A Study on Insurance Rate Decision Making in Disaster Frequent Areas Based on Pearson Correlation Analysis and Random Forest Algorithm

Ruoxi Wang, Yucong Gou, Qiangfeng Li*

College of agriculture and animal husbandry, Qinghai university, Xining, China, 810016

*Corresponding author: 1324985124@qq.com

Abstract. In recent years, earthquakes, floods, droughts, extreme heat, and other natural disasters have occurred more and more frequently in all corners of the earth, and these disasters have caused huge economic losses. The pressure on insurance companies to pay out in natural disaster-prone areas is much greater than in other areas, to make insurance companies operate normally, they need to set insurance rates according to local conditions, to set reasonable insurance rates, this paper obtains the highest correlation between the total disaster loss and the insurance loss through the Pearson's correlation analysis, and then adopts the Random Forest Algorithm, to obtain the highest weight for the number of people affected by the corrected disaster, and then adopts the Maximum Minimum Distance algorithm to classify all models, introduce the disaster damage factor (DDF) and normalize it, and use the BP neural network regression model to establish the evaluation model of the relationship between premiums and the impact factor, and the data visualization results show that the model fit is intuitively good. Combined with the theory of the time value of money, the insurance actuarial model was established from the perspective of insurance companies, and the model was validated by inputting relevant data from Guizhou China, and Brazil.

Keywords: Pearson correlation, Random Forest Algorithm, Premium Rate.

1. Introduction

Every rhythmic movement of the Earth affects the way and rhythm of human production and life, and natural disasters, like a manic dancer in nature, sometimes dance and sometimes wreak havoc, causing serious loss of life and property. In recent years, natural disasters such as earthquakes, floods, dust storms, extreme heat, and other natural disasters are occurring more and more frequently in all corners of the earth, from minor property damage to hundreds of deaths the tragic events, the earth's cold often makes us overwhelmed, the governments of various countries to improve emergency response capabilities, in addition to strengthening the daily preventive drills and exercises, but also to actively take out insurance to cope with the unpredictable and huge losses.

For insurance companies, the lack of sufficient data to support actuarial pricing and the high-cost pressures of catastrophe insurance have made it a challenge to set payout rates. At the same time, the occurrence of catastrophe has the characteristics of low frequency and high loss, protection becomes the center of the topic, as how to balance the insured rate and the payout rate, in the case of insurance companies not leading to excessive losses, but also can widely protect the people's property safety, the expansion of the business to cover the frequency of occurrence of high impact on the production of life storms and floods, the formation of a long-term mechanism, and so the establishment of the conditions of the insured and the compensation amount of the mathematical model is very critical. Mathematical modeling is very critical.

Lin Qiang[1] and other scholars proposed and established the Stackelberg game model by conducting in-depth research on the optimal decision-making of insurance companies under the influence of agricultural natural disasters. In this field, Liu Lixin[2] systematically explores the role of insurance mechanisms in natural disaster management from multiple perspectives and analyzes the order of disaster risk formation as well as the impact of catastrophe risk formation mechanism on the catastrophe insurance system. He achieves insurance payout by adopting a sub-risk decentralized structure. You Yuxuan's[3] research focuses on earthquakes in terms of insurance sub-risk zoning and

pricing and establishes a corresponding analytical framework using the POT-Copula model. Meanwhile, Gang Yang's [4]work focuses on the pricing study of temperature derivatives under extreme weather changes, using the ARMA model for modeling and analysis. Yang's [5]research, which spans the period from 1989 to 2019, systematically identifies the main factors affecting weather index insurance in the Jiamusi region through a comprehensive study of many years of meteorological data in the region and applies the Copula model to derive the most suitable insurance rates. Together, these studies provide profound theoretical and empirical support for decision-making in the field of insurance

So this thesis used Pearson's correlation coefficient analysis, random forest algorithm[6] to correct the data, and BP neural network regression model to establish an assessment model of the relationship between premiums and influencing factors in natural disaster areas, and pricing strategies for certain areas in the next few years due to the impact of various factors on the forecast[7].

2. Actuarial modeling of insurance under consideration of the impact of extreme catastrophes

2.1. Data preparation

The current collection contains information on various types of disasters suffered by different countries and regions at different times, such as earthquakes, floods, and extreme temperatures. This information covers the economic losses, number of casualties, and insurance-related losses of disasters, and has high data integrity and credibility. In our study, the type of catastrophe, the severity of the catastrophe, and the governmental financial loss caused by the catastrophe are identified as the main factors that have a significant impact on the insurance company's revenue, and they become the main reference indicators of this model.

Considering that the number of injuries and deaths caused by a catastrophe is an important reflection of the severity of the catastrophe, these data are introduced as one of the reference indicators in the model. This not only contributes to a more comprehensive assessment of the impact of the disaster but also correlates with the geographic location and urban distribution of the country, making the model more comprehensive and accurate. Given the differences in population density and level of economic development in different regions, the local Consumer Price Index (CPI) was introduced as another reference indicator to improve the accuracy of the model and to eliminate possible data errors due to the number of casualties in different population bases. This combination of local factors makes our model more complete and more accurate.

For the missing content in the data, the method of linear interpolation will be used to complete the processing: first of all, Pearson correlation analysis is used between the above indicators to derive the degree of correlation between the different indicators, to select the highest degree of linearity of the other indicator to interpolate the missing data to complete the missing data, and this process ensures that the difference of the value of the data to the credibility of the data.

Take insurance loss as an example, we can see that the total loss and its highest degree of correlation, choose the total loss of insurance loss in the missing data to fill. The linear interpolation formula is as follows：

$$y = y_1 + \frac{(x-x_1)\cdot(y_2-y_1)}{x_2-x_1} \tag{1}$$

Among them.

X is the position to be interpolated

$x_1$andx_2 are the *x*-values of the known data points

$y_1$andy_2 are the *y*-values of the known data points

The same method is used for other missing data.

Weighting analysis of indicators using the random forest algorithm. Next, the data are normalized and the maximum minimum distance algorithm is used in this model. This method first searches for

clustering centers based on a determined distance threshold, and then divides the model samples into the categories corresponding to each clustering center based on the nearest neighbor rule.

For two samples in a set of data c, we use $d(c_i, c_j)$ to denote the Euclidean distance between them, then we have:

$$d(c_i, c_j) = \max_{c_k}(\min_{c_j}(d(c_k, c_j)) \tag{2}$$

Where c_k denotes the data value of the clustering center where the two data sets are located.

From the above steps, it can be seen that the maximum-minimum distance algorithm can be summarized as follows: take the "maximum Euclidean distance between test classes" as the condition for pre-selecting the initial clustering centers; determine the remaining clustering centers according to the maximum distance among the minimum distances; and after determining all the clustering centers, then classify all the patterns into each class according to the nearest distance. to each category.

Finally, after processing the data of each index, the Disaster Damage Factor (DDF) is introduced in this paper to express the economic loss caused by the disaster in a certain area at different times, and its formula is as follows:

$$\mathrm{DDF} = \sum_{i=1}^{I} w_i d_i \tag{3}$$

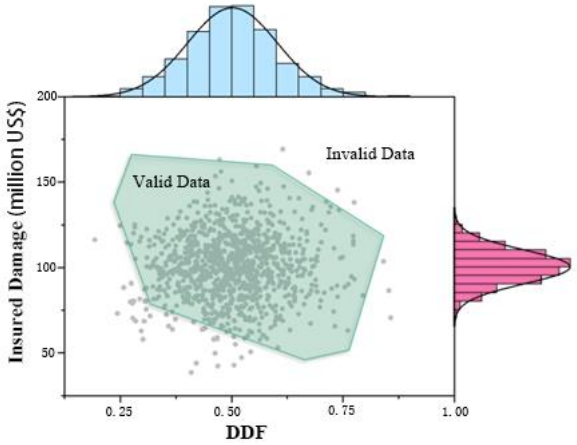


Figure 1 Insurance losses

Among them, for the quantitative value of each indicator, normalization will be used, and its coefficient indicates the corresponding weight of the first indicator, as shown in Fig. 1, and the insurance company can make decisions based on the value. We query the basic business and capital income of a typical insurance company to measure the insurance loss in the above data, and the final situation will be as shown in Fig. 1:

The horizontal coordinate in the above figure is the corresponding DDF value, and the vertical coordinate is the amount of insurance loss, we need to clean the extreme data, using the principle of normal distribution, you can get the normal distribution of the data based on the scatterplot, the green area is the effective date, so the comprehensive consideration of whether the insurance is profitable or not, and ultimately get the insurance strategy as follows:

$$d_{ins} = \begin{cases} 1, & \mathrm{DDF} \le 0.65 \\ 0, else \end{cases} \tag{4}$$

Where d_{ins} denotes the decision variable of whether to insure or not, if it takes the value of 1 it means that it is possible to insure, and vice versa it is not.

2.2. Proposing an assessment model

For the correlation relationship between DDF and known premiums, a BP neural network regression model is used to model the assessment of the relationship between premiums and the influencing factors. The BP [8]neural network minimizes the error between the output of the neural network and the actual target by using a gradient descent algorithm. The cost function of the model is

$$J(W,b) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(a^{(3)}) + (1-y^{(i)})\log(1-a^{(3)})] \tag{5}$$

The gradients of the output layer weights and biases and the gradients of the hidden layer weights and biases are shown

$$dW^{(2)} = \frac{1}{m}dz^{(3)}\cdot a^{(2)T}, db^{(2)} = \frac{1}{m}\sum_{i=1}^{m} dz^{(3)} \tag{6}$$

$$dW^{(1)} = \frac{1}{m}dz^{(2)}\cdot a^{(1)T}, db^{(1)} = \frac{1}{m}\sum_{i=1}^{m} dz^{(2)} \tag{7}$$

By collecting disaster and disaster claims data from Guizhou, China and Brazil as a dataset, after data screening and processing, the dataset is divided into a training set and a test set with a ratio of 9:1, and the training set is used as a dataset in the establishment of the assessment model[9], and the test set will be used in the next section, the evaluation of the assessment model.

2.2.1 Evaluation of assessment models - evaluation with data from two continents

Testing and evaluating the model using a delineated dataset.

2.2.2 pricing strategy

Consider the theory of the time value of money, which is the opportunity cost of holding money or investing in assets. The application of the time value theory involves discounting, which is the process of converting future cash flows or earnings to their current present value by considering interest rates and opportunity costs. Understanding time value helps in making informed decisions in financial planning, investment decisions, and capital budgeting. Considering the different values of money at different points in time to better compare and evaluate different investment opportunities reflecting the accuracy of pricing strategies as well as comprehensive sensitivity analyses[10], we build an optimization model from the insurance company's perspective with an objective function of:

$$\max \sum_{t=1}^{T} d_{ins}(t)\cdot(c_{income}(x_t) - c_{operation}(x_t, t, \mathrm{DDF}_t) - c_{lost}(\mathrm{DDF}_t)) \tag{8}$$

Where T denotes the total number of insured years in *$/year*, t denotes the current year, x_t denotes the insured amount t in $/year.DDF is the disaster factor of the studied area in the year t based on the historical data using machine learning algorithms, and the corresponding $d_{ins}(t)$ is the decision variable of insured in the year t, which can be determined by Eq. $c_{income}(x_t)$denotes the insurance company's income in the year t, which is in the following form:

$$c_{income}(x_t) = x_t\cdot(1+r)^{T-t} \tag{9}$$

Where r denotes the local bank interest rate, for ease of calculation we have chosen 10% here after looking up the data.

$c_{operation}(x_t, t, DDF_t)$denotes the cost of operating the insurance company in a year t, and the expression obtained from fitting the data is as follows:

$$c_{operation}(x_t, t, \mathrm{DDF}_t) = k(1-e^{-\frac{T-t}{\mathrm{DDF}_t}})\cdot x_t^2 \tag{10}$$

The coefficient k is also a least squares estimate based on Eq. DDF_t. The coefficient is estimated as the least squares estimation of the data based on Eq. 1.532. The above equation indicates that whenDDF_t is kept constant, the operating cost tends to decrease as t increases with fewer years of insurance; and if the year t is kept constant, the smaller the value represents the smaller the insurance loss due to the disaster in the region, the lower is the operating cost. $c_{lost}(DDF_t)$It represents the insurance losses for different values of DDF, which can be fitted once by least squares estimation to obtain the linear cost from the data results.

According to the above, it can be seen that the overall objective function is a convex function (convex function), which ensures the accuracy of the model while the model solving efficiency is high.

The constraints that the model adheres to (SUBJECT TO) are:

$$0 \le x_t \le \min(Dem_t, Fin_t) \tag{11}$$

The above equation Dem_tdenotes the demand capacity of the market to ensure that each year's insured revenue does not exceed the market's demand, and its value can be a pre-determined market demand plan, or it may be based on the variables derived from market research and forecasting, for the sake of simplifying the model, we directly use the determined market demand plan here. Fin_tIndicates the financial capacity of the insurance company, and one possible way to express its value is based on the company's annual budget or financial plan.

3. Result

3.1. Random Forest Algorithm Analysis Results

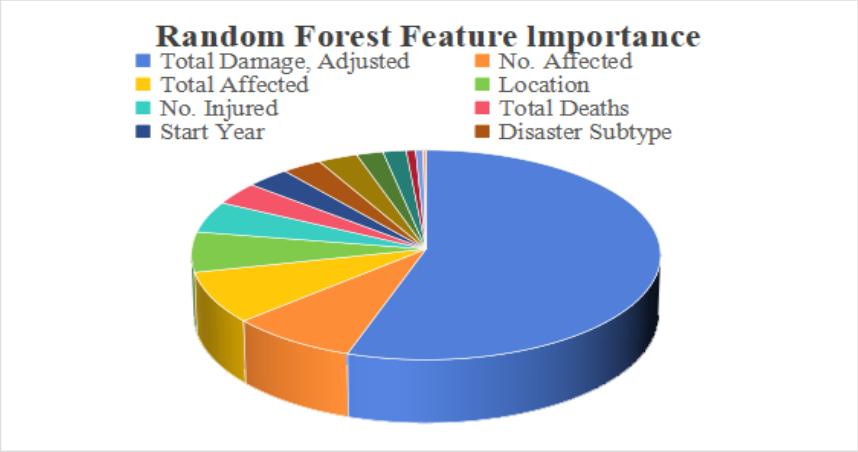


Figure 2 Weighting diagram (insurance losses as dependent variable)

Next, we take the insurance loss as the dependent variable and other indicators as independent variables, and use the random forest algorithm to analyze the weights of the indicators, and the results of its analysis are shown in Figure 2 .

As can be seen in Figure 2, the weight of total disaster loss is the largest, which more than half of the overall, and the total number of affected people No. Affected is the second order of the above figure. The reason is that the correlation between the other indicators and the total loss of disaster is high, which can directly reflect the insurance loss through the total loss of disaster, and the linear correlation is high. The reason is that the correlation between other indicators and the total disaster loss is high, which can directly reflect the insurance loss through the total disaster loss, and the linear correlation is high, so in order to improve the robustness of the model for the analysis of the other indicators, the total disaster loss is removed and the random forest algorithm is used to calculate the new weight distribution graph as in Figure 3:

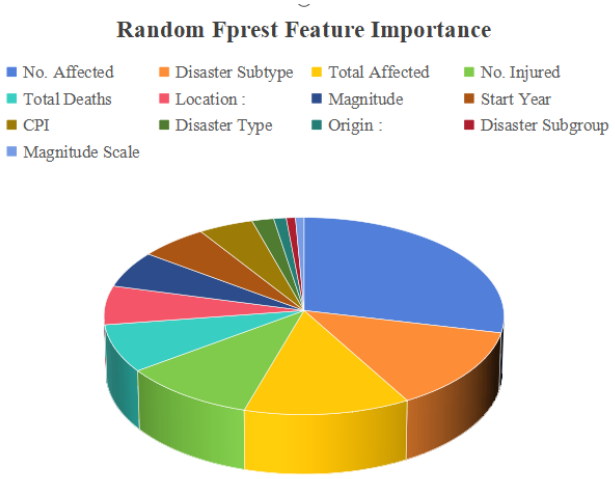


Figure 3 Random Forest algorithm analysis results (total loss removed)

From the Figure 3 it can be seen that No. Affected has the highest weight.

3.2. Assessment of model evaluation results

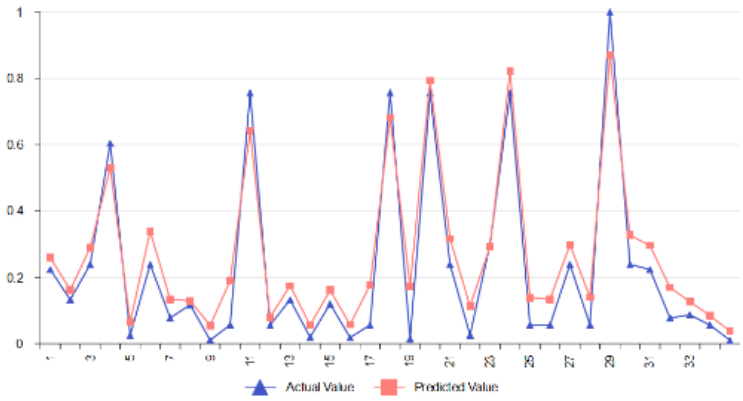


Figure 4 Between the tested results and the real values

The model was tested and evaluated using the delineated dataset, and the comparison between the results of the test and the true values is shown in Figure 4, which shows that the model fits intuitively well.

The results of the quantitative assessment of the model are shown in Table 1.

Table 1 Quantitative assessment results of the model

	MSE	RMSE	MAE	MAPE	R^2
Cross Validation Set	0.005	0.053	0.038	25.869	0.751
Test Set	0.001	0.033	0.027	18.893	0.83

Again, good results were presented.

4. Conclusions

The highest degree of correlation between total and insured losses was obtained through Pearson correlation analysis and other missing data were filled in. Subsequently, the highest weight of the number of people affected by the disaster was determined after correction using the random forest algorithm, the pattern was classified by the maximum minimum distance algorithm, and the disaster damage factor (DDF) was introduced and normalized. Then, a BP neural network regression model was used to establish an assessment model between the premium and the impact factor, and the data visualization results intuitively presented the excellent fitting effect of the model. Considering the theory of the time value of money, the insurance actuarial model was constructed from the perspective

of insurance companies. Finally, our model was verified by inputting relevant data from Guizhou, China and Brazil.

Acknowledgments

This work were supported by "In 2022, the second batch of central forestry grassland ecological protection and restoration fund Datong Beichuan River Park National Nature Reserve subsidy project Qingcai Zi Ring Word (2022) No. 724" and "Kunlun Talents · Leading Talents in Science and Technology" training Project (No. 1 of Qingtalentzi (2022))

References

[1] LIN Qiang, LIU Huang, XU Junxin et al. Research on the insurance mechanism of agricultural natural disasters based on CVaR [J]. Operations Research and Management,2023,32(04):169-176.

[2] LIU Xinli. Catastrophe Insurance System from Comprehensive Risk Management[J]. China Finance,2021,(01):61-62.

[3] You Yuxuan. Research on Risk Zoning and Pricing of Earthquake Catastrophe Reinsurance[D].Yunnan University of Finance and Economics,2023.DOI:10.27455/d.cnki.gycmc.2023.000740.

[4] YANG Gang,TANG Yong,WANG Wenzhuo. Research on Pricing of Temperature Derivatives in the Context of Climate Finance - Based on Machine Learning Algorithm[J]. North China Finance,2023,(07):39-53.

[5] Yang C.W.. Design of weather index insurance based on Copula method [D]. Northwest Agriculture and Forestry University 2023 DOI:10.27409/d.cnki.gxbnu.2023.000932

[6] Jiang Chuan. Long-term health insurance pricing considering repeat hospitalization [D]. East China Normal University,2023.DOI:10.27149/d.cnki.ghdsu.2023.001151.

[7] Qiang Li,Yanjiao Chen. Prediction of commercial pension insurance purchase behavior based on random forest[J]. Technology and Industry,2022,22(08):271-275.

[8] Cai Chao, He Xinyi, Li Li. Distributed neural network regression modeling for large-scale data[J]. Statistics and Decision Making,2023,39(17):34-39.DOI:10.13546/j.cnki.tjyjc.2023.17.006.

[9] SHI Tao,XIE Wusan,ZHANG Li,YANG Yuanjian. GIS and spatialized application of storm flood risk assessment[J].A case study of Wuhu city Journal of Natural Hazards.2015(05).

[10] KONG Yan,GUO Qingkun. Egypt's Climate Change Response Strategy and Its Dilemma[J]. Arab World Studies,2023,(01):70-94+159.