Research on Prediction of Illegal Wildlife Trade Based on PCA and Multiple Linear Regression

Fei Wu*, Siyuan Qi

School of Electronic Information, Soochow University, Suzhou, China

* Corresponding Author Email: 1206586706@qq.com

Abstract. Illegal wildlife trade is a long-standing issue, exerting negative impacts over national economies, regional security, and even global ecosystems. Governments worldwide and numerous environmental and animal protection organizations have been deeply engaged in addressing this issue for many years. Historically, there has been a lack of a quantifiable model to analyze the impact of implementing wildlife monitoring and conservation projects using drones on illegal wildlife trade.In this paper, we primarily propose an Illegal Wildlife Trade Prediction Model. By integrating four secondary indicators related to wildlife trade with the aforementioned 16 indicators, we reduced them to four primary indicators. The methodological tests confirmed their strong correlation, and multiple regression analysis was used to predict illegal wildlife trade, thereby verifying the project's effectiveness. Innovatively, we established a model to measure the probability of project completion, combining data from model predictions, resulting in an approximate completion probability of 60%. In conducting a project sensitivity analysis, we also developed a new model to perturb the three primary indicators, predicting the project's final outcome under random fluctuation of indicators, with variation rates as low as 0.34%, 0.98%, and 1.22%, indicating the project's stability.

Keywords: PCA; Multiple Linear Regression; Wildlife trade; Sensitivity analysis.

1. Introduction

The illegal wildlife trade, a persistent challenge, detrimentally impacts national economic structures, undermines regional security, and poses significant threats to the integrity of global ecosystems. This issue has garnered extensive attention from international governments and a broad spectrum of organizations committed to environmental and wildlife preservation. Valued at an approximate annual sum of 26.5 billion US dollars, it is classified as the fourth most substantial illicit commerce on a global scale. Every year, a variety of related projects are launched; however, identifying the requisite customers and organizations based on the project is a complex task. Only when a project is well-suited to its clients can it achieve optimal results[1].

This study investigates the impact of implementing a project utilizing Unmanned Aerial Vehicles (UAVs) for monitoring and protecting wildlife on illegal wildlife trade[2, 3]. Firstly, by integrating four secondary indicators related to wildlife trade with 16 indicators assessing potential clients of the project, PCA was employed to reduce them to four main indicators, confirming their strong correlation. Multiple regression analysis was then used to predict illegal wildlife trade. Through prediction, it was found that the implementation of the project could significantly reduce illegal wildlife trade over a five-year project period. Finally, a model measuring the probability of project completion was established to validate the effectiveness and stability of the model.

2. Model preparation

2.1. Data Analysis

Data dimensionality reduction is a technique for reducing the number of variables in a dataset, aimed at simplifying data processing and analysis while retaining as much of the original data's important information and structure as possible. This process is particularly useful for dealing with high-dimensional data, that is, data containing a large number of features or variables.

Obtain the results of the KMO test and Bartlett's test using MATLAB.

Table 1. KMO test and Bartlett's test

	Index Name	illegal wild-life trade	power	resources	interset
Bartlett's Test of Sphericity	KMO value	0.641	0.603	0.603	0.634
	Approximate Chi-square	43.703	59.467	91.959	88.723
	df	6	15	15	10
	p	0.00	0.00	0.00	0.00

The KMO values are all greater than 0.6, and the P-values are all less than 0.05, indicating that the correlation coefficients between variables are significantly non-zero, which meets the requirements for principal component analysis (PCA). Through the analysis and calculation of the approximate Chi-square values and degrees of freedom, it can be concluded that the resource and interest indicators are very suitable for principal component analysis, while the power indicators are only moderately suitable for PCA[4].

2.2. PCA-based data transformation and dimensionality reduction

Based on the above, we can utilize Principal Component Analysis (PCA) to reduce the dimensionality of all indicators, facilitating our subsequent analysis. The detailed steps of Principal Component Analysis are as follows:

(1) Construct the sample matrix and standardize the samples within the matrix. All originate from secondary indicators related to power, resources, and interest. Standardized data is obtained by calculating the mean and variance.

Sample Matrix can be calculated as follows.

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}_{m\times n} \tag{1}$$

The mean and standard deviation can be calculated and expressed as follows.

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \tag{2}$$

$$S_j = \sqrt{\frac{\sum_{i=1}^{n}(x_{ij}-\mu_j)^2}{n-1}} \tag{3}$$

Obtain the standardized data can be expressed in the following form.

$$X_{ij} = \frac{x_{ij}-\mu_j}{S_j} \tag{4}$$

(2) Calculate the covariance matrix. This facilitates the subsequent acquisition of eigenvalues and eigenvectors. The covariance matrix can be expressed as follows.

$$R = \begin{bmatrix} r_{11} & r & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{bmatrix}_{m\times n} \tag{5}$$

$$r_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(X_{ki}-\mu_i)\left(X_{kj}-\mu_j\right) \tag{6}$$

(3) Calculate the variance contribution rate and cumulative variance contribution rate. Firstly, based on the covariance matrix in (2), eigenvalues and eigenvectors are obtained. Then, using the eigenvalues and eigenvectors, the variance contribution rate is determined. The cumulative

contribution rate can be obtained by summing up the variance contribution rates.The eigenvalues and eigenvectors can be calculated as follows.

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0 \tag{7}$$

$$a_1 = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{p1} \end{bmatrix}, a_2 = \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{p2} \end{bmatrix}, \cdots, a_p = \begin{bmatrix} a_{1p} \\ a_{2p} \\ \vdots \\ a_{pp} \end{bmatrix} \tag{8}$$

The contribution rates can be calculated as follows.

$$\sigma_j^2 = \frac{\lambda_i}{\sum_{k=1}^{\mathrm{n}} \lambda_{\mathrm{j}}} \tag{9}$$

Further, the cumulative contribution rate can be expressed as follows.

$$l_j = \frac{\sum_{k=1}^{i} \lambda_{\mathrm{k}}}{\sum_{k=1}^{\mathrm{n}} \lambda_k} \tag{10}$$

(4) Analyze the principal components.

Since the process of conducting principal component analysis on the four primary indicators is broadly similar, select the "Resource" indicator, which is most suitable for principal component analysis, to analyze the principal components.

Table 2. Variance contribution rate

Component	Eigenvalue	Eigenvalue Variance Contribution Rate(%)	Cumulative Variance Contribution Rate(%)
1	3.947	65.79	65.79
2	1.083	18.055	83.845

Based on the table above, by calculating the cumulative variance contribution rate, it is found that when the number of principal components reaches 2, the cumulative percentage of variance explained has reached 83.845, which is greater than 80%. Therefore, it is possible to represent the vast majority of the information of the six indicators with two principal components.

2.3. Analysis of Indicators

According to Table 3, the principal components can be identified as non-economic factors and economic factors. The specific analysis process is as follows:

The first principal component has significant positive loading on the second, third, fifth, and sixth indicators, and a significant negative loading on the fourth indicator, allowing the first principal component to be summarized as non-economic factor.

The second principal component has a significant positive loading on the first indicator and a moderate negative loading on the fifth indicator, with relatively small loading on the other indicators, allowing the second principal component to be summarized as economic factors.

Table 3. Factor loading coefficients

	Factor Loading Cofficient		Communality (Common Factor Variance)
	Non-economic Factors	Economic Factors	
External Dependency on Funds	0.309	0.918	0.938
Global Employee Count	0.964	0.032	0.931
Public Awareness	0.965	-0.046	0.934
Employee Compensation as a Percentage	-0.81	-0.189	0.691
Number of Data and Technology Platforms	0.791	-0.44	0.82
Proportion of Professional Teams to Total Staff	0.842	-0.088	0.716

3. Illegal Wildlife Trade Prediction Model

3.1. Multiple Linear Regression

Multiple Linear Regression (MLR) is a statistical method used to analyze the relationship between multiple independent variables and a dependent variable. In the multiple linear regression model, we assume that the dependent variable (or response variable) can be explained by a linear combination of a set of independent variables[5]. Each independent variable has an associated coefficient, which represents the degree to which that variable influences the dependent variable.

The advantages of multiple linear regression include its ability to simultaneously consider the effects of multiple independent variables on the dependent variable and provide a quantitative description of the complex relationship between the independent and dependent variables. When using the multiple linear regression model, it is important to verify the assumptions of the model for reasonableness[6].

3.2. Prediction Model based on MLR

Based on the Kendall consistency test, the P-value is obtained as 0.000, indicating that at the 0.1% significance level, the overall correlation coefficient between variables is significantly non-zero. The Kendall coefficient W=0.524, suggesting that the indicators of power, resources, and interest have good correlation. Therefore, a multiple regression model is established between these indicators and illegal wildlife trade.

Table 4.Kendall's coefficient of concordance

Name	Kendall's W Coefficient	Median	Rank Mean	X^2	P
interest	0.524	0.289	2.304	24.087	0
resource		0.539	2.522		
power		0.053	1.174		

The multiple regression model is as follows.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon_0 \quad (11)$$

Where β represents the coefficients, and epsilon represents the error term. By programming in Python, the results of the model fitting are obtained: the coefficient for power is -0.151, for resources is 0.516, for interest is 0.048, and the constant term is 0.09.

Separate univariate regression analyses were performed on the three indicators: power, resources, and interest, to predict their values over the next five years. These predicted values were then

incorporated back into the multiple regression model. The forecasted values for the next five years, assuming the project is not implemented, are shown in the figure below.

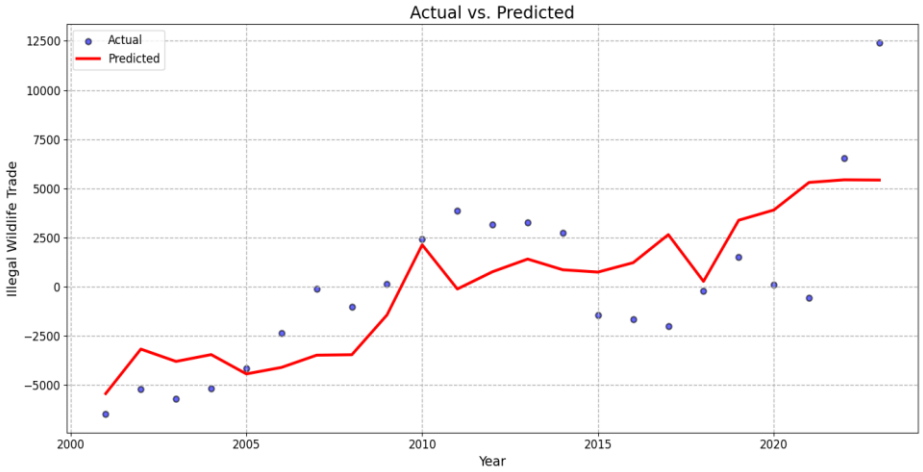

Fig. 1 Data Fitting Graph

Through the project, we can enhance the secondary indicators associated with power, resources, and interest, thereby improving the organization's power, resources, and interest. This, in turn, will reduce the data on illegal wildlife trade over the next five years. The specific results are illustrated in the figure below.

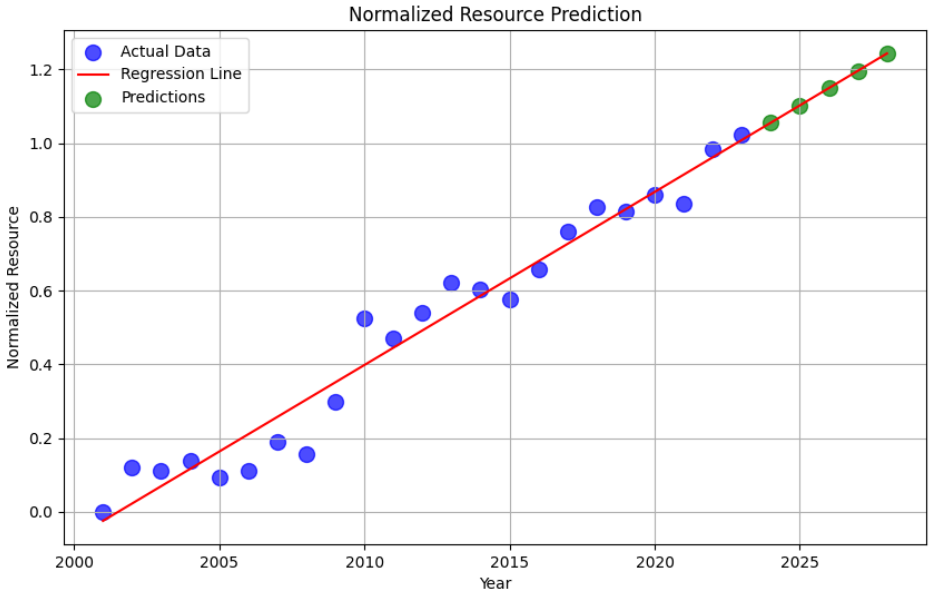

Fig. 2 Indicator Selection

3.3. Results of Model Prediction

Taking the resource indicator as an example, a five-year forecast of the illegal trade indicators is conducted in the absence of project intervention. Here, the illegal wildlife trade indicator is positively correlated with the trade; an increase in this indicator signifies a rise in the quantity of illegal wildlife trade. Combining the aforementioned principal component analysis, The specific results are illustrated in the figure below.

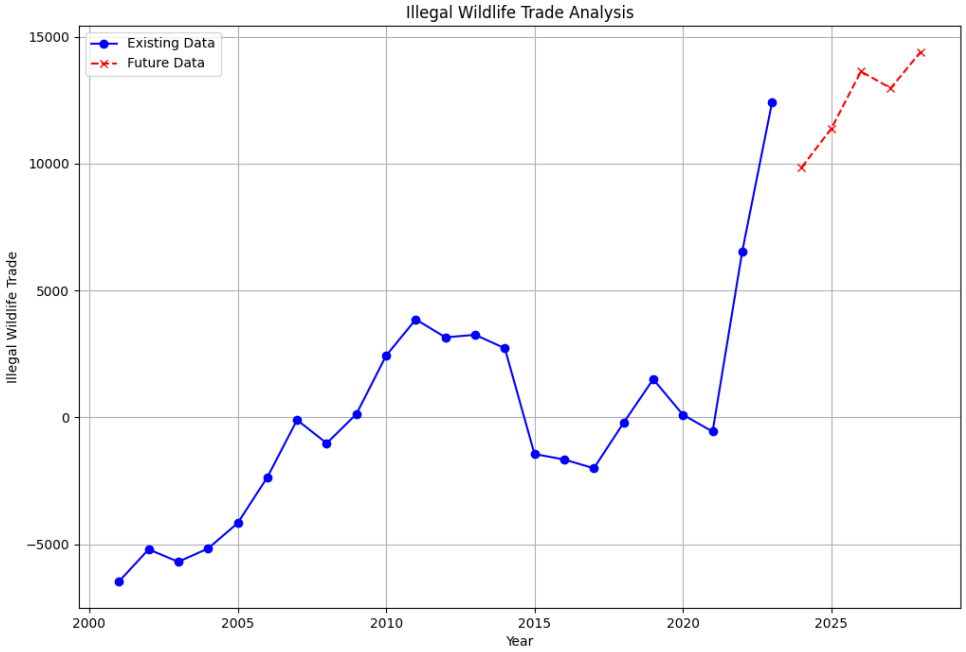

Fig. 3 Trade's predication without project

Continuing with the resource indicator as an example, the impact of project intervention is incorporated into the model, with a primary focus on its effect on the resource indicator. A five-year forecast of the illegal trade indicators under the condition of project intervention is conducted.The specific results are illustrated in the figure below:

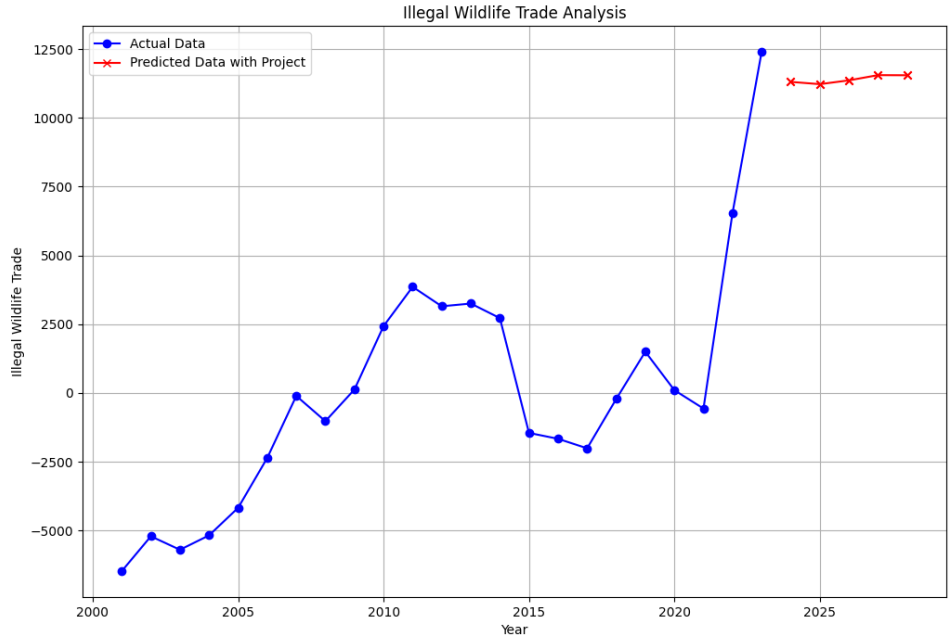

Fig. 4 Trade's predication with project

The data prior to and including 2023 are repetitive. To more intuitively reflect the impact of project implementation on illegal wildlife trade, indicators of illegal wildlife trade within the next five years are selected to construct a three-dimensional bar chart for comparison. The specific results are illustrated in the figure below.

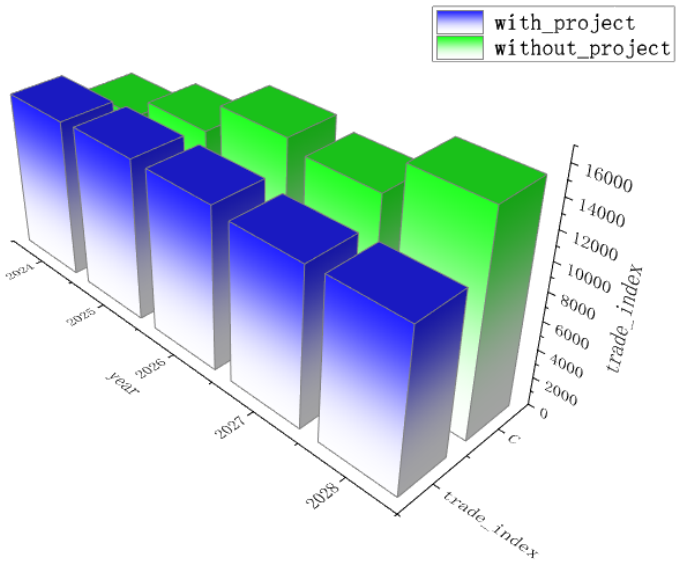

Fig. 5 With vs without project(3D)

3.4. Model Evaluation

3.4.1 Feasibility Analysis

Regarding the probability of project completion, a simple probability model is established for analysis. The illegal wildlife trade with and without project intervention, as mentioned in Task 2, is used as the indicator for analysis:

$$Y = P = \frac{2T_y}{T_y + T_n} \tag{12}$$

where T_n represents the volume of illegal wildlife trade without project intervention, and $T_y$represents the trade volume with project intervention. Taking the predicted value for 2028 as an example, the probability of project completion is approximately 60%.

3.4.2 Sensitivity Analysis

Python is used to generate random numbers between -15% and 15% (with three significant digits) and randomly disturb the primary indicators after dimensionality reduction by principal component analysis. The disturbance formula is as follows:

$$V_{new} = (1 + N_{random})V_{old} \tag{13}$$

Selected typical results are shown in the table below:

Table 5. Sensitivity analysis to random impact

Random Number	Random Indicator	Change Rate in Forecast Result
2.43%	resources	0.34%
7.75%	power	0.98%
14.60%	interest	1.22%

Notably, when the indicator labeled as "resource" underwent a 2.43% adjustment, there was a corresponding modest adjustment in the forecast outcome by 0.34%.

Further examination revealed a more pronounced but still proportionately restrained response when the "power" input was altered by 7.75%, leading to a forecast shift of 0.98%.

A critical observation was made with the "interest" input, which experienced a substantial 14.6% change. Surprisingly, this significant alteration resulted in a mere 1.22% variation in the predicted outcome.

In summary, minor variations in any of the variables will not significantly affect the model's prediction results, indicating that the model's predictions are relatively stable. However, if there are larger changes in the variables, the stability of the model's predictions may be affected.

4. Summary

This study adopts the PCA-MLR model for prediction, reducing a large number of indicators to a few for regression analysis, allowing for consideration of more indicators' impacts during the prediction process and enhancing the model's comprehensive analytical capability. The model not only considers illegal wildlife trade but also analyzes related ecosystem indicators, thus achieving effectiveness and accuracy in model prediction. It is of significant positive relevance in curbing illegal wildlife trade, creating suitable habitats for these species, thereby preserving biodiversity and enhancing ecosystem stability.

Certainly, the proposed model in this study heavily relies on the accuracy of data. Indicators of illegal wildlife trade are challenging to collect and prone to errors, which could affect the model's outcomes. In future research, we will ensure the accuracy of the data to the fullest extent possible.

References

[1] Warchol G L. The transnational illegal wildlife trade[M]//Transnational environmental crime. Routledge, 2017: 379-396.

[2] Lee W Y, Park M, Hyun C U. Detection of two Arctic birds in Greenland and an endangered bird in Korea using RGB and thermal cameras with an unmanned aerial vehicle (UAV)[J]. PLoS One, 2019, 14(9).

[3] Shaffer MJ, Bishop JA. Predicting and Preventing Elephant Poaching Incidents through Statistical Analysis, GIS-Based Risk Analysis, and Aerial Surveillance Flight Path Modeling. Tropical Conservation Science. 2016;9(1):525-548.

[4] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.

[5] Uyanık G K, Güler N. A study on multiple linear regression analysis[J]. Procedia-Social and Behavioral Sciences, 2013, 106: 234-240.

[6] Tranmer M, Elliot M. Multiple linear regression[J]. The Cathie Marsh Centre for Census and Survey Research (CCSR), 2008, 5(5): 1-5.