Ridge regression-based modeling for assessing the extent of response to biological invasions

Zhaokun Li *

School of Transportation, Beijing Jiaotong University, Beijing, China, 100044

* Corresponding Author Email: 23241032@bjtu.edu.cn

Abstract. This study aims to construct a ridge regression-based model for assessing the threat level of biological invasion security. By comprehensively collecting data from economic, medical, scientific research, social and ecological domains and combining correlations, key factors closely related to the threat of biological invasion security are filtered out. Compared with linear regression, ridge regression can effectively solve the problem of covariance in the data, prevent the model from overfitting, and improve the stability and generalization ability of the model through the regularization term. The innovation is that this paper is the first time to apply the ridge regression technique to the assessment of biological invasion security threat, which fills the research gap in this field. Pearson correlation was used in the study, and several factors such as nominal GDP, total population, average annual temperature, energy conservation and environmental protection expenditures, and scientific research output were screened as independent variables, and the quantitative relationship with the number of biological invasive species species and invasion was constructed by using the ridge regression model. Through the validation of the model and data testing, the results show that the model is able to accurately assess the situation of different regions in response to biological invasions, providing a scientific basis for the development of targeted prevention strategies. In addition, another innovation of this study is the comprehensive consideration of economic, ecological, scientific research and other factors from the degree to ensure that the model can accurately reflect the multiple influencing factors of biological invasions, which provides an important reference for policy making and risk management in related fields.

Keywords: Biological Invasion, Ridge Regression, Correlation, Assessment Modeling, Influencing Factors.

1. Introduction

Biological invasions, as a global problem, have become a major challenge for ecological conservation and agricultural production in many countries. With the high speed of global trade and transportation, the phenomenon of biological invasions has increased and caused serious negative impacts on ecology, agricultural production and socio-economics. Invasive species not only affect the local ecological balance, but also may lead to native species, crop yield reduction and pest outbreaks, resulting in huge economic losses. Therefore, accurately assessing the extent of a region's response to biological invasions and taking timely preventive and control measures have become an urgent problem in related fields.

Currently, research on biological invasion risk assessment mostly focuses on the identification of risk factors and the quantification of single factors, and lacks an in-depth exploration of the combined effects of multiple factors. Early studies mainly relied on qualitative assessment and judgment, and with the risk, quantitative models were gradually introduced. Li Zhihong's team (2018) systematized the quantitative assessment model of pest risk (PRA) in their review, pointing out that most of the existing quantitative assessments use linear regression to explore the relationship between invasion risk and variables such as climate adaptation and host distribution. Although linear regression modeling can provide predictive results to a certain extent, it tends to face the problem of covariance when dealing with high-dimensional data, leading to model instability and low prediction accuracy. Linear regression assumes that independent variables are independently and linearly correlated with each other, an assumption that is often difficult to hold in the real world, especially in the case of multivariate covariate coexistence, which makes the regression perform poorly in dealing with biological invasion security threats.

In recent years, to address the problem of data covariance, scholars have attempted to use regularization such as ridge regression. Ridge Regression (Ridge Regression) reduces the size of model parameters by introducing the L2 regularization term, which effectively mitigates the covariance problem and improves the stability and generalization ability of the model. Although ridge regression has achieved remarkable results in some fields, such as economics and science, the application in the field of bioinvasion security threat assessment is still relatively small, and most of the related literature lacks systematic modeling and quantification of bioinvasion prevention. Therefore, how to solve the covariance problem and construct accurate assessment models using ridge regression is an important gap in the current research in this field. This study aims to fill this research gap by applying ridge regression for the first time to assess the degree of biological invasion security threats.

In addition, most of the models in the existing literature focus only on a small number of single factors and lack a comprehensive consideration of degree factors. For example, some studies only look at climate change and species invasion from a single perspective, ignoring the combined effects of socioeconomic factors, ecology, and scientific research and innovation. In this context, exhaustive predictions provide valuable information, but these models often fail to fully reflect the complexity and diversity of biological invasions due to the neglect of multifactor interactions. Therefore, how to incorporate multiple influencing factors into the model to form a more accurate and comprehensive framework for assessing the risk of biological invasions is a pressing issue in the field.

Through this study, we have comprehensively collected data related to biological invasions, including data from economic, social, ecological and scientific research fields, and combined them with Pearson's correlation to screen out the key factors that are closely related to the threat of biological invasions. The innovation of this study is that a ridge regression model is used to solve the problem of covariance in regression, and a biological invasion risk assessment model that integrates multiple factors is constructed. By quantifying the influence of each factor on the degree of biological invasion, the model provides a scientific basis for the formulation of preventive strategies in different regions, and helps to improve the precision and effectiveness of biological invasion prevention and control.

2. Materials and Methods

2.1. Data acquisition and pre-processing

Data on specific factors of a country are selected for the last ten years. The data for this study come from public reports issued by scientific research organizations of statistical departments and field survey data (https://www.stats.gov.cn/) to ensure the comprehensiveness and reliability of the data. The data include many aspects related to biological invasions, economic data, ecological data, scientific research data, social data, etc. All the data are pre-processed and planned to ensure the stability and accuracy of the model.

2.2. Method Introduction

BP neural network is back propagating, mainly composed of three parts: input layer, middle layer and output layer. The number of nodes in the input and output layers is relatively easy to determine, but the determination of the number of nodes in the hidden layer is a very important and complex problem.

2.2.1. Description of symbols

Year: Year, Number of Invasive Species: Inv_Species,Invasion (Estimated): Inv_Freq, Nominal GDP (Billions of Yuan): GDP, Real GDP Growth Rate (%): GDP_Growth, Total Population (10,000s of people): Population, Annual Mean Temperature (℃):Avg_Temp, Annual Precipitation (mm).

Precipitation, EQI: EQI, Energy Saving and Environmental Protection Expenditures (Billion Yuan): Eco_Exp, Number of Papers Published in the Field of Biology: Bio_Papers, Number of Patents Filed in the Field of Biology: Bio_Patents.

2.2.2. Screening of influencing factors

Pearson correlation is used, Pearson Correlation Coefficient (Pearson Correlation Coefficient) is a statistical indicator used to measure the strength and direction of the linear relationship between two continuous variables. Its value is between -1 and 1: +1 means perfect positive correlation: one variable and the other variable are perfectly linear. 0 means no linear correlation: there is no significant linear relationship between the variables. -1 means perfect negative correlation: one variable, the other variable, and a perfect linear relationship [1]. Calculation formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \cdot \sqrt{\sum(y_i - \bar{y})^2}} \tag{1}$$

r: Pearson's correlation coefficient. x_i, y_i: Observed values of variables X and Y. $\bar{x}, \bar{y}$: Means of X and Y.

This formula essentially calculates the ratio of the covariance to the difference and measures the linear correlation of two variables by quantifying the covariance [2].

The advantages of Pearson correlation are simple and intuitive: the calculation formula is simple and easy to understand, and the results are intuitive and clear, which is suitable for quickly assessing the relationship between variables. Quantifying the strength of correlation: the result ranges from -1 to 1, which can clearly reflect the degree of positive, negative or no correlation, and it is convenient to compare the relationship between different variables. Widely applicable: applicable to most continuous type data, especially widely used in economics, ecology, biostatistics and other fields. High degree: Especially for linear relationships, it can effectively capture small change trends. Support bidirectional relationship: not only can the relationship between the dependent variable and the independent variable, but also applies to the peer-to-peer correlation between the variables, without having to define the causal direction in advance [3].

Using the correlation, the correlation between the many factors involved and the degree of threat of biological invasion security is calculated, and according to the size of the correlation coefficient, the key factors with strong correlation are screened out, and these key factors will be used as the input variables for the subsequent modeling, in order to ensure that the model can accurately respond to the main influencing factors of the threat of biological invasion security[4]. As shown in Figure 1:

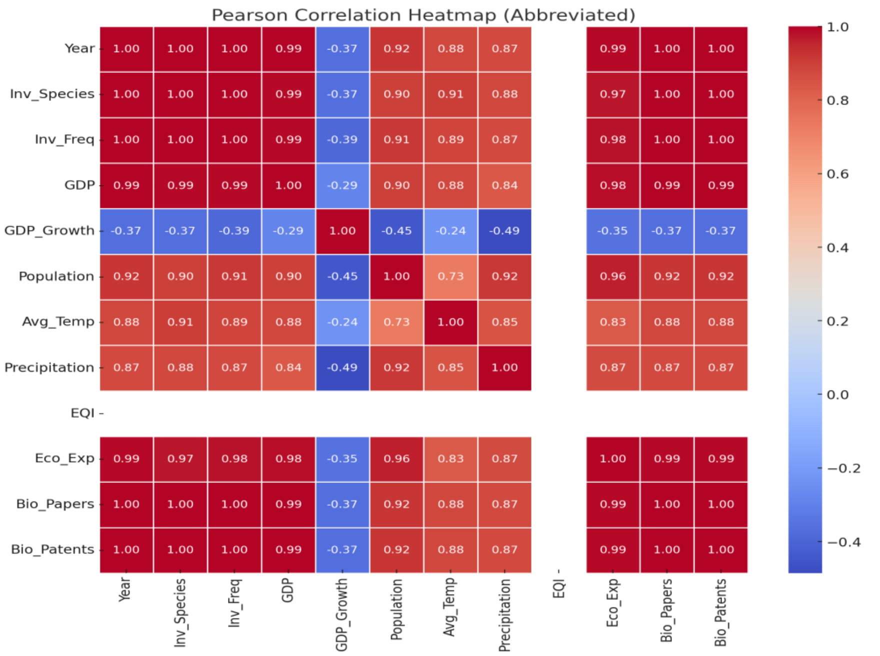

Figure 1. Correlation between variables

The variables and their correlation coefficients are shown separately for variables that are more strongly correlated with invasion (Inv_Freq) and the variables and their correlation coefficients that are more strongly correlated with the number of invasive species types (Inv_Species) [5]. As shown in Figure 2.

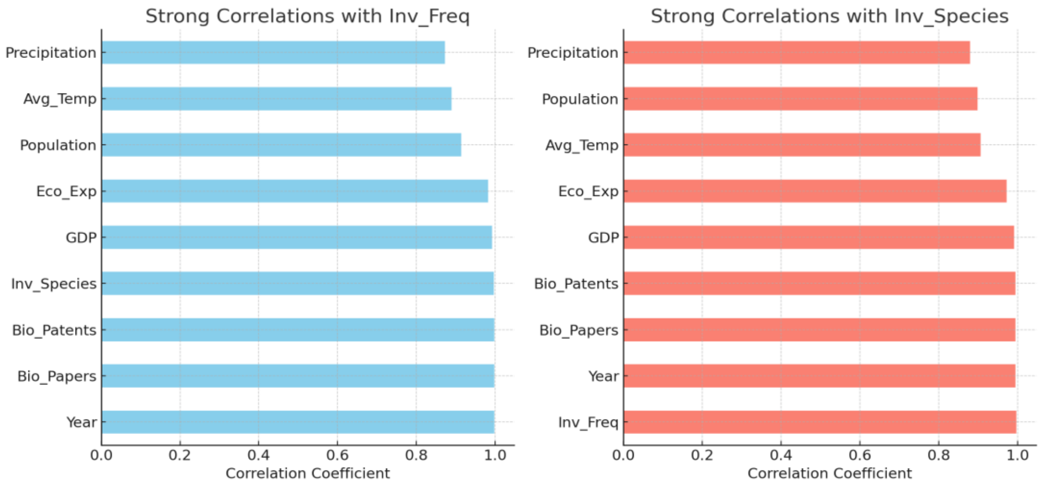

Figure 2. Stronger correlation factors

The trend of invasion (Inv_Freq) and number of invasive species types (Inv_Species) with respect to year is shown [6]. This is shown in Figure 3.

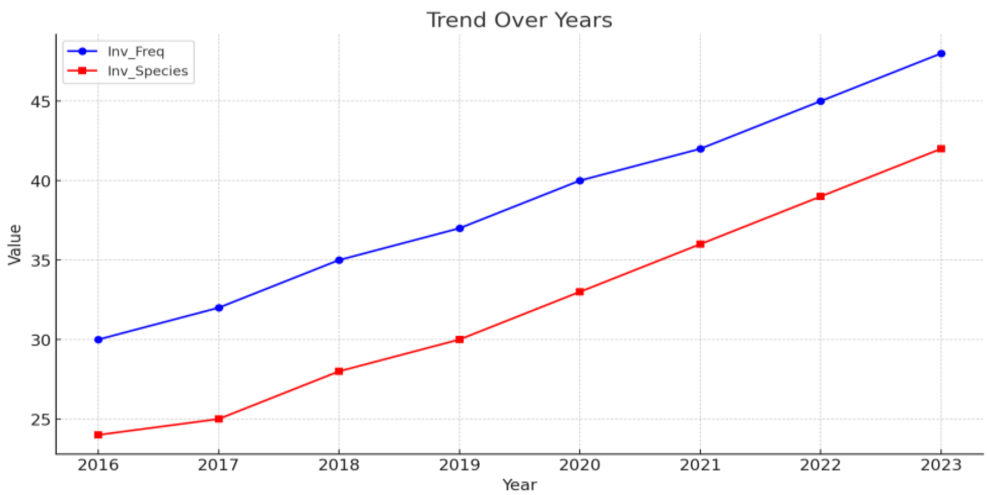

Figure 3. Relationship between biological invasions and year

In summary, the following factors were selected: Nominal GDP (billions of yuan): GDP, Total population (10,000 people): Population, Average annual temperature (°C): Avg_Temp, Expenditure on energy conservation and environmental protection (billions of yuan): Eco_Exp, Number of papers published in the field of Biology (one): 'Bio_Papers, Number of patents applied in the field of Biology (one): Bio_Patents as the independent variables. Patents factors were used as independent variables, species invasion and number of species as dependent variables, and assessment was modeled using ridge regression.

2.2.3. Construction of the model

Choosing the Ridge Regression model, Ridge Regression (RR) is a regularization for linear regression whose main purpose is to deal with the problem of multicollinearity and prevent overfitting by introducing a regularization term. It is based on Ordinary Least Squares regression (OLS) and has an L2 regularization term to reduce the size of the model parameters and to improve the generalization ability of the model by controlling the complexity of the model[7]. The goal of ridge regression is to minimize the following loss function.

$$J(\beta) = \sum_{i=1}^{n}(y_i - x_i \cdot \beta)^2 + \alpha \sum_{j=1}^{p} \beta_j^2 \quad (2)$$

y_i is the dependent variable (target variable) for the ith observation. x_i is the independent variable (characteristic) for the ith observation. β_j is the regression coefficient, indicating the weight of each independent variable. α is the regularization parameter (controlling for model complexity), larger means a stronger penalty for model coefficients. n is the number of samples and p is the number of features.

The regularization term $\alpha \sum_{j=1}^{p} \beta_j^2$ prevents overfitting by penalizing excessively large regression coefficients, especially in the presence of multicollinearity, and ridge regression performs more robustly than ordinary least squares regression [8].

The advantage of ridge regression is that it solves the problem of multicollinearity: in the presence of multicollinearity in the data (i.e., there is a strong linear relationship between the features), ordinary least squares (OLS) regression may lead to instability of the regression coefficients. Ridge regression shrinks the regression coefficients by introducing the L2 regularization term, thus the covariance problem affects the model. Preventing overfitting: ridge regression imposes constraints on the regression coefficients through the regularization parameter, thus avoiding model overfitting. In complex models, overfitting often results in models that perform well on training data but generalize poorly to new data. The regularization term helps to control the model complexity and improve the generalization ability. Stability: Due to the regularization term, ridge regression is more stable than ordinary least squares regression, especially in high-dimensional data (the number of features is much larger than the number of samples), ridge regression can provide more robust prediction results. Applicable to high-dimensional data: For high-dimensional data with a large number of features, ordinary least squares regression may not be able to calculate stable regression coefficients. By introducing regularization, ridge regression can effectively deal with high-dimensional data with interference between features. Solve the overfitting one: By choosing the regularization parameter, ridge regression can balance the fit and complexity of the model, helping us to build a model that performs well on both training and testing sets. Simple and easy to implement: ridge regression is relatively simple to implement and does not require complex algorithms. It requires only one regularization parameter, which can be chosen by the user through cross-validation and so on [9].

3. Modeling and solving

In order to assess the extent of biological invasions, we can construct a predictive model through Ridge Regression (Ridge Regression), which utilizes a series of factors affecting the extent of biological invasions to predict the level of risk or threat of biological invasions [10].

3.1. Modeling

3.1.1. Independent variables (characterizing variables)

These independent variables are factors that influence the level of biological invasions and can be selected as relevant variables based on the data you provide. For example: GDP: economic level of the region (may be related to the investment in biological invasion control). Population: population size (may affect the allocation of resources for control efforts). Avg_Temp: average annual temperature (climatic conditions affect the adaptability of biological invasions). Eco_Exp: energy saving and environmental protection expenditures (may reflect the ecological conservation of the region). Bio_ Papers: number of papers published in the field of biology (reflecting the extent of research related to biological invasions). Bio_Patents: number of patents filed in the field of biology (reflecting the innovation and application of biological invasion technologies).

3.1.2. Dependent variable (target variable)

The target variable is the extent of biological invasion, which may be represented by the following dependent variables: Inv_Freq: invasion (estimated). inv_Species: number of invasive species species.

3.2. Ridge regression model

The goal of ridge regression is to predict the dependent variable (extent of biological invasion), which is modeled with the following formula [11]:

$$y = \beta_0 + \sum_{i=1}^{p} \beta_i \cdot x_i + \alpha \sum_{i=1}^{p} \beta_i^2 \tag{3}$$

y is the target variable indicating the degree of biological invasion (which can be the number of invasive or invasive species types). β_0 is the intercept term of the model. β_i is the regression coefficient, which indicates the strength of the effect of each independent variable on the dependent variable (e.g., β_1 is the regression coefficient for GDP). x_i is the ith independent variable (e.g., GDP, population size, temperature, etc.). α is the regularization parameter, which controls the complexity of the model. p is the number of characteristic variables (in this model there are 6 characteristic variables) [12].

3.3. Evaluation indicators

Based on the predicted number of invasive species species (Inv_Species), we can assess the level of biological invasion based on the magnitude of the predicted value: High: If the predicted value is greater than a certain threshold (e.g. 80), it means that the area has a high level of preparedness. Medium: If the predicted value is within a certain range (e.g., 40 to 80), the area has a medium level of preparedness. Low: If the predicted value is less than a certain threshold (e.g., 40), it indicates that the area is deficient and at high risk.

3.4. Model Validation

Plot a line graph of the number of invasive species types (Inv_Species) over years based on the predictions from the ridge regression model. It shows how the predicted values change over time, helping us to observe trends in the level of biological invasions. This is shown in Figure 4.

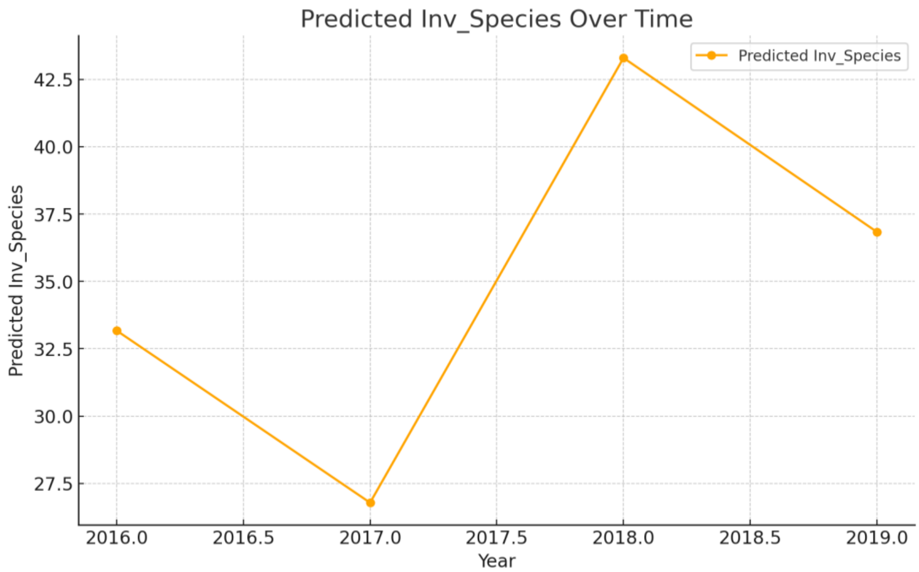

Figure 4. Predicted values derived from the model

The following shows the actual values versus the predicted values for each year. As well as the trend of the actual value and the predicted value with the year, it is easy to observe the relationship between their changes. As shown in Figure 5, Figure 6.

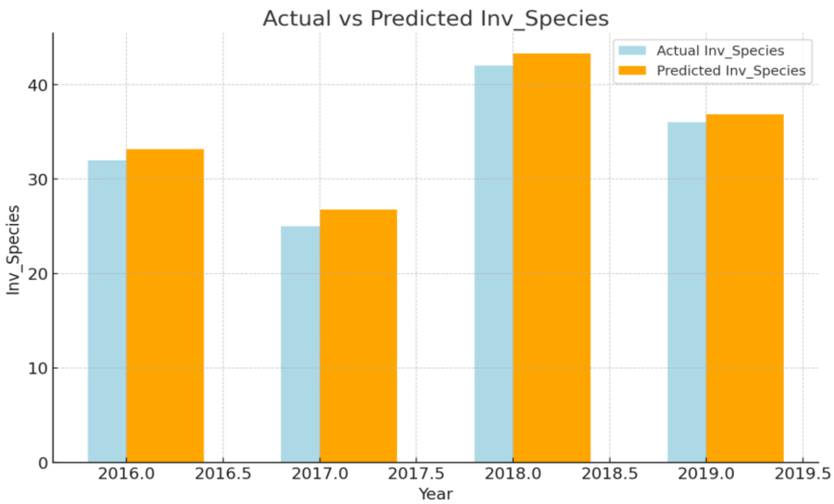

Figure 5. Comparison of predicted and true values

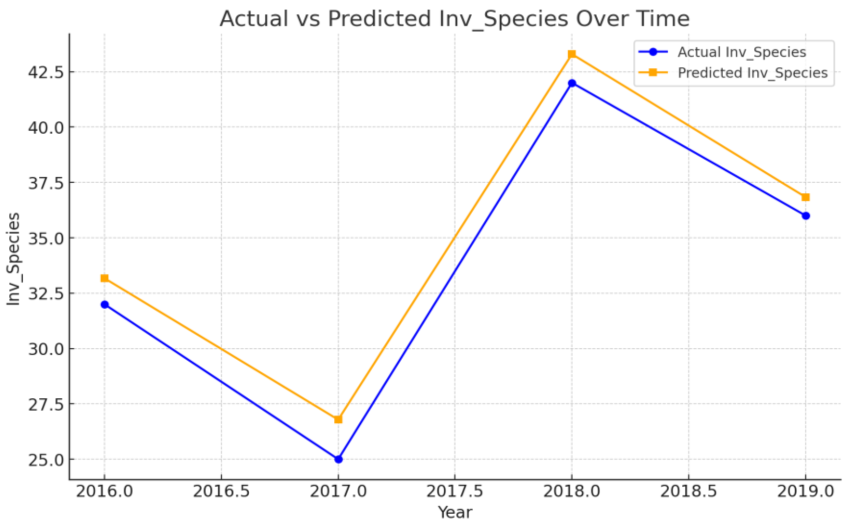

Figure 6. Trend comparison of projected and actual values

4. Conclusion

The aim of this study is to construct an assessment model based on ridge regression to cope with the threat level of biological invasion security. It comprehensively collects data from economic, medical, scientific research, social, ecological and other fields, filters out key factors closely related to the threat of biological invasions through correlation, establishes the relationship equation between the key factors and the number of invasive species species, and finally, evaluates the degree of biological invasions based on the predicted number of invasive species species. It provides a scientific basis for countries to formulate targeted preventive strategies. In the future, more data, more countries, and more data sources will be introduced, and more influencing factors will be taken into account to improve the comprehensiveness and robustness of the model.

References

[1] Maliwan T, Do T T Q, Nguyen M C, et al. Exploring the co-occurrence of microplastics, DOM and DBPs inside PVC pipes undergoing chlorination by correlation analysis and unsupervised learning [J]. Chemosphere, 2025, 373144171-144171.

[2] Rahman H, Anggadiredja K, Sasongko L. Mechanisms of oral ciprofloxacin-induced depressive-like behavior and the potential benefit of lactulose: A correlation analysis [J]. Toxicology Reports, 2025, 14101920-101920.

[3] Liu D, Ma X, Li X, et al. Correlation study between the microstructural abnormalities of medial prefrontal cortex and white matter hyperintensities with mild cognitive impairment patients: A diffusion kurtosis imaging study [J]. Psychiatry Research: Neuroimaging, 2025, 348111958-111958.

[4] Ren Y, Qi F, Li J, et al. Expression and correlation analysis of VISTA in peripheral blood mononuclear cells of myasthenia gravis patients [J]. International Immunopharmacology, 2025, 148114096-114096.

[5] Zhang S, Ma Y, Ren X, et al. Combined Lactobacillus plantarum fermentation and heat-moisture treatment: Correlation analysis of physicochemical properties of rice flour and quality of rice noodles [J]. International Journal of Gastronomy and Food Science, 2025, 39101120-101120.

[6] Li J, Cui X, Yang J. Multi-scale correlation analysis between geometric parameters and solar radiation in high density urban environment——Case study in Nanjing [J]. Frontiers of Architectural Research, 2025, 14(1): 248-266.

[7] Dai D, Javed F, Karlsson P, et al. Nonlinear forecasting with many predictors using mixed data sampling kernel ridge regression models [J]. Annals of Operations Research, 2025, (prepublish): 1-20.

[8] Amin M, Amanullah M, Aslam M, et al. Influence diagnostics in gamma ridge regression model [J]. Journal of Statistical Computation and Simulation, 2019, 89 (3): 536-556.

[9] Aamir S, Muhammad A, Walid E, et al. New ridge parameter estimators for the quasi-Poisson ridge regression model [J]. Scientific Reports, 2024, 14(1): 8489-8489.

[10] Muhammad S, Sohail C, Golam M B K. Quantile based estimation of biasing parameters in ridge regression model [J]. Communications in Statistics - Simulation and Computation, 2019, 1-13.

[11] Renlong H, Qingshan L, Guiyu X, et al. Correcting MODIS aerosol optical depth products using a ridge regression model [J]. International Journal of Remote Sensing, 2018, 39 (10): 3275-3286.

[12] Yuting L, Zhiyao S, Ruijie L, et al. A hybrid mechanism and ridge regression model to separate the effects of advection and resuspension on suspended sediment concentration [J]. Ecological Indicators, 2023, 156.