A study of property insurance based on ARFLGB-XGBoost modeling

Guancheng Chen[1, *], Xiang Qi[1, #], Yuxin Wang[1, #] and Wenzhuo Du[2, #]

[1]School of Criminal Investigation, People's Public Security University of China, Beijing, China

[2]School of Information and Network Security, People's Public Security University of China, Beijing, China

*Corresponding author: 1600639753@qq.com

#These authors contributed equally to this work

Abstract. Extreme weather events have become a crisis for property owners and insurance companies, and insurance companies have changed the way they are willing to underwrite policies. The purpose of this report is to develop a comprehensive assessment model for the multiple factors that influence underwriting policies, in the hopes of providing the community with strategies for preserving historic landmarks in the future. In this paper, we first collected data on six climate hazard impact indicators in China and the United States after 2000, then preprocessed the nulls and outliers in the original dataset with linear and exponential fitting, and then evaluated the indicators using the CVM-CRITIC model. In order to analyze how community and demographic factors affect the application of the model, this paper uses the Spearman algorithm to analyze the correlation between community and demographic indicators and climate risk indicators, and uses the ARFLGB-XGBoost model for regression prediction in order to demonstrate the feasibility of the HICD model from different perspectives.

Keywords: XGBoost; property insurance; CVM-CRITIC; Spearman.

1. Introduction

Extreme weather events are becoming a huge challenge for real estate owners and insurance companies. More than 1,000 extreme weather events have reportedly occurred in the past few years, causing more than $1 trillion in damages globally. According to the insurance industry, natural disaster claims in 2022 are 115 percent higher than the average of the past 30 years. The situation is expected to become even more dire as losses from extreme weather events such as floods, hurricanes, cyclones, droughts and wildfires are likely to increase further. As a result, insurance premiums are rising rapidly, with climate change expected to increase premiums by 30-60% by 2040 [1].

Property insurance is not only becoming more expensive, but also more difficult to obtain, as insurers have changed how and where they write policies. Weather-related events are already driving up the cost of property insurance, and the extent of this impact depends on your location. Additionally, the insurance coverage gap averages 57% globally and is still growing. This highlights the dilemma facing the insurance industry - the profitability of insurers and the affordability of homeowners are both severely challenged [2].

2. Strategy development for insurance companies based on the CHR model

2.1. Research framework and indicators

In this section, we will discuss climate hazard property insurance, and for this purpose, we will focus on the following 2 models, with model information and indicator selection as shown in the Figure 1:

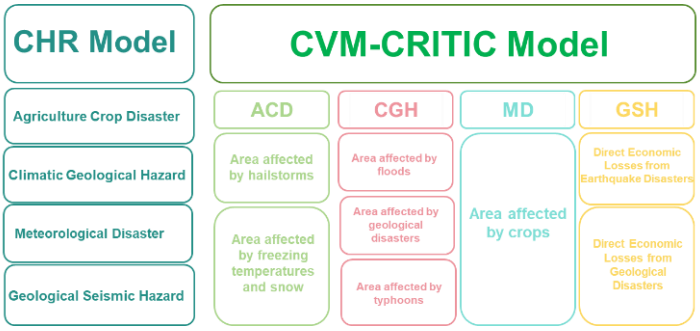

Fig. 1 Model structure for climate hazard risk assessment

In this section, we will discuss the extreme weather factors that affect property insurers' underwriting, and for this purpose we will focus on the following two models, with model information and metrics selected as shown in Figure 2:

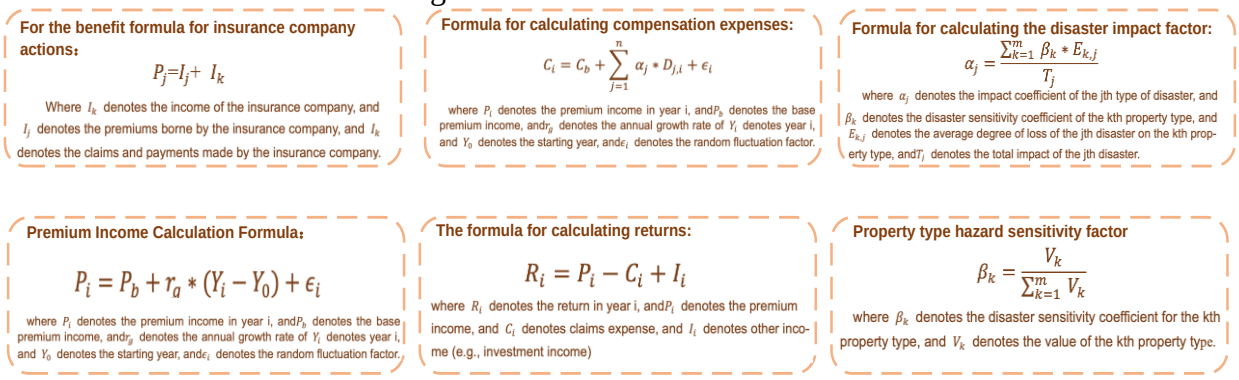

Fig. 2 Methodology for calculating CHR model indicators

We will analyze China on the Asian continent and the United States on the American continent. Each of these indicators can be roughly divided into natural and economic indicators, under which we analyze a number of secondary and tertiary indicators based on the actual selection. When calculating the impact of climate hazards on property insurance, we can define a series of formulas to quantify the different impacts [3].

2.2. CVM-CRITIC Model

We set the scoring results of the CVM model to M_{i1}, the scoring results of the CRITIC model to M_{i2} , the CVM-CRITIC model and the scoring results to M_{i3}. The combined model is solved step-by-step, the formula is as follows:

$$M_{i3} = \alpha M_{i1} + \beta M_{i2} \quad (1)$$

Empirically, α is 0.6 and β is 0.4 and the results are shown in Figure 3:

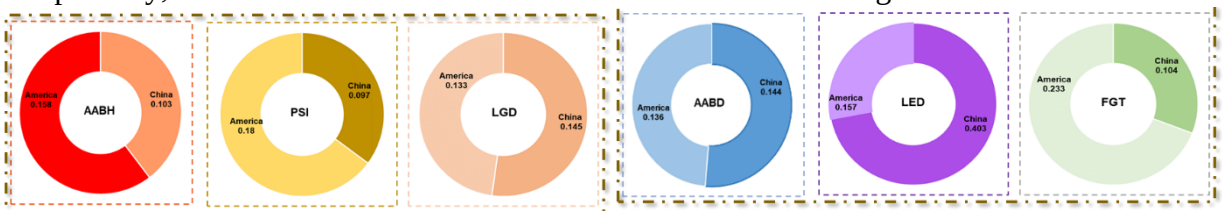

Fig. 3 Comparison of Six Indicators between China and the United States

The darker colours in the chart above represent China and the lighter colours represent the United States.ACD there is no significant difference between two countrys. In the second indicator, Climatic Geological Hazards, the United States is about twice as high as China. In the third indicator Meteorological Disaster, the US is significantly higher than China. Because the U.S. is closer to the depths of the oceans. In the fourth indicator, Geological Seismic Hazards, China's indicator is about twice as high as that of the U.S., due to that China has frequent crustal activity area.

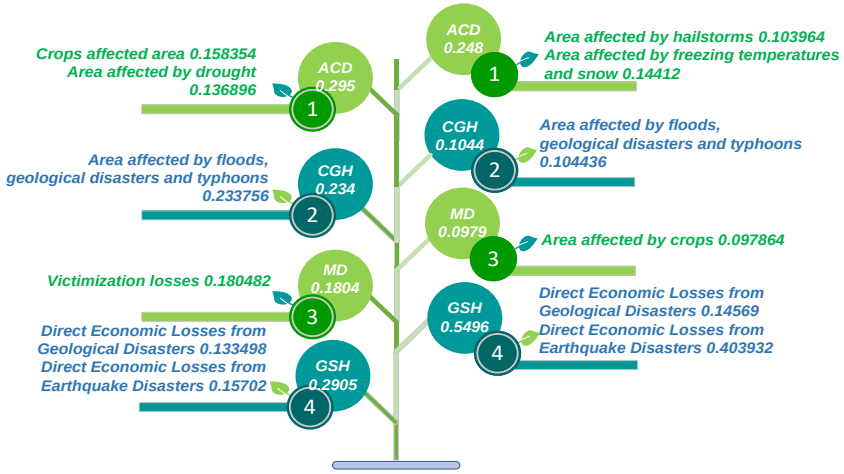

Fig. 4 CHR model indicator weights for China and the United States

In the Figure 4, China's weights are on the right and the US's are on the left. The index of hail-affected areas is similar in both countries. Floods and typhoons are similar in China and the United States, mainly because both countries are susceptible to tropical cyclones. Economic losses due to geologic disasters are also very similar in both countries. Economic losses due to earthquakes are much higher in China than in the U.S. This is because China is located in the Mediterranean-Himalayan Volcanic Seismic Belt, where the plates are susceptible to earthquakes.

2.3. Determination of the CRI index

In order to facilitate the quantification of the results of climate hazard risk, a Climate Risk Assessment Index (CRI) has been established, which is calculated using the following formula:

$$CRI = \alpha * M_{i3x1} * ACD + \beta * M_{i3x2} * CGH + \gamma * M_{i3x3} * MD + \delta * M_{i3x4} * GSH \quad (2)$$

Empirically, α is 100, β is 200, γ is 300 and δ is 400, the results are shown in Tables 1 and 2:

Table 1. China's CRI

Year	CRI	Ranking
2009	197.102	1
2013	110.848	2
2010	99.458	3
…	…	…
2022	4.182	20

Table 2. United States’ s CRI

Year	CRI	Ranking
2022	167.329	1
2018	162.624	2
2020	156.815	3
…	…	…
2006	69.245	20

2.4. FCM clustering score

The purpose of FCM clustering is to maximize the data similarity within a cluster while minimizing the data similarity between different clusters.

Applying n= 20 to the FCM algorithm for China and the U.S., the clustering results and the corresponding non-derivative policies are shown in Figure 5.

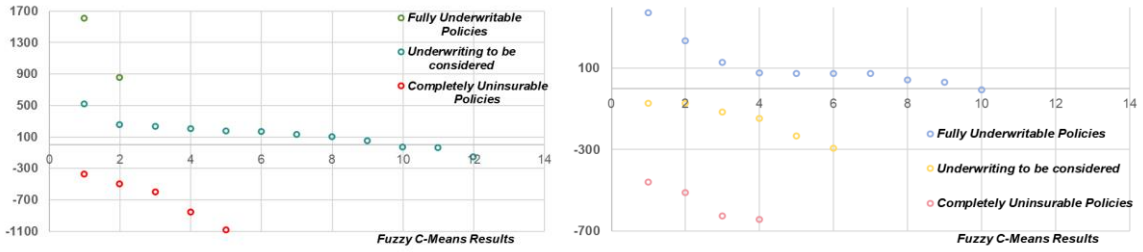

Fig. 5 Clustering results of Chinese and American CHR models

The FCM analysis allows us to derive three clusters regarding the underwriting situation. In the cluster of fully underwritten policies, U.S. insurance is basically in a profitable or break-even position, and Chinese insurance has high profits; in the cluster of underwriting to be considered, Chinese insurance is basically in a profitable or break-even position, and there are a large number of insured losses in the U.S., but due to the U.S. government's policy of insurance subsidies, it is still possible to achieve profitability. In the group of policies that cannot be underwritten at all, insurance in both countries is in a serious loss position, and it is very difficult to achieve profitability even after taking into account government subsidies.

In order to analyze whether the indicator system has a significant differentiating effect, we clustered the data using FCM. Using the elbow method, we determined that the 20 data could be categorized into three groups. We calculated the coordinates of the center points of two of the three clustering centers and used this value as the critical value. The specific formula is as follows:

$$\lambda_1 = \frac{\mu_1+\mu_2}{2},\ \lambda_2 = \frac{\mu_2+\mu_3}{2} \tag{3}$$

Where λ_1, λ_2 represents the threshold, represents the three cluster centers.

The thresholds calculated using the above methodology are shown in Table 3. It can be seen that CRIs and profits are categorized.

Table 3. Threshold of each indicator

Mark	Insurable	To be considered	Uninsurable
CCRI	(97.15,197.1]	(54.28, 97.15]	(4.18,54.28]
CP	(681.245,1603.9]	(-275, 681]	(-1083, -275]
UCRI	(132.1,167.32]	(102.63,132.1]	(69.24,102.63]
UP	(-226,370]	(-359, -226]	(-644,0.361]

Combined with the above table, we believe that Chinese insurance companies can choose to underwrite when the CRI index is greater than 97.15, and give up coverage when it is lower than 54.28, and can decide whether to underwrite the property in light of its historical and cultural heritage when it is between 54.28 and 97.15; U.S. insurance companies can underwrite when the CRI is greater than 132.1, and give up coverage when it is lower than 102.63, and can decide whether to underwrite when the CRI is greater than 132.1, and can decide whether to underwrite when the CRI is greater than 132.1 or less than 102.63. The U.S. insurance company can underwrite the property if the CRI is greater than 132.1, and will drop the property if it is lower than 102.63, and will decide whether to underwrite the property between 102.63 and 132.1 by taking into account the U.S. government's subsidy policy and the economic value of the property.

3. Impact of community and demographic factors on model application

3.1. Data collection and pre-processing

Prior to adopting our approach, we selected climate hazard and property insurance data after 2000, standardized the dataset to ensure that each variable contributed equally to the clustering process, and that clustering was based on relative distances between data points rather than absolute values of variables. Once standardization is complete, we will adjust all assessment metrics using orthogonalization techniques.

The Spearman's correlation coefficient is used to measure the nonlinear relationship between two variables and is usually used to deal with non-normally distributed data or ordered data. It is calculated as follows: where Ri and Si are the ranks of the values taken by observation i, respectively, R and S are the mean ranks of the variables x and y, respectively, and is the total number of observations.

Communities and populations are more associated with climate and geological hazards because climate impacts are positively correlated with the economy. The CRI formula at this point is as follows:

$$CRI = \alpha * M_{i3x1} * ACD + \beta * M_{i3x2} * CGH + \gamma * M_{i3x3} * MD + \delta * M_{i3x4} * GSH + \varepsilon * M_{i4x5} * Communities + \epsilon * M_{i4x6} * demographic \quad (4)$$

As a rule of thumb, the values of the coefficients here are 100, 150, 200, 250, 300 and 350.

3.2. ARFLGB-XGBoost model

The ARFLGB-XGBoost framework, as illustrated in Figure 6, is introduced in this section. These variables are fed into the Random Forest and LightGBM models. Feature engineering is employed to obtain feature importance and classification metrics individually for each model. Subsequently, an adaptive algorithm consolidates the regression metrics from Random Forest and LightGBM to generate weighted feature importance, which is then integrated into the XGBoost model. Finally, the performance of the ARFLGB-XGBoost model is assessed using a validation set [4] [5].

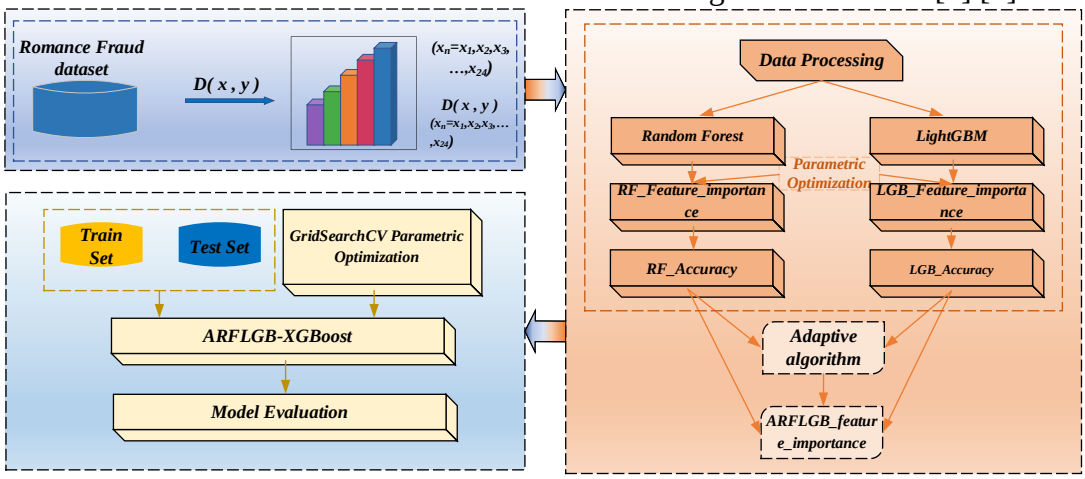

Fig. 6 ARFLGB-XGBoost design framework

Based on the above ARFLGB-XGBoost model, the property valuation indicators and CRI indexes for the next 50 years are as follows:

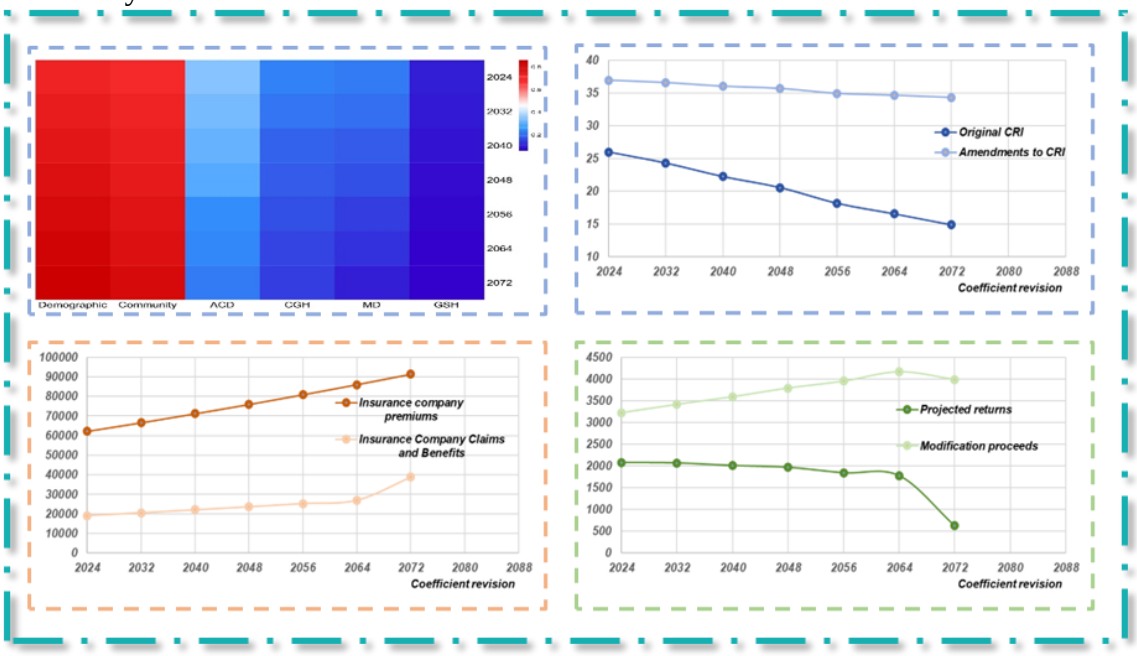

Fig. 7 ARFLGB-XGBoost design framework

As can be seen from the Figure 7, the corrected CRI is lower than the original CRI, the corrected gain is higher than the original gain, and as the economy develops, the climatic impact diminishes, but the geological impact is basically unchanged.

4. Summary

The indicators selected in this paper are more comprehensive, covering various aspects such as history, culture, community, population, climate and ecology. , considering a wide range of scenarios, focusing on the comprehensiveness and high accuracy of the indicators.The ARFLGB-XGBoost model can be used to predict the future trend of the CRI by simply utilizing the evolution of the historical state of the stock itself, which is a simple and reliable approach compared to traditional mathematical and statistical methods.

In this paper, we use the CVM-CRITIC model to evaluate the indicators, propose that the CRI refers to climate hazards, and utilize the FCM clustering algorithm to classify the underwriting policies of China and the U.S. into three categories, and find that when the CRI of China and the U.S. is more than 97.0%, the underwriting policies of China and the U.S. are fully insurable. When the CRIs of China and the U.S. exceed 97.15 and 132.1, respectively, they are found to be fully insurable; when the CRIs are lower than 54.28 and 102.63, they are found to be exempted from insurability.

For HICD model II, we first use Spearman's algorithm to solve the correlation between community and demographic indicators and climate hazard indicators, and use the ARFLGB-XGBoost model to make regression predictions to demonstrate the feasibility of the HICD model from different perspectives.

References

[1] Liu, Y., Wang, B., & Lv, S. (2014). Using Multi-class AdaBoost Tree for Prediction Frequency of Auto Insurance. Journal of Applied Finance and Banking, 4, 1-4.

[2] Barthel, F., & Neumayer, E. (2011). A trend analysis of normalized insured damage from natural disasters. Climatic Change, 113, 215-237.

[3] Willoughby, H. (2012). Distributions and Trends of Death and Destruction from Hurricanes in the United States, 1900-2008. Natural Hazards Review, 13, 57-64.

[4] Li, Yude. (2022). The Application of Big Data in the Insurance Industry-with Potential Risks and Possible Solutions. 10.2991/aebmr.k.220603.069.

[5] Meng Shengwang.The Prediction of Automobile Insurance Claim Probability and Aggregated Losses Based on Machine Learning Algorithm[J]. Insurance Studies, 2017.