = 1,2,...,m)$ for $ith (i = 1,2,...,m)$ prediction method. The weighted prediction values for these m methods are:

$$J = x_1 \hat{y}_1 + x_2 \hat{y}_2 + \cdots + x_m \hat{y}_m \tag{6}$$

In the formula, $\sum_{i=1}^{m} x_i = 1, x_i \geq 0 (i = 1,2,...,m)$.

$$e_{ij} = \hat{y}_{ij} - \hat{y}_i \tag{7}$$

$$E_j = \begin{pmatrix} e_{1j}^2 & e_{1j}e_{2j} & \cdots & e_{1j}e_{mj} \\ e_{1j}e_{2j} & e_{2j}^2 & \cdots & e_{2j}e_{mj} \\ \vdots & \vdots & \ddots & \vdots \\ e_{1j}e_{mj} & e_{2j}e_{mj} & \cdots & e_{ij}^2 \end{pmatrix} \tag{8}$$

Next, calculate the error function L:

$$\begin{aligned} L &= \sum_{j=1}^{n} (\hat{J}_j - y_j)^2 = \sum_{j=1}^{n} (x_1\hat{y}_{1j} + x_2\hat{y}_{2j} + \cdots + x_m\hat{y}_{mj})^2 \\ &= \sum_{j=1}^{n} (x_1e_{1j} + x_2e_{2j} + \cdots + x_me_{mj})^2 = \sum_{j=1}^{n} X^T E_j x \end{aligned} \tag{9}$$

Therefore, an optimization model is constructed with the objective function of minimizing error and the constraint condition of a sum of weights of 1. To reduce computational complexity, particle swarm optimization algorithm is used for solving. Through the model, each particle determines the optimal solution for each individual particle and finds an optimal value from these individual optimal solutions. The formulas for the position and distance of particles are shown in (13) and (14) [7].

$$z_{ij} = z_{ij} + d_{ij} \times rand() \times (pbest_{ij} - x_{ij}) + d_{2ij} \times rand() \times (pbest_{ij} - x_{ij}) \tag{10}$$

$$x_{ij} = x_{ij} + z_{ij} \tag{11}$$

rand() is a random number between zero and one.

$$z_{ij} = \omega \times z_{ij} + d_{ij} \times rand() \times (pbest_{ij} - x_{ij}) + d_{2ij} \times rand() \times (pbest_{ij} - x_{ij}) \tag{12}$$

Adopting a linear decreasing weight strategy as shown in formula (13).

$$\omega^{(t)} = \frac{(\omega_{ini} - \omega_{end})(G_k - g)}{G_k} + \omega_{end} \tag{13}$$

G_k is the maximum number of iterations, ω_{ini} is the initial inertia weight, ω_{end} is the inertia weight when iterating to the maximum evolution number. The weighted prediction result can be obtained using formula (13).

2.7. Analysis of evaluation indicators and project impact

In order to test the prediction results of the forecasting model, this paper uses root mean square error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination (R^2) as evaluation indicators for establishing a time-based economic forecasting model. The results are shown in Table 1:

Table 1. Analysis of Evaluation Indicators for Predictive Models

Evaluation indicators (value)	RMSE	MAPE	$\boldsymbol{R^2}$
Multiple linear regression	0.627	1.528	0.843
Seasonal ARIMA series forecasting	0.084	0.103	0.833
Weighted regression model	0.316	0.541	0.870

From Table 1, it can be seen that the weighted model R^2 is closest to 1, indicating the most accurate prediction.

From the previous analysis, it can be seen that the implementation of this project can effectively reduce the phenomenon of illegal wildlife trade. The measurable impact is based on four data points: the percentage of wildlife trade shipments rejected or found to be illegal as a percentage of total shipments, estimated number of elephant illegal killed per year, number of world wildlife seizures, and number of poverty incidents in Africa, which are reduced by principal component analysis.

3. Results and discussion

3.1. Results of principal component analysis

By using principal component analysis, the results of equations (14) and (15) can be obtained as follows:

$$F1 = 0.232 \times PS + 0.286 \times NA + 0.297 \times WS - 0.312 \times E \tag{14}$$

$$F2 = 1.165 \times PS - 0.486 \times NA - 0.119 \times WS + 0.309 \times EY \tag{15}$$

From formulas (14) and (15), this paper obtains formula (16):

$$F = (0.779 / 0.926) \times F1 + (0.147 / 0.926) F2 \tag{16}$$

PS represents the percentage of wildlife trade shipments rejected or found to be illegal as a percentage of total shipments, *NA* represents the number of poverty incidents in Africa, *WS* represents the world wildlife seizures, and *EY* represents the estimated number of elephant illegally killed each year.

Subsequently, this paper incorporates various national indicators to obtain the rankings of each country as shown in Table 2.

Table 2. Ranking table of principal component analysis

	Country	Composite score	Ranking	Country	Composite score
1	China	1.161	9	Italy	0.086
2	USA	1.07	10	Thailand	0.01
3	Germany	0.829	11	Russia	0.006
4	Japan	0.514	12	Vietnam	-0.016
5	Korea	0.472	13	Poland	-0.07
6	Turkey	0.463	14	France	-0.111
7	Malaysia	0.151	15	Spain	-0.143
8	India	0.124			

The rating results in Table 2 show that China, the United States, Germany, Japan, and other countries are the largest clients of this project. These countries usually have better economic conditions, policy support, and technological levels, which can support the smooth operation of the project. They also have more wild species, and this project can help them better protect local wild species.

Afterwards, we used the T-SNE method for nonlinear dimensionality reduction, and the results are shown in Figure 1:

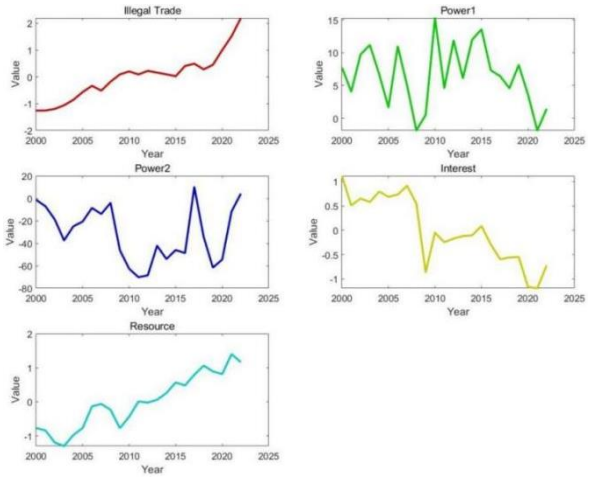


Figure 1. Diagram of data dimensionality reduction results

From Figure 1, it can be seen that the various types of data images after dimensionality reduction exhibit different distribution patterns over time. The two types of data, Illegal Trade and Resource, have increased over time [8], indicating an upward trend in illegal trade and resource volume. The Power1 and Power2 data show significant temporal fluctuations, reflecting the instability of power factors in combating illegal wildlife trade.

3.2. Results of ridge regression and univariate linear regression

In the regression model, the heat map obtained from correlation analysis is shown in Figure 2:

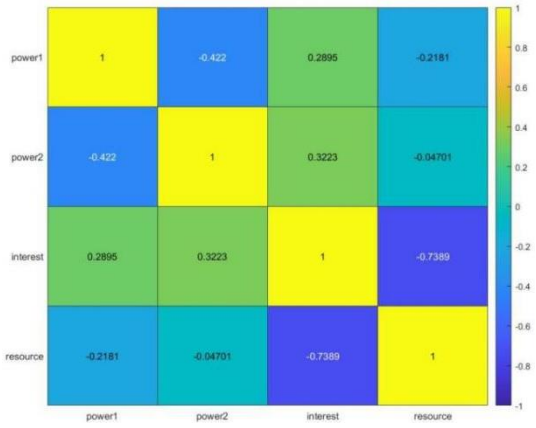


Figure 2. Correlation analysis heat map

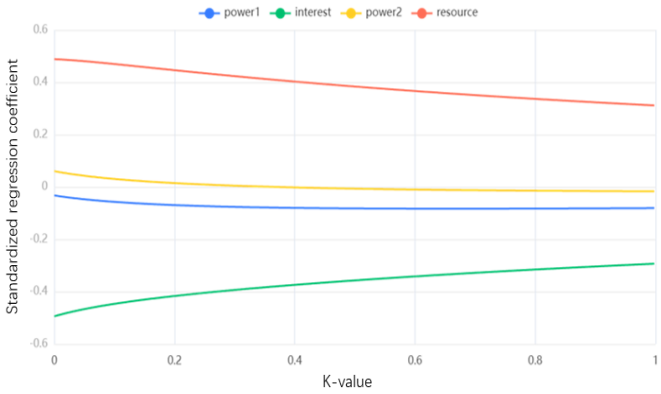


Figure 3. Map of the Ridge Trail

Figure 2 illustrates the correlation between the four variables: power1, power2, interest, and resource. The analysis reveals a moderate negative correlation between power1 and power2. There is a weak positive correlation between power1 and interest, while power1 and resource exhibit a weak

negative correlation with a coefficient of -0.2181. Similarly, power2 shows a weak positive correlation with interest, but there is no significant linear relationship between power2 and resource. Notably, the correlation coefficient between interest and resource is -0.7389, indicating a strong negative correlation, which suggests a significant inverse relationship between these two variables.

The ridge trace map is drawn as shown in Figure 3.According to the variance expansion factor method, K=0.122 was determined, and the results of ridge regression were obtained: based on the F (F=22.91) test, the significance P value was 0.000**, indicating a regression relationship between the independent variable and the dependent variable. Meanwhile, the goodness of fit R^2 of the model is 0.836, indicating that the model performs relatively well.

From Figure 3, it can be seen that the regression coefficient of the resource variable shows a slow decreasing trend with the increase of K value, and eventually tends to stabilize. This indicates that the influence of resource variables on the model gradually weakens during the regularization process. The regression coefficient of the interest variable shows a slow upward trend with the increase of K value, indicating that as the degree of regularization increases, the contribution of this variable to the model gradually increases. The regression coefficients of power1 and power2 power variables change relatively smoothly, with power1 slightly decreasing and power2 remaining relatively unchanged.

Meanwhile, the formula for the model derived from the above analysis is:

$$iTrade = 0.099 - 0.011 \times p1 - 0.551 \times intrest + 0.001 \times p2 + 0.499 \times resource \tag{17}$$

In the formula, *p1* represents the first power indicator, *p2* represents the second power indicator, interest represents the interest indicator, and resource represents the resource indicator. The visualization results are shown in Figure 4:

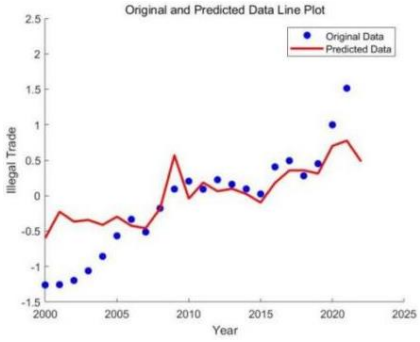


Figure 4. Realized versus projected value of illegal trade

Figure 4 shows that illegal transactions have shown a significant upward trend from 2000 to 2025. During the forecast period, the number of illegal transactions continues to rise, indicating that this problem may become increasingly severe in the coming years.

Afterwards, five univariate linear regression results were obtained using linear regression, as shown in Figure 5:

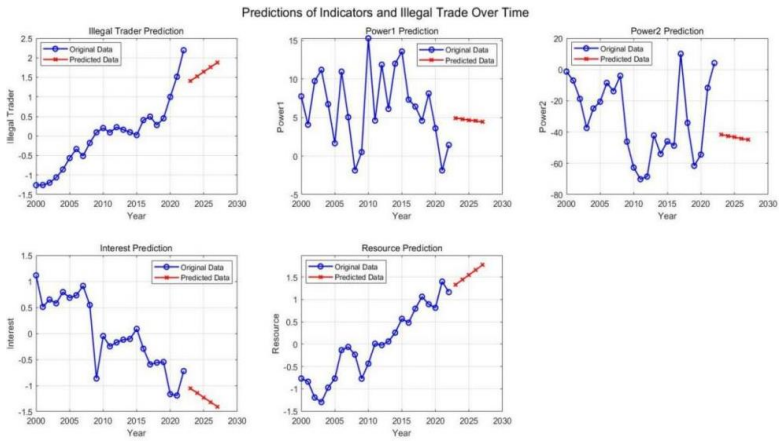


Figure 5. Projected results for the next five years of data

In Figure 5, the predicted graphs of Power1 and Power2 indicate that although the actual data has fluctuated greatly in the past, the volatility of these two indicators will decrease in the future, showing a stable downward trend. Secondly, the Interest prediction chart shows that the interest variable will continue to decline in the future, suggesting that the interest of the country and the public in cracking down on illegal trade may gradually weaken. The Resource Forecast chart indicates that resource investment will continue to increase in the coming years to cope with the growth of illegal trading activities. Perform three rounds of fitting using Illegal Trade data and use the fitting function for prediction. In order to compare the predicted results, plot the two prediction results on the same graph, as shown in Figure 6.

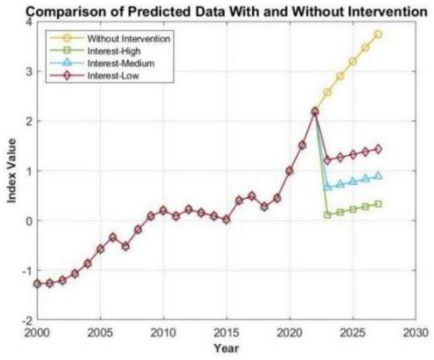


Figure 6. Comparison of forecast results

Figure 6 shows a comparison of predicted data for the illegal trading index at different intervention levels. Figure 6 contains four curves representing the predicted results under four scenarios: no intervention, high intervention, medium intervention, and low intervention. It can be seen that after intervention, the predicted results all showed a decrease, but as the degree of intervention decreased, the final indicator also increased, which is consistent with the actual content. This indicates that if the protection of wildlife is only the primary goal, the more interventions, the better the effect may be.

3.3. Prediction results of multiple linear regression and ARIMA models

Using SPSSPRO to establish a seasonal ARIMA prediction model [9], the results are shown in Figure 7.

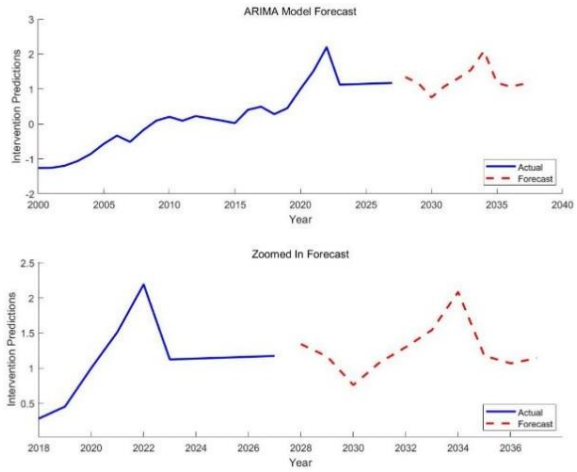


Figure 7. Forecasting results based on time series

From Figure 7, it can be seen that within the five years of project implementation, the representative value of illegal wildlife trade rapidly decreased and tended to stabilize. However, ten years after project implementation, illegal wildlife trade increased, but the peak value decreased compared to before, indicating the immediate and long-term impact of the project.

Meanwhile, the results obtained by constructing a prediction model using data before 2027 are shown in Figure 8, including the comparison between the predicted and actual values of the multiple

linear regression model combined with data from the following decade, as well as the prediction of related time series.

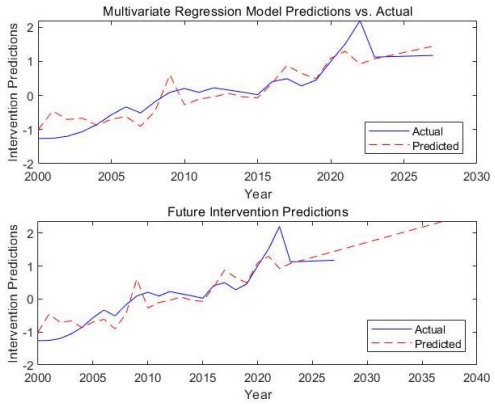


Figure 8. Multiple linear regression model

As shown in Figure 8, the blue solid line represents the actual value, and the red dashed line represents the predicted value. The multiple regression model can fit the actual data well in most years, especially during the period from 2000 to 2015, where the predicted value is highly consistent with the actual value. However, around 2020, there was a significant peak in the actual value, which the model's predicted values failed to fully capture. This may be due to the model's lack of consideration for certain unexpected events or special factors during this period. Overall, the model's prediction of major trends is relatively accurate, demonstrating its good fitting and predictive abilities.

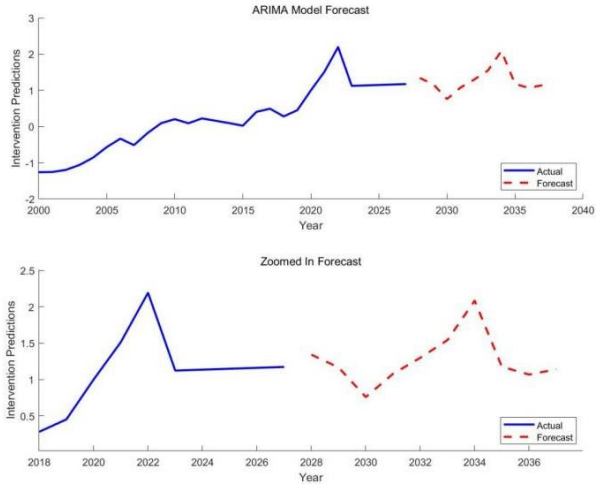


Figure 9. Forecasting results based on time series

Using multiple linear regression and ARIMA models as shown in Figure 9, the illegal wildlife trade data for the next ten years were predicted. The results showed that the illegal wildlife trade volume significantly decreased within the five years of project implementation and remained at a low level for the ten years after project completion. The significant short-term effect is mainly due to the significant investment of resources and the adoption of strong measures to combat illegal transactions in the early stages of the project. The maintenance of long-term effects benefits from the continuous monitoring and management measures of the project. Combining multiple linear regression and ARIMA models can improve the accuracy of predictions, making project implementation more forward-looking and scientifically based. Scientific predictive models can help decision-makers adjust their strategies in a timely manner, maintaining the sustainability and effectiveness of combating illegal trading.

The prediction results indicate that the project not only has short-term immediate effects, but also brings long-term positive impacts. The combination of multiple linear regression model and ARIMA

model has improved the accuracy and reliability of prediction, provided strong scientific basis for policy makers, and promoted the promotion and implementation of similar projects.

3.4. Sensitivity analysis

Sensitivity analysis [10] shows that the model has good adaptability and robustness to different countries and conditions, and even under conditions of data fluctuations, the rate of change in the predicted results is not very high, as shown in Table 3.

Table 3. Table of results of sensitivity analysis

	percentage	percentage	percentage
Rate of change in new resource data	5%	10%	15%
Rate of change of projected results	1.74%	3.48%	6.96%

This good adaptability and robustness mean that the project can be promoted globally, applicable to countries with different economic levels and natural environments. This enhances the universality and feasibility of the project, indicating that it has a high probability of success in both economically developed and developing countries, and contributes to the global fight against illegal wildlife trade.

Through the analysis and discussion of the above results, it can be seen that this study not only provides scientific basis and effective strategies for combating illegal wildlife trade, but also demonstrates the feasibility and effectiveness of promoting the project globally. These research findings provide important guidance for policy makers, international organizations, and relevant stakeholders, helping to promote global cooperation and jointly protect wildlife and biodiversity.

4. Conclusion

Illegal wildlife trade has become the fourth largest illegal trade in the world. This paper constructs ridge regression model, multiple linear regression model, and ARIMA prediction model based on illegal wildlife trade data, and performs weighted optimization on the latter two. Ridge regression found that the intensity of intervention measures is positively correlated with the effectiveness of combating illegal wildlife trade. According to the optimized prediction model, an increase in resource investment and an increase in national interest can effectively reduce illegal wildlife trade, and the volatility of power factors also has a significant impact on such activities. Based on sensitivity analysis, the models developed in this paper demonstrate high robustness and can serve as a valuable reference for global anti-wildlife trade efforts.

References

[1] Jones C, Newsome D, et al. Understanding the conflicting values associated with motorized recreation in protected areas [J]. Ambio, 2016, (45): 323-330.

[2] Rytwinski T, Muir J M, Miller B R J, et al. What is the evidence that counter-wildlife crime interventions are effective for conserving African, Asian and Latin American wildlife directly threatened by exploitation? A systematic map [J]. Ecological Solutions and Evidence, 2024, 5 (2).

[3] Amankwa J, Damnyag L, Banunle A. Determinants of forest law compliances among inhabitants of local communities surrounding tropical wildlife and forest reserve in Ghana [J]. Journal of Cleaner Production, 2024, (443): 964-974.

[4] Yue K K, Ruifang C, Chieh C W, et al. Towards effective wildlife protection law for marine species in China: A stakeholders' perspective [J]. Frontiers in Marine Science, 2023, (10).

[5] Wang Yan. Research on Nature, Humans, and Power within the National Park System [D]. Central University for Nationalities, 2023, (03).

[6] MacGillivray Alexander O, et al. A functional regression analysis of vessel source level measurements from the Enhancing Cetacean Habitat and Observation (ECHO) database [J]. The Journal of the Acoustical Society of America, 2022, 152 (3): 1547-1547.

[7] Yang Wei, Li Qiqiang. Overview of Particle Swarm Optimization Algorithm [J]. Chinese Engineering Science, 2004, (05): 87-94.

[8] Wu Zhaoliang, Jin Min. Research on residents' participation in community environmental governance: Based on the analysis of "situation-process-impact" [J]. Transactions of Beijing institute of Technology (Social Sciences), 2023, 25 (01): 55-66.

[9] Li Lingling, Xin Hao. Application of Time Series Model ARIMA in Data Analysis [J]. Fujian Computer, 2024, 40 (04): 25-29.

[10] Li Zhongsheng, Qin Hongxing. Optimized design of reliability growth planning based on sensitivity analysis [J]. Mechanical Engineering and Automation, 2021, (03): 46-49.