Introduct Risk Prediction Protection Study Based on GA-BP and K-means Clustering

Hanqing Cao [1,†], Yichen Wei [2,†,*] and Binying Fang [1,†]

[1] School of International Economics and Trade, Central University of Finance and Economics, Beijing, China

[2] Faculty of Computing, Harbin Institute of Technology, Harbin, China

[†] These authors also contributed equally to this work

* Corresponding Author Email: vic_yichen_wei@163.com

Abstract. This paper mainly focuses on the risk management problems caused by frequent extreme weather events. Firstly, a grey bp neural network model is used as a heuristic optimization algorithm to construct an IWM model, which is used to decide whether insurance company provides insurance under extreme weather conditions. Secondly, based on the Comprehensive evaluation method (RSR), a DDI evaluation model was constructed to analyze the data of 100 buildings in 10 provinces in China to help developers evaluate the rationality of decisions. Finally, a PM evaluation model is constructed based on AHP and EWM-CRITIC, and K-means clustering is used to conduct comprehensive rating and grading evaluation of 30 low-insurance landmarks, and corresponding protection strategies are designed. The comprehensive application of these models provides useful suggestions and decision support for insurance company underwriting, real estate bidding, community building protection and so on.

Keywords: Grey correlation method; GA-BP algorithm; AHP; RSR.

1. Introduction

Since its commercial operation, the insurance industry has pursued its own economic growth while also aiming to protect the economic assets of its customers. However, in recent years, the frequency of extreme weather events has caused a rapid rise in economic losses. To assist insurance companies and communities in making better insurance decisions, risk management issues arising from frequent extreme weather events were studied.

Using official public data from national statistical offices and weather stations, this paper collects a large amount of data for 30 countries and territories from 1990 to the present. Based on the quantitative analysis of the above data, the relevant insurance intention model is established to help insurance companies make decisions on whether to insure areas in extreme weather. On this basis, through the further expansion of the model, reasonable investment and protection policies are provided for real estate development and community building protection.

2. Insurance Intention Model (IWM) Construction

2.1. Initial Insurance Model (IIM) Construction Based on Entropy Weight Method

This section uses the entropy weight method (EWM) to determine the weights of five relevant indicators [1]. The EWM is an objective assignment method that calculates the information entropy of indicators to determine their degree of variability and corresponding weights [2].

1. Calculate the proportion of indicators using the standardized data provided.

$$P_{ij} = \frac{Y_{ij}}{\sum_{i=1}^{n} Y_{ij}} \tag{1}$$

2. According to the concept of self-information entropy in information theory, the information entropy of each evaluation index can be calculated [3], thus getting:

$$E_j = -(ln\ n)^{-1} \sum_{i=1}^{n} P_{ij} ln\left(P_{ij}\right) \tag{2}$$

3. The weights of the metrics used in the model will be calculated based on information entropy.

$$W_j = \frac{1-E_j}{m-\sum_{j=1}^{m} Ej} \tag{3}$$

Based on the established evaluation model, the composite score of the insurance market's willingness to underwrite index in regions experiencing extreme weather is calculated and ranked. The willingness to underwrite is then classified according to quartile and median in the results, as shown in Table 1. Grade I: Not Recommended for Underwriting (0.596~0.695), Grade II: Careful Selection (0.695~0.79), Grade III: Recommended (0.79~0.889), Grade IV: Highly Recommended (0.885~0.988).

Table 1. Indicator weights for WIU model

Index	weight
Insurance Density	28.287%
Insurance Depth	27.787%
Size of Local Insurance Market	17.474%
Projected Premium Income	16.864%
Claims Losses	9.688%

2.2. Data Processing Based on Grey Correlation Method and Random Forest Method

2.2.1. Gray correlation method

To evaluate weather indicators with a high correlation to the insurance willingness index, we first applied the gray correlation method. This method is less computationally intensive, objective, reliable, and equally applicable to data irregularities [4].

1. Determine the parent sequence and the feature sequence.

First, specify the parent series (reference series), which serves as the ideal standard of comparison and is composed of the optimal values of each indicator in the model. The characteristic sequence refers to the correlation with the parent sequence.

2. Raw Data Processing.

Since the raw data are all greater than 0 and have large values, the homogenization method was chosen to normalize the data.

3. Calculate the gray correlation coefficient between the parent sequence and the feature sequence.

Next, calculate the difference sequence, bipolar difference, and correlation coefficient using a resolution factor of $\xi = 0.5$. Use the obtained correlation coefficient results for further calculations [5].

$$\Delta_i(k) = |x_1'(k) - x_i'(k)| \tag{4}$$

$$M = \max_i \max_k \Delta_i(k),\ m = \min_i \min_k \Delta_i(k) \tag{5}$$

$$\gamma_{1i}(k) = \frac{m+\xi M}{\Delta_i(k)+\xi M} \tag{6}$$

Analyze the results by calculating correlation values and ranking them.

$$\gamma_{1i} = \frac{1}{n} \sum_{k=1}^{n} \gamma_{1i}(k) \tag{7}$$

In this paper, indicators with correlations greater than 0.5 are filtered for subsequent model construction.

2.2.2. Random forest prediction

The model starts by using Random Forests to create a basic regression prediction of the time series [6]. It utilizes time-characterized data with the year uniquely coded and includes annual precipitation, carbon dioxide emissions, historical extremes, and the number and intensity of natural hazards as predictors. The target variable is the average extreme weather index for the upcoming year. The average extreme weather index is a weighted average of the number and intensity of extreme weather and natural disasters.

The comparison of the model's predicted results to the actual results was performed on the test set. Fig.1 shows the sample number on the horizontal axis and the average weather index on the vertical axis. The prediction accuracy is relatively high, with an R-square of 0.876 on the test set and 0.747 on the cross-validation set.

2.3. IWM Model Construction Based on GA-BP Algorithm

After introducing a new index: forecast average climate index, BP neural network model is used to regression it. WIU represents the score calculated for this indicator in IWM.

Genetic algorithm is used as heuristic optimization algorithm and BP neural network is used to calculate and evaluate the underwriting intention index [7]. The learning process of the BP neural network model involves two processes: forward propagation of the signal and back-propagation of the error. The weights from neuron i to j are represented as W_{ji}.

Neuron j receives input $v_j(n)$ during forward propagation in the following manner:

$$v_j(n) = \sum_{i=0}^{m} w_{ji}(n) y_i(n) \tag{8}$$

φ_j is the activation function, and the output $y_j(n)$ of neuron j is:

$$y_j(n) = \varphi_j\left(v_j(n)\right) \tag{9}$$

If the output of the output layer during forward propagation does not match the desired output, proceed to the backpropagation stage to calculate the error. The error is then propagated backwards through each layer, distributed to all units, and used to calculate the error signal of each unit. The error signal produced by neuron $j's$ output is determined by the difference between the actual output $y_j(n)$ and the desired output $d_j(n)$.

$$e_j(n) = d_j(n) - y_j(n) \tag{10}$$

To ensure continuous differentiability, we minimize the root mean square difference and define the instantaneous error energy of neuron j as:

$$E_j(n) = \frac{1}{2} e_j^2(n),\ E_n = \sum_{j \in C} E_j(n) = \frac{1}{2} \sum_{j \in C} e_j^2(n) \tag{11}$$

To optimize weight parameters, the network calculates the total instantaneous error energy by adding the error energies of all output layer neurons. The conjugate gradient method is used to correct the weights of each unit instead of gradient descent to avoid local optimization. Genetic algorithms are utilized as a heuristic optimization technique to determine the optimal initial weights and thresholds for neural network initialization [8]. This is important because these values have a significant impact on network training but cannot be precisely obtained. The process is shown in Fig.2:

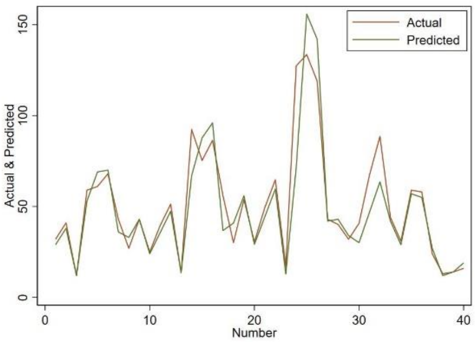

Fig. 1 Comparison between predicted and actual results

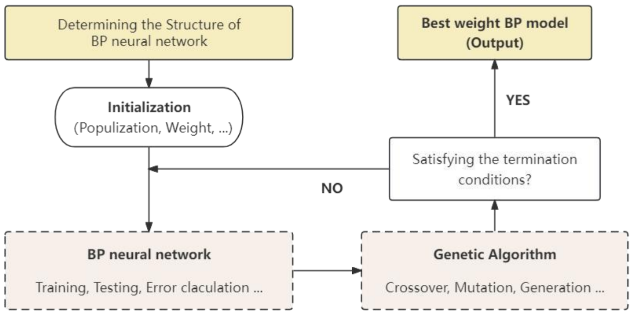

Fig. 2 GA/BP neural network model

2.4. Model Application and Analysis

China and the United States have rapidly developed their insurance markets and have stable national conditions. This can help minimize stochastic errors caused by unexpected factors. The index of insurers' willingness to underwrite and tiering was calculated using data from 2011-2020 for both the United States and China.

Table 2. WIU Classification Result

Year	US	Classification	CN	Classification
2011	0.66	I	0.54	I
2012	0.69	I	0.59	I
2013	0.71	II	0.67	I
2014	0.86	III	0.74	II
2015	0.92	IV	0.80	III
2016	0.95	IV	0.89	IV
2017	0.9	IV	0.93	IV
2018	0.87	III	0.87	III
2019	0.78	II	0.84	III
2020	0.73	II	0.77	II

Table 2 suggests that in both countries, the WIU increases with the frequency of extreme weather events and then decreases after reaching a certain threshold. This implies that insurers are initially willing to meet the demand for insurance when the likelihood and number of claims are lower, but as claims increase, they become less willing to provide coverage. However, as the severity of extreme weather events increases, insurers face a greater economic threat. Although there may be an increase in demand for insurance, insurers are no longer willing to take on long-term risks in exchange for short-term premium income.

3. Construction of Developer Decision Guarantee Model (DDM)

3.1. Indicator Selection and Calculation

To ensure the accuracy and comprehensiveness of the indicators, this paper has selected 10 secondary indicators that have positive and negative impacts on developers' decision-making. Among them, the disaster resilience indicator, as a secondary indicator, is divided into two tertiary indicators due to the lack of quantifiable data sources for disaster resilience: the type of wall material and the cost of building structure design. The weights were determined by averaging the independence weight method and the coefficient of variation method, based on a literature review. This approach considers both the volatility and correlation of the data, ensuring that the model results are comprehensive and accurate.

3.1.1. Independence weighting method

The independent weight method is applicable to the indicators themselves with some correlation, belonging to the same system under the indicators to calculate the weights. The strength of the covariance between indicators is utilized to determine the weights. If an indicator is said to have a strong correlation with other indicators, it means that there is a large overlap of information, implying that the weight of the indicator will be lower, and vice versa.

1. Calculate the complex correlation coefficient R and its reciprocal $1/R$.

$$R = \frac{\sum (y-\bar{y})(\hat{y}-\bar{y})}{\sqrt{\sum (y-\bar{y})^2(\hat{y}-\bar{y})^2}} \tag{12}$$

2. Sum the complex correlation coefficients to normalize the calculation to obtain the weights, output the weight matrix.

$$\frac{x_i}{\sum_{i=1}^{n} x_i} \tag{13}$$

3.1.2. Coefficient of variation method

Applicable to treating data variability as a piece of information, using the degree of data volatility as a measure of indicator weights.

1. Both variables are positive variables, no need for data regularization, directly based on standardization to calculate the coefficient of variation V_j can be.

$$V_j = \frac{S_j}{A_j} \tag{14}$$

2. Calculate weights and scores, output weight matrix.

$$\omega_j = \frac{V_j}{\sum_{j=1}^{n} V_j} \tag{15}$$

3. The result of the weights was obtained as follows:

$$\text{Score}_i = \sum_{j=1}^{n} \omega_j r_{ij} \tag{16}$$

3.2. DDM Model Construction Based on RSR Method

1. Construct the data matrix to obtain the rank order

Total rank sum ratio method:

$$R = \left(R_{ij}\right)_{m*n} \tag{17}$$

Non-integral rank sum ratio method:

$$R_{ij} = 1 + (n-1)\frac{X_{ij}-\min\left(X_{1j},X_{nj},\cdots,X_{nj}\right)}{\max\left(X_{1j},X_{nj},\cdots,X_{nj}\right)-\min\left(X_{1j},X_{nj},\cdots,X_{nj}\right)} \tag{18}$$

$$R_{ij} = 1 + (n-1)\frac{\max\left(X_{1j},X_{nj},\cdots,X_{nj}\right)-X_{ij}}{\max\left(X_{1j},X_{nj},\cdots,X_{nj}\right)-\min\left(X_{1j},X_{nj},\cdots,X_{nj}\right)} \tag{19}$$

2. Calculate rank sum ratio RSR value and ranking

The data has been initially solved for the weights of the variables by the entropy weighting method, so at this point the weights are different and the weighted rank sum ratio should be calculated as WRSR with the following formula:

$$WRSR_i = \frac{1}{n}\sum_{j=1}^{m} W_j R_{ij} \tag{20}$$

3. Determine the distribution of RSR and analyze the results

The distribution of RSR is the specific cumulative frequency of values expressed in probability units Probit. Linear regression was performed using the RSR distribution values as the independent variable and the Probit value as the dependent variable to obtain the equation as follows [9]:

$$y = -0.31 + 0.137 \times \text{ Probit} \tag{21}$$

Based on the F-test analysis, the P-value was found to be 0.000^{***}, indicating significant results and rejecting the null hypothesis that the regression coefficient is 0. Additionally, the model's goodness-of-fit ($R2$) is 0.869, indicating excellent performance and meeting the necessary requirements.

4. Sorting and grading the evaluation targets

According to the estimated value of RSR (WRSR) obtained from the regression equation, the evaluation objects are ranked. The grades are 1,2, and 3 , corresponding to the degree of recommendation of the community and real estate developers to build at the site: low, medium, and high,as shown in Table 3.

Table 3. RSR Classification Result

Level	Percentile Threshold	Profit	RSR Critical Value
I	< 15.866	< 4	< 0.2363
II	$15.866 \sim$	$4 \sim$	$0.2363 \sim$
III	$84.134 \sim$	$6 \sim$	$0.5094 \sim$

4. Protected Mode (PM) Construction

4.1. Calculation of Indicator Weights for Protection Model (PM)

The precision of weighting is critical. We utilized a combination of the Analytic Hierarchy Process (AHP) for subjective weighting and the Entropy Weight Method (EWM) and Criteria Importance Though Intercriteria Correlation (CRITIC) method for objective weighting. We named this objective combination method the EWM-CRITIC method.

4.1.1. The AHP method

AHP is the method of dividing the elements that are always related to decision-making into a hierarchical structure, and based on which qualitative and quantitative analyses are carried out, and the structure and main steps are as follows:

1. Construct a hierarchical model, mainly divided into three levels of objectives, guidelines and programs, containing five primary indicators and 11 secondary indicators.

2. Construct the pairwise comparison matrix using the pairwise comparison method and the 1-9 comparison scale for factors at the same level that belong to or affect each factor of the previous level, up to the lowest level. The matrix was constructed in accordance with the diagram of a subjective hierarchical single ranking with reference to the expert evaluation criteria, as shown in Fig.3.

3. The algorithm calculates the final weight vector (weights) using the sum-product method. The judgment matrix columns are first normalized and then summed up by columns.

$$b'_{ij} = \frac{b_{ij}}{\sum_{i=1}^{n} b_{ij}} \; i,j = 1,2,\ldots,n \;\; W_i = \sum_{j=1}^{n} b_{ij} \; i = 1,2,\ldots,n \tag{22}$$

Next, normalize $W' = (W'_1, W'_2, \ldots, W'_n)^T$, and obtain W as an approximate solution for the desired eigenvector.

4. Conduct a consistency test to ensure the correct hierarchical ordering for AHP. Calculate the maximum characteristic root and corresponding eigenvectors for each pairwise comparison array. The consistency index (CI) of 0.027 indicates a value close to 0, and the consistency ratio (CR) of 0.018 is less than 0.1, which means that the test has passed. Therefore, the eigenvectors can be used as weight vectors.

4.1.2. The EWM-CRITIC method

Since objective weight determination methods have a different focus on the data, our team decided to synthesize the Entropy Weight Method (EWM) and the CRITIC weight method so as to focus on the volatility and correlation of the data in a comprehensive manner. CRITIC utilizes the volatility of the data or the correlation between the data to calculate the weights, which is suitable for the data with certain correlation and volatility of the indicators themselves, so it is more suitable for the original data of this model.

1. Calculate the variability of the indicator in the form of standard deviation, Sj is the standard deviation of the j_{th} indicator.

2. Calculate the indicator conflictiveness, which is reflected in the form of correlation coefficient as follows, where r_{ij} is the correlation coefficient between i and j.

$$R_j = \sum_{i=1}^{p} \left(1 - r_{ij}\right) \tag{23}$$

3. Combine the two indicators in 1 and 2 to calculate the amount of information Cj with weights Wj as follows:

$$C_j = S_j \sum_{i=1}^{p} \left(1 - r_{ij}\right) = S_j \times R_j \tag{24}$$

$$W_j = \frac{C_j}{\sum_{j=1}^{p} C_j} \tag{25}$$

4.1.3. Calculation of composite weights

The assignment decision is calculated based on both subjective and objective factors, with a weight of 40% and 60% (30%+30%), respectively, as shown in Table 4.

Table 4. All kinds of weights of the Protection Model

Unit %	CRITIC	AHP	EWM	Combination Weight
Internet Search Volume	4.926	15.005	24.184	14.735
Scenic Area Grading	11.177	10.566	6.772	9.6111
History to Date	6.977	25.669	15.417	16.9858
Repair Cycle	11.471	4.165	10.133	8.1472
Functionality	16.683	17.068	9.751	14.7574
Floor Place	7.757	1.697	1.343	3.4088
Repair Difficulty	7.849	4.165	1.851	4.576
Residential	8.898	2.578	2.198	4.36
Store	9.333	2.613	3.153	4.791
Traffic	8.146	9.37	2.619	6.9775
Ticket Revenue	6.783	7.103	22.579	11.6498

4.2. Protection Level Model

4.2.1. K-Means clustering

Because of the need to inform community leaders' decisions on whether to protect and how to protect, we decided to target strategies by clustering the Community Protection Index (CPI) calculated with the weights described above and grading them [10].

Firstly, the final conservation model score, CPI, is calculated through the above weights as a data base to be used for clustering and secondly the optimal number of clusters is explored. Here the elbow method is used to determine the better number of clusters as shown in Fig. 4. Since the loss value plateaus and basically does not change anymore after the number of clusters $>= 3$,3 is considered as the optimal number of class clusters. And the landmarks are categorized into high, medium and low categories based on the CPI, and the insurance model is combined to give the final strategy.

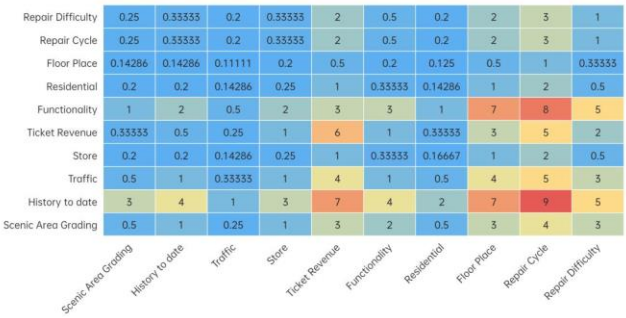

Fig. 3 Pairwise comparison matrix

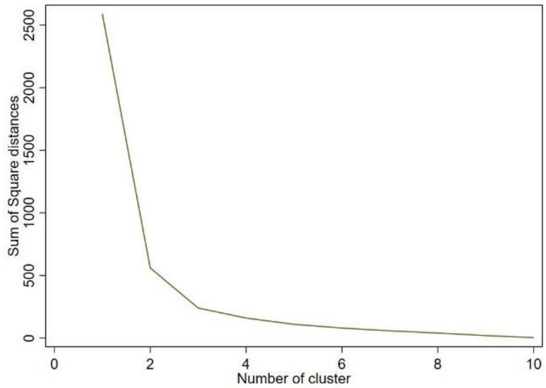

Fig. 4 Line graph of clusters' number versus loss

4.2.2. Rating results and strategy

To facilitate the use of our model and provide community leaders with targeted strategies that maximize community benefits, we developed a comprehensive evaluation model. This model is based on a weighted average of the above tiered protection model combined with the willingness to insurance model, which guides the final strategy development. This paper divides the policy overview into three levels of protection models: low, medium and high, as shown in Table 5:

Table 5. RSR Classification Result

Strategy	High protection strategy	Appropriate protection strategy	Low protection strategy
Score	≥80	80~70	≤70

Taking the example of the Bell and Drum Tower in Xiuning County, Anhui Province, China, constructed during the Yuan Dynasty (1345), this classical landmark possesses significant historical and cultural value. This study applies protection and insurance models to obtain a PM score of 86.43 (high) and a WIU score of 60.11 (low). It is evident that the Bell and Drum Tower has a high protection level, while insurance companies show a lower willingness to underwrite. Therefore, the surrounding community can adopt the high protection strategy mentioned earlier.

5. Summary

First, the IWM model is applied to the decision of whether insurance company provides insurance under extreme weather conditions. The entropy weight method (EWM) is used to select the relevant data and calculate the insurance company's underwriting willingness index (WIU). Through the study of grey BP neural network model, the relationship between extreme weather and WIU is revealed.

Secondly, the DDI evaluation model is constructed. The model uses comprehensive evaluation method (RSR) to analyze the data of 100 buildings in 10 provinces in China, and uses independent weight method and coefficient of variation method to calculate the weights of indicators. Three evaluation levels were built to help developers evaluate the reasonableness of decisions.

Finally, a PM evaluation model is constructed based on AHP and EWM-CRITIC, five key indicators are evaluated, and 30 low-insurance landmarks are evaluated by K-means clustering analysis, which provided substantial guidance for the protection of low-insurance land targets.

References

[1] Ming Liang; Gen Yang; Xiaojun Zhu; Hua Cheng; Liugen Zheng; Hui Liu; Xianglin Dong;Yanhai Zhang.AHP-EWM Based Model Selection System for Subsidence Area Research [J]. Sustainability, 2023, Vol.15(7135): 7135.

[2] Yan Li;Chengpeng Lu;Guang Liu;Yufeng Chen;Yong Zhang;Chengcheng Wu;Bo Liu;Longcang Shu.Risk assessment of wetland degradation in the Xiong'an New Area based on AHP-EWM-ICT method[J].Ecological Indicators,2023,Vol.153: 110443.

[3] Cong Cao;Tianlan Wei;Shengyuan Xu;Fan Su;Haiquan Fang.Comprehensive evaluation of higher education systems using indicators: PCA and EWM methods[J].Humanities and Social Sciences Communications,2023,Vol.10(1): 1-12.

[4] Xuesheng Li;Lihang Gao.State Assessment Method of Electricity Meter Based on Grey Correlation Analysis[J].Process Integration and Optimization for Sustainability,2023,Vol.7(5): 1149-1156.

[5] Chen, Cang;Zhao, Zhiqin;He, Suizhi.Analysis of water quality of Guangzhou waterworks based on grey correlation method[J].Journal of Physics: Conference Series,2021,Vol.2012(1): 012103.

[6] Benjamin Goehry;Hui Yan ;Yannig Goude ;Pascal Massart ;Jean-Michel Poggi .Random Forests for Time Series[J].Revstat Statistical Journal,2023,Vol.21(2): 283-302.

[7] Libiao Bai;Yuqin An;Yichen Sun.Measurement of Project Portfolio Benefits With a GA-BP Neural Network Group[J].IEEE Transactions on Engineering Management,2024,Vol.71: 4737-4749.

[8] Junqi Zhu;Haotian Zheng;Li Yang;Shanshan Li;Liyan Sun & …Jichao Geng.Evaluation of deep coal and gas outburst based on RS-GA-BP[J].Natural Hazards,2023,Vol.115(3): 2531-2551.

[9] Fangxing Zhao;Changjun Zhu;Budong Li;Shiyan Wang;Wenlong Hao;Xiaobo Liu.Application of PCA-RSR Model in Reservoir Water Quality Evaluation[J].Nature Environment and Pollution Technology,2022,Vol.21(2): 571-579.

[10] Anurag Goel;Angshul Majumdar.Contrastive deep convolutional transform k-means clustering [J]. Information Sciences,2024,Vol.661: 120191.