Research on the Integrated Impact Assessment Model of Extreme Weather in Agro-tourism Based on ARIMA and Entropy Weight Approach

Xupei He [1, *], Lige Deng[2, #], Ye Deng[2, #]

[1]School of Electronics and Communication Engineering, Guangzhou University, Guangzhou, China, 510006

[2]School of Mathematics and Information Sciences, Guangzhou University, Guangzhou, China, 510006

*Corresponding author: hxp358358@163.com

#These authors contributed equally.

Abstract. Agro-tourism has been developing rapidly in China and globally, but agro-tourism development is deeply affected by the frequency of extreme weather in the region and other constraints on tourism development. Based on the above viewpoints, this paper synthesizes the frequency of extreme weather and the key factors of regional tourism development, and discusses the development of agritourism layout in mainland China. Firstly, based on ARIMA time series model to construct the risk assessment system of agro-tourism industry affected by extreme weather, to predict the boundary value of the risk of damage to agro-tourism industry caused by natural disasters, to provide reference for the development of agro-tourism industry. Secondly, the Spearman-entropy weighting method model is used to determine the weight of each influencing factor by combining the six key factors restricting the development of agro-tourism industry. Finally, the Gaussian mixed clustering model combines the extreme weather risk assessment index and the weights of each influencing factor to derive the results of the assessment of the development of agro-tourism in the Chinese region, which further determines the feasibility of the development of the region, and provides inspiration and basis for the development of agro-tourism.

Keywords: Agro-tourism, Extreme weather, ARIMA, Entropy Weighting, Gaussian Mixture Clustering Model.

1. Introduction

In recent years, as countries have increasingly focused on the transformation and upgrading of rural economies, Maria-Irina Ana's research on European agro-tourism has revealed that this emerging sector has emerged as a promising avenue for sustainable development, offering a novel approach to village development worldwide[1]. The question of agro-tourism positioning and development has consequently emerged as a pivotal area of inquiry within the academic community.

Researchers from different countries have provided insights into the development of agrotourism. Absalon proposed a methodological framework combining Fuzzy Delphi, Fuzzy Best Worst Method (BWM) and Fuzzy Simple Weighted Approach (SAW) to create a comprehensive sustainability index for agro-tourism zones [2]. Puška used the FUCOM method to identify the importance of evaluation criteria and the Marginal Compromise Solution (MARCOS) to measure and rank rural tourism potential in different areas [3]. Wang emphasized the dependence of integrated agriculture and tourism on natural resources and demonstrated its impact on the efficiency of agrotourism through a PSTR model [4]. Ayu Andayani used an adaptive neuro-fuzzy inference system (ANFIS) to develop a predictive model of local government and community support for agritourism to aid decision making [5]. Zhang studied the spatial differentiation of rural tourism in China using methods such as the nearest neighbors index and kernel density analysis, and explored various influencing factors [6]. Liao analyzed the influence of spatial resource heterogeneity on tourism resources from a geological perspective through various analytical methods [7]. In addition, Joshi and Sharma used AHP-TOPSIS modelling to identify nine critical success factors (CSFs) for sustainable agritourism cluster development [8].

Previous studies have concentrated on the development of location-based agro-tourism, with a particular focus on the influence of natural resources, government support, land conditions and other pertinent factors. However, the development of this sector is subject to the impact of extreme weather conditions. Accordingly, this paper employs the ARIMA model to forecast the potential for extreme weather-related losses in China. It also utilizes the Spearman-entropy weighting method to determine the relative importance of six key tourism development factors and employs a Gaussian mixed clustering model to assess the feasibility of agro-tourism.

Data for this article comes from: https://data.stats.gov.cn/.

2. Theory and Method

2.1. ARIMA model

The Autoregressive Integrated Moving Average (ARIMA) model is a sequence of numbers presented in chronological order (ARIMA prediction model plot shown in Figure 1.). A time series is most commonly defined as a sequence of data points taken at consecutive, equally spaced intervals over time. Time series analysis comprises a set of techniques for examining time series data with a view to deriving meaningful statistics and other characteristics of the data. Time series forecasting is the utilization of models to predict future values based on previous observations. Predictive study of vegetable related problems using ARIMA model with reference to Haoxiang Cui [9]. The ARIMA model is constituted of two fundamental components: the Autoregressive (AR) model and the Moving Average (MA) model. The AR model elucidates the interrelationship between the present and lagged values, thereby anticipating future values through the utilization of historical data. In contrast, the MA model monitors the future residuals through the application of linear combinations of past residual terms. The ARIMA prediction model can be expressed as follows:

$$\widehat{p^{\{t\}}} = p_0 + \sum_{j=1}^{p} \gamma_j \, p^{\{t-j\}} + \sum_{j=1}^{q} \theta_j \, \varepsilon^{\{t-j\}} \tag{1}$$

where p is the autoregressive model (AR) order, q is the moving average model (AM) order, $\varepsilon^{\{t\}}$ is the error term between time t and t-1, γ_j and γ_j are the fitting coefficients, and p_0 is a constant term.

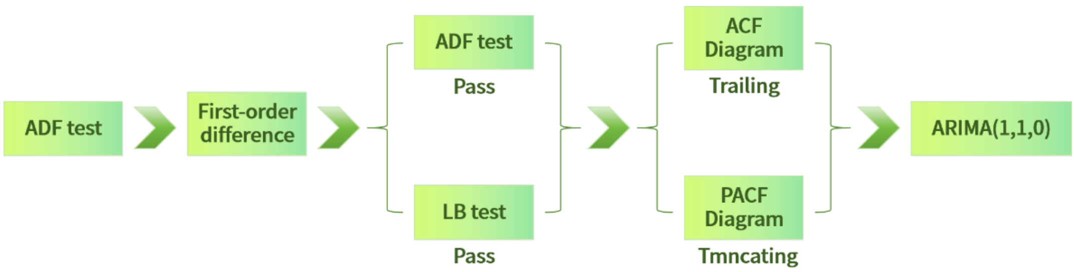

Figure 1. ARIMA prediction model plot

When making statistical inferences about the structure of a stochastic process from the observational record, it is often necessary to make certain simplifying (and approximately reasonable) assumptions about it. The most important of these assumptions is smoothness. The basic idea of smoothness is that the statistical laws that determine the nature of the process do not change over time. With the smoothness of the model ensured, we can determine the value of d to ensure that the value of d does not affect the accuracy of the model. After determining the value of d, referring to the research on ARIMA model by Zhou Xu and Guo Liwen et al, we can finally arrive at the form of ARIMA model [10].

The ARIMA(p, d, q) model takes the form of :

$$\left(1-\sum_{i=1}^{p}\alpha_i L^i\right)(1-L)^d y_t=\alpha_0+\left(1+\sum_{i=1}^{q}\beta_i L^i\right)\epsilon_t \tag{2}$$

where P is the order of the auto-regressive model, d is the order of the difference, and q is the order of the moving average model.

By plotting the ACF and PACF plots of the time series, based on the historical data, the AIC values of the model at different orders are calculated through a computer programming cycle, and the optimal orders of the model are found by finding the orders p and q that minimize the AIC. After determining the optimal order, we perform parameter estimation and test the data found to be statistically significant.

2.2. Spearman-entropy weight model

The Spearman correlation coefficient method is a good statistical method for evaluating the correlation between two variables, which has the most significant feature of not needing to examine the sample size of the variable or the overall distributional characteristics, there only needs to be two sets of data observations of the rank of the data assessment data are pairs of and correspond to the corresponding $i(1\leq i\leq n)$ elements, and according to the same kind of ascending or descending way of arranging the and, to get a new sequence of variables x and y and accordingly, x and y correspond to the $i(1\leq i\leq n)$ elements, with $d_i = x_i - y_i$ as the set of elemental differences, and the formula for calculating the Spearman's correlation coefficient between the random variables and is as follows.

$$\rho=1-\frac{6\sum_{i=1}^{n}\mathrm{m}d_i^2}{n(n^2-1)} \tag{3}$$

The entropy weighting method is an objective assignment method for determining the weights of multiple indicators in a comprehensive evaluation. Its basic principle is based on the concept of information entropy, which is used in information theory to measure the degree of disorder in a system or the amount of effective information provided by data. Referring to the improved entropy weight method used by Gexu Liu and Yichen Zhang et al. in the study of flood risk assessment of metro system due to extreme weather, the entropy weight method (EWM) is used to derive the weights of each evaluation index among various levels according to the constructed hierarchical evaluation system of the impact factors of the development of agro-tourism [11].

Firstly, the construction of evaluation matrix based on indicators, assuming that there are m samples to be evaluated and n evaluation indicators, forming the raw indicator data matrix(matrix of raw indicators shown in Formula 4). The data is then standardized, $\mathrm{x}_{ij}'=(\mathrm{x}_{ij}-\overline{\mathrm{x}_j})/S_j$. The third step is to calculate the weight, calculate the numerical weight of the jth indicator of the ith item $\mathrm{P}_{ij}=\mathrm{x}_{ij}^1/(\sum_{i=1}^{m}\mathrm{x}_{ij}^1)$. The fourth step is to calculate the entropy value $\mathrm{e}_j=-\mathrm{k}\sum_{i}^{m}\mathrm{p}_{ij}\mathrm{l\,n\,p}_{ij}$, where $\mathrm{k}>0,\mathrm{k}=1/\ln\mathrm{m}$, to ensure that $0\leq \mathrm{e}_j\leq 1$, the maximum of e_j is 1. So the entropy value of the jth indicator is $\mathrm{e}_j=-(1/\ln\mathrm{m})\sum_{i}^{m}\mathrm{p}_{ij}\mathrm{l\,n\,p}_{ij}$. In the fifth step, the degree of difference of the jth indicator is defined, and the entropy value method assigns weights according to the degree of difference of the sign value of each indicator, so as to derive the corresponding weight of each indicator $\mathrm{d}_j = 1 - \mathrm{e}_j$, and in the sixth step, the weights are defined as $\mathrm{w}_j = \mathrm{d}_j/(\sum_{j=1}^{n}\mathrm{d}_j)$. Finally, a comprehensive evaluation $\mathrm{F}_i=\sum_{j=1}^{n}\mathrm{w}_j\mathrm{p}_{ij}$.

$$\mathrm{X}=\begin{pmatrix}\mathrm{x}_{11} & \cdots & \mathrm{x}_{1n}\\ \vdots & \ddots & \vdots\\ \mathrm{x}_{m1} & \cdots & \mathrm{x}_{mn}\end{pmatrix} \tag{4}$$

A review of the literature revealed several factors that have been identified as influencing the growth of the agro-tourism industry. These include the local GDP per capita level, urban population density, the incidence of geological disasters, the crime rate, and the urban employment rate.

Accordingly, a scoring system was devised based on these six factors (scoring system model framework diagram shown in Figure 2.).

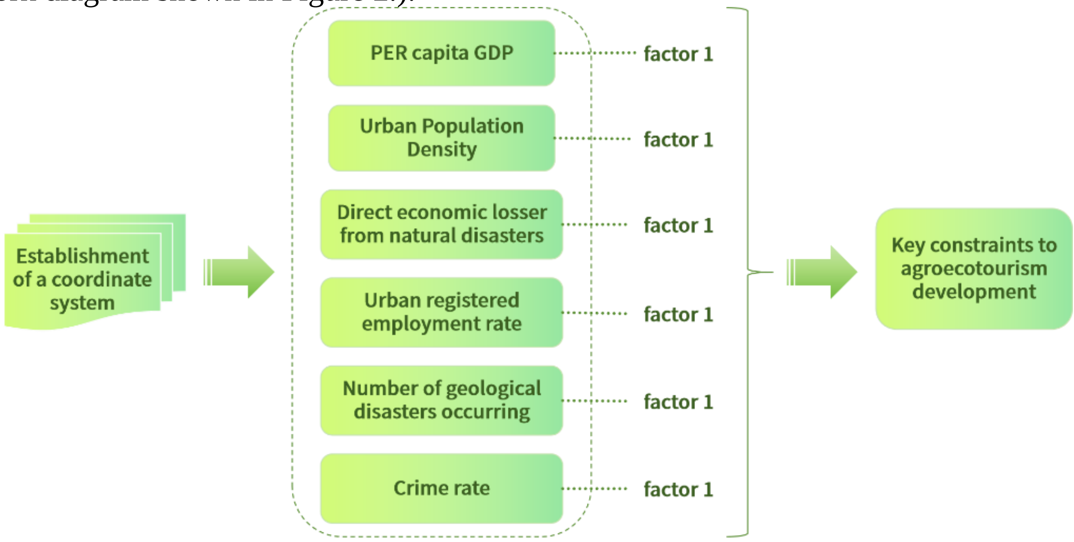

Figure 2. Framework diagram of the scoring system consisting of six key influencing factors

2.3. Gaussian mixture clustering model

A Gaussian mixture model is the precise quantification of things in terms of Gaussian probability density functions (normal distribution curves), which is a model that breaks things down into several formations based on Gaussian probability density functions (normal distribution curves).

Initialize the Gaussian distribution parameters for each cluster: μ、 $\sum$、 α, and construct the probability density of the multivariate Gaussian distribution as:

$$f(X)=\frac{1}{(2\pi)^{\frac{n}{2}}|\sum|^{\frac{1}{2}}}e^{-\frac{1}{2}(X-\mu)^{T}-1(X-\mu)} \tag{5}$$

where μ is the n-dimensional mean vector with shape $(n, 1)$ and $\sum$ is the covariance matrix of the random variable X.

The Gaussian mixture distribution function is:

$$p(X)=\sum_{i=1}^{k}\alpha_i p(X|\mu_i,\textstyle\sum_i) \tag{6}$$

where x represents the probability that we choose each cluster, then their sum is equal to 1.

Next, referring to the Gaussian Mixture Model established by Hangyu Yan in his study of Collaborative Filtering , several variables z_j are introduced to represent the random variable of our choice of a certain distribution, which then takes values between 1 and k. Obviously it corresponds to our x above, that is $p(z_j = i) = \alpha_i$ $p(z_j = i) = \alpha_i$ [12].

So our final conditional probability formula is:

$$p(z_j=i|x_j)=\frac{P(z_j=i)p(x_j|z_j=i)}{p(x_j)}=\frac{\alpha_i p(x_j|\mu_i,\sum)}{\sum_{l=1}^{k}\alpha_l p(x_j|\mu_1,\sum)} \tag{7}$$

The formula means the probability that one of our samples conforms to a certain distribution, denoted γ_{ji}, meaning the probability that data j con forms to distribution i. So we are requiring: $\lambda_{j=\mathrm{argmax}_{i\in 1,2,\ldots k}\gamma_{ji}}$. We will find the probability that the data matches each of the distributions of j, and then obtain the largest probability among them, then the data j will be classified into the clusters corresponding to them. So we perform the maximized likelihood:

$$LL(D) = ln\left(\prod_{j=1}^{m} p(x_j)\right) = \sum_{j=1}^{m} l\,n\left(\sum_{i=1}^{k} \alpha_i p(x_j | u_i, \Sigma)\right) \tag{8}$$

Derive the above function and then the other derivative is 0. Find the corresponding μ and $\sum$, constraints respectively:$a_i \geq 0, \sum_{i=1}^{k} \alpha_i = 1_{\leftrightarrow}$. After that we can find $\alpha_i = \frac{1}{m}\sum_{j=1}^{m} \gamma_{ji}$.

Finally, use the new parameters to update our distribution model to obtain a new probability density, use the new distribution model to calculate the conditional probability of each sample, and continuously proceed to update the model distribution parameters until the model converges and reaches a certain accuracy, stop iterating, obtain the conditional probability of each sample based on our new model parameters, and then obtain the maximum of the value, as λ_j.

3. Results and Discussions

3.1. Solution of ARIMA Model

Firstly, the Augmented Dickey-Fuller test was utilized to test whether a given time series is smooth or not, and it was found that the data of the two-time series before the differencing process were found to be unstable, then, the time series were differenced to the first order (the data are shown for the pre-differenced and post-differenced data before and after differencing shown in Figure 3.).

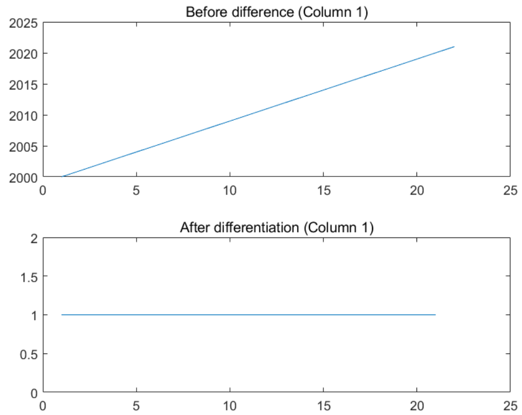

Figure 3. Plot of data before and after first-order differencing of time series

Based on this analysis, we can see that the time series after the first-order differencing, the line in the figure below gradually stabilizes. From this, we can approximate the parameter d of the ARIMA model to be 1. Determining the value of d to be 1 does not affect the accuracy of the model, so d is sufficiently feasible for d to be 1 under the conditions allowed for error, and then the ARIMA model can be determined.After performing first-order differencing on the time series, the ACF and PACF plots can be drawn separately (the ACF and PACF plots shown in Figure 4.).

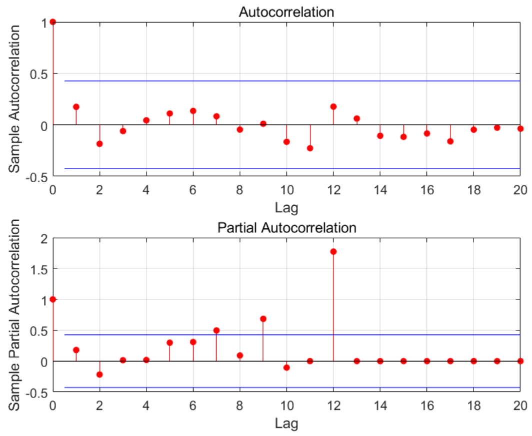

Figure 4. ACF and PACF plots after first-order differencing of time series

After determining the ARIMA optimal solution model, the number of future reported outcomes can be predicted using data on the number of historical extreme weather occurrences in mainland China and data on property damage due to natural disasters (ARIMA optimal solution model prediction results shown in Figure 5.).

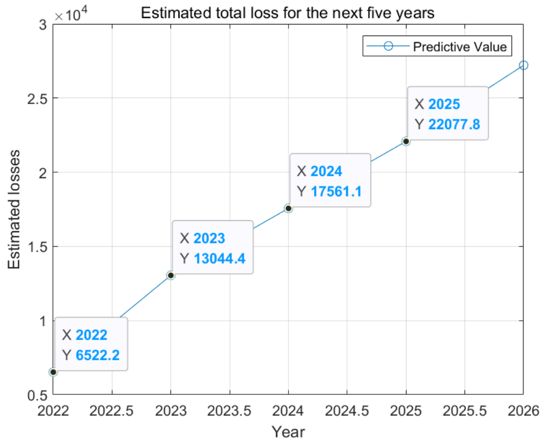

Figure 5. Forecasts of losses due to extreme weather in mainland China for the next five years using the ARIMA optimal solution model

After analyzing the climate models for China, we obtained risk rating projections for the years 2022 to 2026, respectively. Having utilized the historical data from 2000 to 2021, we made predictions by ARIMA time series model and determined the accuracy of the model. The forecasted data are as follows[6522.2,13044,17561,22078,27211]. According to our risk assessment model, the boundary value between low-risk and medium-risk is derived as: 9292.77, and the boundary value between medium-risk and high-risk is derived as: 25273.93. If the predicted value is under the lower limit of low risk, it will be considered that the development of agro-tourism industry in this place is at low risk of loss due to extreme weather; if the predicted value is more than the threshold of high risk, it may mean the development of If the projected value exceeds the high-risk threshold, it may mean that the risk of loss from extreme weather for the development of the agro-tourism industry in the area is high.

3.2. Spearman-entropy weight model solution

Through the National Bureau of Statistics of China (NBS) and other official sources, we have collected data on six key factors: GDP per capita, urban population density, the number of geo-hazards, the crime rate, and the urban employment rate in mainland China over the past few years.

According to the Spearman correlation coefficient matrix plot (the Spearman coefficient matrix heatmap shown in Figure 6.), we find that there is a weak correlation between the variables, so our indicator system is still more rigorous.

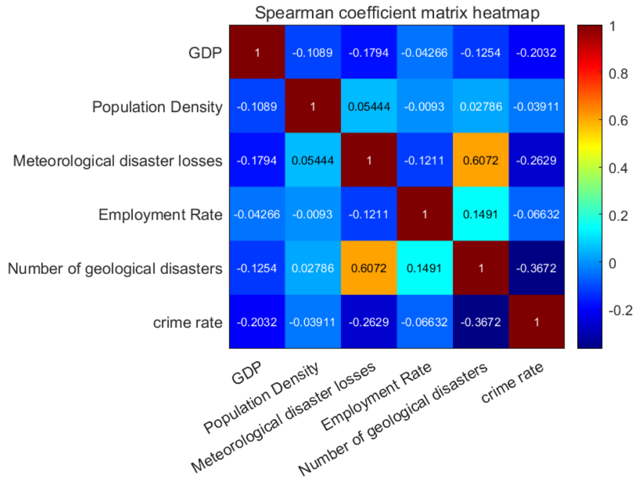

Figure 6. The Spearman coefficient matrix heatmap

The weights of the factors were then derived using the entropy weighting method (weighting charts for the six factors shown in Figure 7. and Figure 8.), and of the six factors affecting the maximization of the benefits of the agro-tourism industry, the GDP per capita accounted for 5%, the urban population density was 21%, the direct economic loss caused by natural disasters was 24%, the employment rate of the urban population was 16%, the number of geologic hazards occurred was 24%, and the regional crime rate was 10%. The location of agro-tourism can be judged on the different factors of the region based on the weights derived from this scoring system, which can further lead to the selection of whether or in which specific location the development will take place.

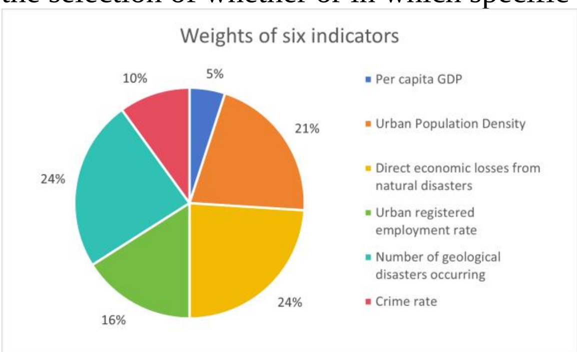

Figure 7. Pie chart of weights of the six indicators

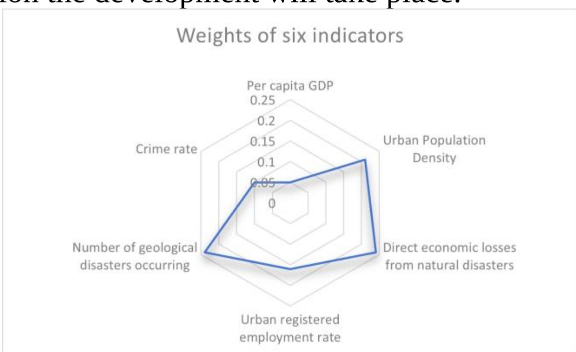

Figure 8. Radar chart of the weights of the six indicators

3.3. Gaussian mixture clustering model solution

According to the constructed evaluation model of agro-tourism development, 100 data collected from six groups of factors were processed for modeling. By analyzing the output data of the Spearman-entropy weight model we found that the data of the six groups of factors satisfy the multivariate Gaussian distribution.

Now the parameters of Gaussian distribution for each cluster are estimated by maximum likelihood estimation based on the predicted evaluations for extreme weather derived from the ARIMA model and the weights of the factors derived from the Spearman-entropy weight model. Firstly, using the elbow method (line graph of data using the elbow method shown in Figure 9.), a line graph is made, as shown in the figure, it is observed that the data decreases significantly at 1-3, so we choose the number of clusters to be 3.

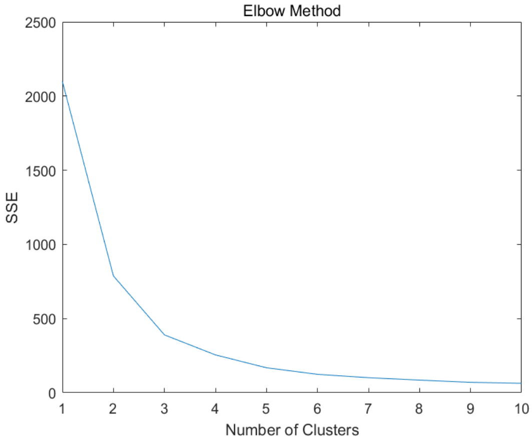

Figure 9. Line graph of data using the elbow method

The results from the Gaussian mixture clustering model were then used to plot bar charts (clustering results) and make scatter plots (Gaussian mixture clustering results shown in Figure 10. and Figure 11.).

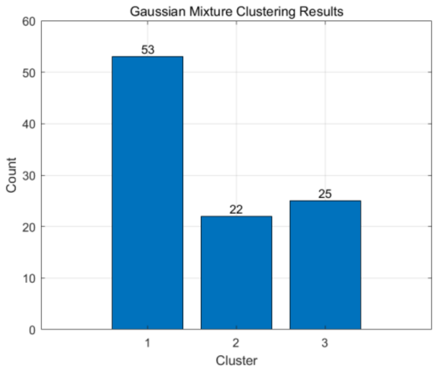

Figure 10. Histogram of Gaussian mixture clustering results

Figure 11. Scatterplot of Gaussian mixture clustering results

Based on the Gaussian mixture clustering model, we visually assessed the regional suitability for agro-tourism development. Bar charts show that high potential regions (53%), medium potential regions (22%) and low potential regions (25%) form three different clusters, respectively. The model results reflect the heterogeneity of these regions in terms of extreme weather risk and impact factor weights. Scatter plots detail the distribution of 100 data for each of the six clusters of factors. After integrating the predicted values of extreme weather and the weights of the six factors, mainland China has the largest percentage of high-potential regions suitable for agro-tourism development, fewer medium-potential regions, and low-potential regions with higher development risks due to the frequency of extreme weather.

4. Conclusion

This paper assesses the potential for agro-ecotourism development by incorporating the predicted values of extreme weather events in mainland China and the weighted values of six key impact factors. High-potential areas with low probability of extreme weather events, low economic losses, high GDP per capita, low disaster frequency and strong economic resilience are suitable for prioritized development and can be actively promoted by governments and investors. Areas with medium potential need to strengthen infrastructure and emergency response to manage risks. For low-potential areas, it is recommended that risk assessment and market research be conducted first, and if necessary, development should be postponed or alternative industries explored, and development should be considered when effective wind control or climate resilience is ensured.

References

[1] Ana M I. Ecotourism, agro-tourism and rural tourism in the European Union[J]. Cactus Tourism Journal, 2017, 15(2): 6-14.

[2] Absalon J A T, Blasabas D K C, Capinpin E M A, et al. Impact assessment of farm tourism sites using a hybrid MADM-based composite sustainability index[J]. Current Issues in Tourism, 2022, 25(13): 2063-2085.

[3] Puška A, Pamucar D, Stojanović I, et al. Examination of the sustainable rural tourism potential of the brčko district of bosnia and herzegovina using a fuzzy approach based on group decision making[J]. Sustainability, 2021, 13(2): 583.

[4] Wang J, Zhou F, Xie A, et al. Impacts of the integral development of agriculture and tourism on agricultural eco-efficiency: A case study of two river basins in China[J]. Environment, Development and Sustainability, 2024, 26(1): 1701-1730.

[5] Andayani S A, Umyati S, Dinar, et al. Prediction model for agro-tourism development using adaptive neuro-fuzzy inference system method[J]. Open Agriculture, 2022, 7(1): 644-655.

[6] Zhang Y, Li W, Li Z, et al. Spatial distribution characteristics and influencing factors of key rural tourism villages in China[J]. Sustainability, 2022, 14(21): 14064.

[7] Liao C, Zuo Y, Law R, et al. Spatial differentiation, influencing factors, and development paths of rural tourism resources in Guangdong Province[J]. Land, 2022, 11(11): 2046.

[8] Joshi S, Sharma M, Singh R K. Performance evaluation of agro-tourism clusters using AHP–TOPSIS[J]. Journal of Operations and Strategic Planning, 2020, 3(1): 7-30.

[9] Cui H. Innovative Research on Pricing and Replenishment Decision Making of Vegetable Commodities Based on ARIMA Prediction Modeling[J]. Highlights in Business, Economics and Management, 2023, 20: 542-548.

[10] Xu Z, Liwen G, Jiuling Z, et al. Prediction of mine dust concentration based on grey Markov model[J]. Shock and Vibration, 2021, 2021(1): 5859249.

[11] Liu G, Zhang Y, Zhang J, et al. Geographic-information-system-based risk assessment of flooding in Changchun urban rail transit system[J]. Remote Sensing, 2023, 15(14): 3533.

[12] Yan H, Tang Y. Collaborative filtering based on Gaussian mixture model and improved Jaccard similarity[J]. IEEE Access, 2019, 7: 118690-118701.