An Exploration of Tourism Development Based on Genetic Algorithm and Multiple Linear Regression Modeling

Yang Sun*

School of Mathematics and Computer Science, Northwest Minzu University, Lanzhou, China

*Corresponding author: weiki886@163.com

Abstract. This paper focuses on tourism development based on genetic algorithm and multiple linear regression model. Firstly, the data are preprocessed, including identifying outliers and processing them through QR algorithm, as well as standardizing and normalizing the data for transformation. Second, polynomial regression models were constructed for tourism development analysis, defining target variables and input features, and ridge regression models were built for prediction through polynomial feature generation, etc., and model effects were evaluated through residual and error analysis; meanwhile, multiple linear regression models were constructed to analyze glacier recession, defining relevant variables and evaluating parameters. Finally, in order to achieve the comprehensive optimization objectives of maximizing tourism revenue and minimizing glacier recession, genetic algorithm is used to optimize it, and through a series of genetic operations, synergistic optimization of multiple parameters is achieved to effectively deal with the trade-off situation between tourism development and glacier recession.

Keywords: Polynomial regression models; ridge regression; multiple linear regression models; genetic algorithms.

1. Introduction

In the field of tourism development, previous studies of the same type have mostly used a single model or simple data analysis methods. These methods are often difficult to comprehensively capture the complex variable relationships and have limitations in dealing with the actual situation of multiple factors interacting with each other and are unable to realize the comprehensive optimization of tourism economy and ecological environment[1].

In order to break through the above limitations, this paper comprehensively utilizes genetic algorithm[2] and multivariate linear regression model[3] to carry out in-depth research. On the one hand, the polynomial regression model is used to process the data related to tourism development[4], covering multiple links, and accurately analyze the relationship between tourism income and various influencing factors[5]; on the other hand, the integrated objective function of maximizing tourism income and minimizing glacier retreat is optimized by genetic algorithm to analyze the multi-parameter synergistic optimization. The research in this paper can fully consider the complex relationship between multiple factors in the two key situations of tourism development and glacier recession, realize the comprehensive optimization of multiple objectives, and provide a more scientific basis for tourism development planning and environmental protection decision-making[6].

2. Data Preprocessing

2.1. Data Cleaning

2.1.1 Missing value processing

In the actual data collection process, some features (such as temperature, number of tourists, precipitation, etc.) may have missing values. The existence of missing values will affect the training and prediction effect of the model, so it needs to be processed. For missing values, the mean filling method is used in this model.

Assume that the data set contains n samples, and the variables X have missing values. The study first X express the observed values of as vector form:

$$X = [X_1, X_2, \ldots\ldots, X_n]^r \tag{1}$$

Some X_i values may be missing, and the subscript set of non-missing values is defined as:

$$\Omega = \{i \mid X_i \text{ Not-missing}\} \tag{2}$$

Next, calculate X the observed mean of the variable μ and replace the missing X_i values with the observed mean μ :

$$\mu = \frac{1}{|\Omega|}\sum_{i\in\Omega} X_i \tag{3}$$

Where μ is the observed mean, $\sum_{i\in\Omega}X_i$ is the sum of the non-missing values in the column, Ω and is the number of non-missing values.

2.1.2 QR outlier identification

First, implement data sorting and quantile definition:

Assume that the data set is $X = x_1, x_2, \ldots, x_n$, and then arrange it in ascending order to get the ordered sequence: $x_{(l)} \le x_{(2)} \le \cdots \le x_{(n)}$.

Then implement the quantile definition:

First quartile (Q_1): 25% of the observations in the data are less than or equal to it.

$$Q_1 = \text{Quantile - function}\,(X, p = 0.25) \tag{4}$$

Third quartile (Q_3): 75% of the observations in the data are less than or equal to it.

$$Q_3 = \text{Quantile-function}\,(X, p = 0.75) \tag{5}$$

Next, calculate the interquartile range (IQR):

$$IQR = Q_3 - Q_1 \tag{6}$$

IQR Describes the dispersion of the middle 50% of the data.

The outlier boundary formula is given as:

Nether:

$$\text{Lower Bound} = Q1 - k \times IQR \tag{7}$$

Upper bound:

$$\text{Upper Bound} = Q3 + k \times IQR \tag{8}$$

Where is the adjustment coefficient ($kk = 1.5$ in this model).

The outlier determination condition is $x_i <$ Lower Bound or $x_i >$ Upper Bound. After determining it as an outlier, it will be filled as a missing value.

2.2. Data Conversion

Since different features have different dimensions, standardization or normalization can effectively prevent certain features from having too much influence on model training, especially for distance-based algorithms (such as regression analysis). The formula is as follows:

$$x' = \frac{X-\mu}{\sigma} \tag{9}$$

Among them: X' is the standardized data, X is the original data, μ is the mean, and σ is the standard deviation.

3. Tourism Development Analysis

3.1. Target Variable and Input Feature Definition

The target variable is tourism revenue y, which is defined by the product of the number of tourists T and their spending S:

$$y = T \times S \tag{10}$$

The model directly reflects the scale of tourism economy through tourism revenue, which is the core objective of the model's prediction.

The input features x includes the following:

$$x = \left[T_{\log}, S_{\log}, \theta, p, c, e, r\right] \tag{11}$$

Where: $T_{\log} = \ln(1 + T)$ represents the logarithmic transformation of the number of tourists (alleviating dimensional differences and heteroscedasticity), $S_{\log} = \ln(1 + S)$ represents the logarithmic transformation of tourist spending, θ represents the average temperature $(^{\circ}C)$, p represents the precipitation (mm), c represents the cloud cover $(\%)$, e represents the potential evapotranspiration (mm/day), and r represents the poverty rate (unit: number of poor people/total population).

3.2. Data Standardization and Normalization

Data standardization has been described in the data preprocessing section. The characteristics after standardization are as x_{scalke} follows. Here the study focuses on describing the normalization of the target variable.

In this model, the target variable needs to y be scaled to the interval $[0,1]$:

$$y_{norm} = \frac{y - y_{\min}}{y_{\max} - y_{\min}} \tag{12}$$

The role of normalization is to accelerate model convergence and avoid gradient explosion.

3.3. Polynomial Feature Generation

Then, the standardized features x_{saaled} are expanded by cubic polynomial to generate nonlinear features:

$$x_{poly} = \left[1, x_i, x_i x_j, x_i^2, x_i^2 x_j, x_i^3\right](i, j \in \{1,2, \ldots \ldots, 7\}, i < j) \tag{13}$$

Polynomial expansion was performed to capture interactions and nonlinear relationships between features (e.g., the joint effects of temperature and precipitation).

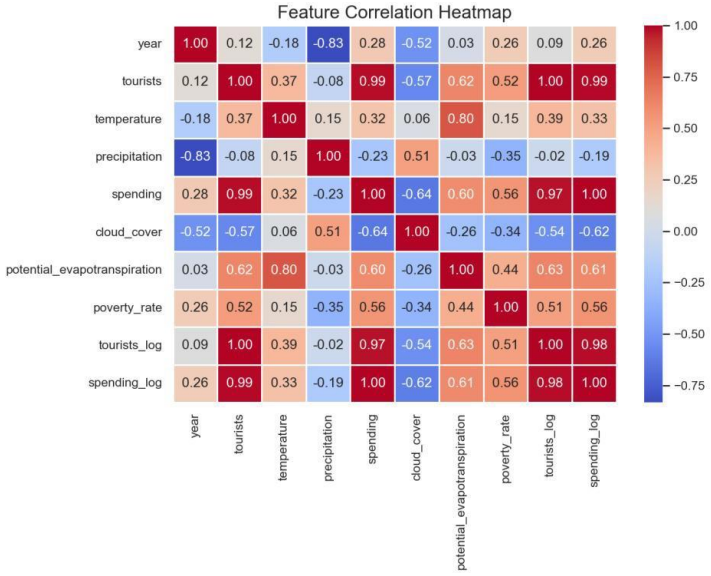

Fig. 1 Feature correlation heat map

This is a feature correlation heat map. The horizontal and vertical axes of this figurerepresent the feature names, which respectively represent the year, the number of tourists in Juneau, the temperature, etc. The depth of the color represents the strengthof the correlation. As can be seen from the above Fig. 1, the closer the color is to red, thestronger the correlation is. The closer the color is to blue, the weaker the correlationis

3.4. Establishing Ridge Regression Model

Ridge regression minimizes the loss function with L2 regularization:

$$\min_{w} \| y_{\text{norm}} - X_{\text{poly}} w \|^2 + \alpha \| w \|^2 \tag{14}$$

Where: w is the regression coefficient vector, dimension is the polynomial feature number. X_{poly} is the polynomial feature matrix. α is the regularization strength, used to control the model complexity.

The optimal value is selected through 5-fold cross validation (GridSearchCV) α^* to minimize the prediction error.

3.5. Model Prediction and Inverse Transformation

Predict the normalized target variable:

$$\hat{y}_{\text{norm}} = X_{\text{ploy}} w^* + b \tag{15}$$

Among them: b is the intercept term.

Inverse normalization:

$$\hat{y} = \hat{y}_{\text{norm}} \cdot (y_{\max} - y_{\min}) + y_{\min} \tag{16}$$

The goal is to restore the predicted values to their original scale.

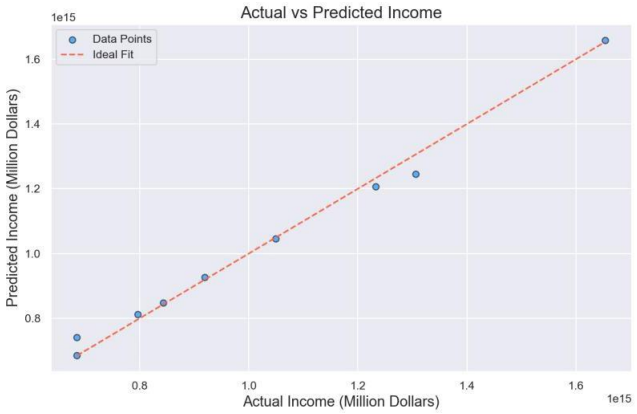

Fig. 2 Actual income VS predicted income

The above is the actual income VS predicted income chart. This scatter plot showsthe relationship between the actual income and the income predicted by the Ridgeregression model. Each point represents a year, the horizontal axis is the actual income, and the vertical axis is the predicted income. All the points in the figure are roughlydistributed along the diagonal line (ideal fitting line), that is, the predicted incomeisalmost exactly the same as the actual income. As can be seen from the above Fig. 2, the model prediction effect is very good.

3.6. Residual and Error Analysis

Residuals r (analyze the difference between actual and predicted values, used to assesslocal deviations of the model):

$$r = y - y \tag{17}$$

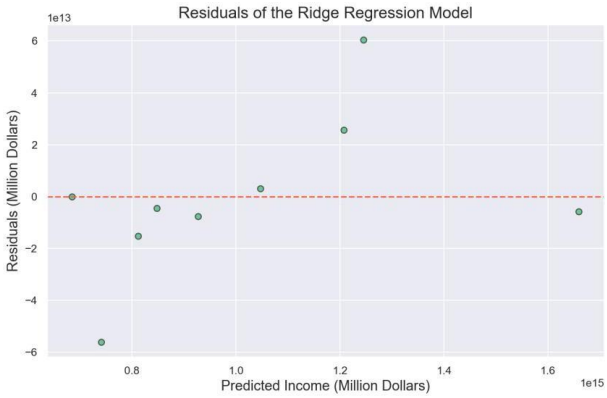

Fig. 3 Residual between the predicted revenue and the actual revenue

Above Fig. 3 shows the residual (i.e., forecast error) between the predicted revenue and the actual revenue. The horizontal axis is the predicted revenue, and the vertical axis is the residual (the difference between the actual revenue and the predicted revenue). A horizontal line $(y = 0)$ is also drawn in the figure, which indicates that the forecast error is zero in the ideal case. This figure is used to check the fit of the model.

Average relative error (MAE_{rel}):

$$MAE_{\text{rel}} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{18}$$

The average prediction error of the model is 2.09%.

3.7. Feature Correlation Analysis

Analyzed by Pearson correlation coefficient:

$$p_{ij} = \frac{\mathrm{Cov}(x_i, x_j)}{\sigma_{x_i}\sigma_{x_j}} \tag{19}$$

4. Glacier Retreat Analysis

4.1. Symbol Definition and Target Variable

Target variable G : represents the glacier area loss in square kilometers (km^2)
Input features u : Includes the following six environmental and human factors:

$$u = [u_1, u_2, u_3, u_4, u_5, u_6] \tag{20}$$

Where: u_1 is the average annual temperature ($^\circ C$), u_2 is the annual number of tourists (person-times), u_3 is the carbon dioxide emissions (unit: kilotonnes), u_4 is the methane emissions (unit: kilotonnes), u_5 is the total greenhouse gas emissions (unit: kilotonnes), u_6 and is the nitrous oxide emissions (unit: kilotonnes).

4.2. Building a Multiple Linear Regression Mode

The linear relationship between glacier retreat and various factors can be expressed as:

$$G = \beta_0 + \beta_1 u_1 + \beta_2 u_2 + \beta_3 u_3 + \beta_4 u_4 + \beta_5 u_5 + \beta_6 u_6 + \varepsilon \tag{21}$$

Where: β_0 is the intercept term, representing the baseline glacier loss. $\beta_1, \ldots, \beta_6$ is the regression coefficient, quantifying the contribution of each factor to glacier loss. ε is the random error term, which follows a normal distribution with a mean of 0 $\varepsilon \sim N(0, \sigma^2)$.

4.3. Parameter Estimation and Model Training

Estimate the regression parameters using the least square method:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} \left(G_j - \left(\beta_0 + \sum_{j=1}^{6} \beta_j u_{i,j}\right)\right)^2 \tag{22}$$

Among them: N is the number of observed samples, β is the optimal coefficient vector.

4.4. Model Evaluation Metrics

Mean Square Error (MSE)

$$MSE = \frac{1}{N}\sum_{i=1}^{N} \left(G_i - \hat{G}_i\right)^2 \tag{23}$$

Coefficient of determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \left(G_i - \hat{G}_i\right)^2}{\sum_{i=1}^{N} \left(G_i - \hat{G}\right)^2} \tag{24}$$

Where: $\bar{G}$ is the mean of glacier loss, reflecting the proportion of variance explained by the model.

4.5. Visualization Analysis

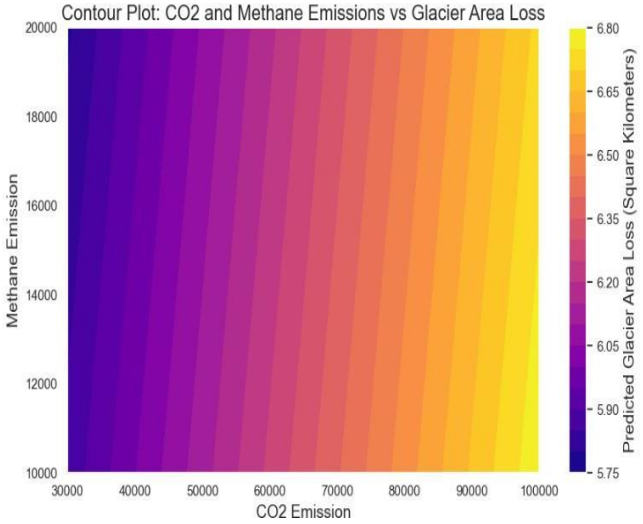

Fig. 4 Contour map

The horizontal axis of this Fig. 4 represents the range of carbon dioxide emissions, andthe vertical axis represents the range of methane emissions. The contour lines represent the glacier ablation values predicted by the model under different carbon dioxide andmethane emissions. The color map represents the range of glacier ablation values usingplasma color mapping. The color bar represents the specific numerical range of glacier ablation values. This figure analyzes the joint impact of carbon dioxide and methaneemissions on glacier ablation. The density and color changes of the contour linesintuitively show the impact of these two features on the target variable.

5. Analysis of Synergistic Optimization Between Glacier Retreat and Tourism

5.1. Symbolic Definition and Objective Function

Objective function H:

Comprehensive optimization objectives, maximizing tourism revenue R and minimizing glacier retreat penalties P:

$H = -R + P$ (Minimizing H is equivalent to maximizing R and minimizing P).

Input variables v (including the following 11 parameters):

$$v = \left[N, S, \theta, \rho, \kappa, \eta, \phi, \varepsilon_{\mathrm{CO_2}} \varepsilon_{\mathrm{CH_4}}, \varepsilon_{GHG}, \varepsilon_{\mathrm{N_2O}}\right] \tag{25}$$

Where: N is the number of tourists (person-times), S is the tourists' spending (US$), θ is the average annual temperature $(^\circ\mathrm{C}), \rho$ is the precipitation (mm), κ is the cloud cover (%), is the η potential evapotranspiration (mm/day), ϕ is the poverty rate (per 1,000 people), $\varepsilon_{\mathrm{CO_2}}$ is the carbon dioxide emissions (thousand tons), $\varepsilon_{\mathrm{CH_4}}$ is the methane emissions (thousand tons), ε_{GHG} is the total greenhouse gas emissions (thousand tons), ε_{N_2O} and is the nitrous oxide emissions (thousand tons).

5.2. Tourism Income Model

Basic income:

$$R_0 = N \cdot S \tag{26}$$

Environmental factor adjustments:

$$\Delta R = -0.01(\theta - 5.0)N - 0.005(\rho - 110.0)N + 0.002(\kappa - 75.0)N - 0.001(\eta - 1.3)N - 0.0001\phi N \tag{27}$$

Total Income:

$$R = R_0 + \Delta R \tag{28}$$

5.3. Glacier Retreat Penalty Model

Predicted glacier retreat ($\hat{G}$ based on a multivariate linear regression model withcoefficients of coefficients of γ_1):

$$\hat{G} = \gamma 0 + \gamma_1 \theta + \gamma_2 N + \gamma_3 \varepsilon_{CO_2} + \gamma_4 \varepsilon_{CH_4} + \gamma_5 \varepsilon_{GHG} + \gamma_6 \varepsilon_{N_2O} \tag{29}$$

Penalty items:

$$P = \max(0, \hat{\mathrm{G}} - 4.0) \tag{30}$$

When the predicted glacier retreat $\hat{G}$ exceeds a threshold of 4.0 km^2, a linear penalty is applied.

5.4. Genetic Algorithm Optimization Framework

Chromosome code:
Each individual is represented as $v \in R^{11}$ containing all optimization variables.
Fitness function:

$$\text{Fitness}\ (v) = H = -R(v) + P(v) \tag{31}$$

Genetic Operations:
Selection: retain 50% of individuals with the best fitness.
Crossover: Single-point crossover generates offspring.
Mutation: Add uniform perturbations to random genes with probability 0.1:

$$v_i \leftarrow v_i + \delta, \delta \sim u(-0{,}05, 0.05) \tag{32}$$

5.5. Error Analysis

The fitness change curve shows the fitness value change of the genetic algorithmineach generation. From the Fig. 5, the study can see that the rapid convergence: in the first fewgenerations, the fitness value drops rapidly, indicating that the algorithmquickly finds agood solution. Stabilization: as the number of iterations increases, the fitness valuegradually stabilizes, indicating that the algorithm is close to the optimal solution. Local optimum: in the later stage, the change in fitness value is very small, indicating that thealgorithm may fall into a local optimum, but still finds a more reasonable solution.

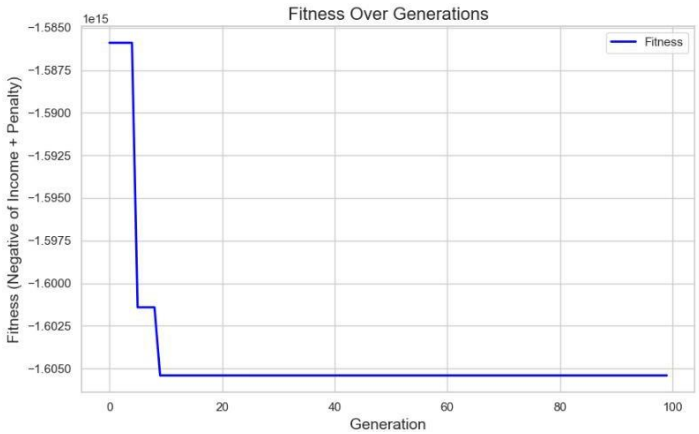

Fig. 5 Fitness change curve

6. Conclusion

In this paper, the study has conducted an in-depth investigation on tourism development and glacier recession by using genetic algorithm and multiple linear regression model. In terms of data processing, the mean padding method and QR algorithm are used to deal with missing values and outliers, and data standardization and normalization are completed to effectively improve data quality. In terms of model construction, polynomial regression model was constructed to analyze tourism development, and multiple linear regression model was constructed to study glacier recession, and the models showed good prediction and explanation ability after evaluation from various aspects. Meanwhile, with the help of genetic algorithm, tourism income and glacier recession are co-optimized to cope with the multi-objective optimization situation to a certain extent. Compared with the traditional single model or simple analysis method, this study integrates multiple algorithms, which can reveal the complex relationship between the factors more comprehensively and accurately.

References

[1] LIU Zhizhi, WANG Hai, TANG Leilei. Tourism development, tourism ecological compensation efficiency and ecological civilization construction[J]. Statistics and Decision Making,2024,40(22): 87-91.DOI: 10.13546/j.cnki.tjyjc.2024.22.015.

[2] Xu Xiangrong. Research on travel recommendation algorithm based on improved genetic algorithm with multi-source heterogeneous data[D]. Xi'an University of Technology, 2023.DOI: 10.27398/ d. cnki. gxalu.2023.001863.

[3] Xie Yanlin, Zhou Yan, Liang Jiajia. Empirical analysis of the main influencing factors of domestic tourism income - based on multiple linear regression model[J]. Journal of Beijing Institute of Printing,2020,28(01): 64-66.DOI: 10.19461/j.cnki.1004-8626.2020.01.021.

[4] YANG Chen, LIAN Kaicheng, XU Hao, et al. A weighted polynomial regression color characterization algorithm with residual correction[J]. Computer Application Research,2024,41(10): 3188-3193.DOI: 10.19734/j.issn.1001-3695.2023.11.0597.

[5] Hu Jiayu, Huang Zhenna, Ni Yuchen. Research on the influencing factors and development dynamics of domestic tourism income under the perspective of econometric analysis[J]. Industrial Innovation Research,2024, (09):8-10.

[6] Zhang Feng. Application of green development concept in the development of tourism management in urban scenic spots[J]. Engineering Seismic Resistance and Reinforcement Retrofitting,2025,47(01):199.