Development of the best investment strategy for insurance companies based on ARIMA-SVM

Yanxin Long [*, #], Jingyi Wang [#], Ke Dong [#]

Mathematical &Physical Sciences, Wuhan Textile University School, Wuhan, China, 430200

* Corresponding author: 19839399622@163.com

[#]These authors contributed equally.

Abstract. The purpose of this article is to develop the best investment strategy for insurance companies in the context of extreme weather, aiming to improve the profitability of insurance companies and reduce investment risks. This research focuses on Asia and North America, and constructs an ARIMA model to accurately predict the frequency of extreme weather in the next decade. At the same time, a comprehensive evaluation model was constructed by combining the entropy weight method to quantitatively evaluate the payment ability of local residents. On this basis, the SVM classification model was established to predict whether the insurance company should be insured in a certain place, and the performance of the model was deeply analyzed, and the recall rate and accuracy of the classification model were 0.93 and 0.94, respectively, indicating that the model had good reliability. After a series of rigorous model analysis and data validation, it was finally concluded that China in Asia is the best choice for insurance companies. This strategy not only helps insurers optimize resource allocation, but also achieves stable and sustainable development in the context of complex and volatile extreme weather.

Keywords: Extreme Weather, Insurance Investment, ARIMA, Entropy Weight Method, SVM.

1. Introduction

In recent years, extreme weather has caused more than $1 trillion in losses worldwide. Looking at the data for 2022 alone, natural disaster claims in the global insurance industry have increased by 54% compared with the average of the last 10 years, and by 115% compared with the average of the last 30 years [1]. At the same time, insurance costs are increasing. It is expected that by 2040, in the face of increasingly severe extreme weather, weather-affected insurance premiums will increase by 30-60% This impact involves the entire process from original insurance, reinsurance to insurance-linked securities. The sharp increase in catastrophe risks has brought challenges to these levels[2].Thinking from an insurance company's perspective, when a risk can be taken on, when a policy is covered, and whether the area is insured is a tricky question.

Our team studied and consulted a large number of literatures, and found that currently scholars have few studies on the combination of the evaluation results of whether insurance companies insure in the local area under the background of frequent extreme weather[3-5].Most of the researches focus on how to reduce losses with the help of extreme weather insurance, but do not give the circumstances under which insurance companies should choose to avoid risks.

SVM model is a method with good classification effect and strong generalization ability[6-7]. Asking the profits of insurance companies as the entry point, this paper selected the occurrence of various extreme weather (such as hurricanes, sandstorms, and wildfires) in 51 countries in Asia and North America in the past 10 years to conduct a comprehensive assessment of the affordability of local residents. Based on this, the SVM model of insurance company is established, and the decision is given to reduce the risk of insurance in the local insurance.

2. Establishment of the best investment strategy model for insurance companies in extreme weather

The data are from the following websites, as detailed in Table 1:

Table 1. Data source

data source	Data Websites
data source 1	https: //ourworldindata.org/natural-disasters
data source 2	https: //www.emdat.be/
data source 3	https: //www.gddat.cn/newGlobalWeb/#/home
data source 4	http: //www.tjnjw.com/niandu/diqunianjian-2023.html

2.1. ARIMA prediction model for extreme weather

Our team collected detailed information on the occurrence of various extreme weather events in Asia and North America from 1980 to 2023. Statistics on the number and frequency of extreme weather occurrences in different years. Use ARIMA's time series prediction model to predict the number and frequency of extreme weather occurrences in the region in the next ten years.

2.1.1. Establishment of the ARIMA prediction model

The basic idea of ARIMA is to treat the data sequence formed by predictions that change over time as a random sequence and use a model to approximately describe this sequence. The basic principle is to convert non-stationary time series models into stationary time series. Then regress the dependent variable only on its lagged value, the present value and lagged value of the random error term. Models can predict future values from historical data in a time series. In this model, we selected the number of extreme weather occurrences in each year as a random sequence. Based on this, the predicted number of extreme weather occurrences can be obtained.

ARIMA (p, d, q) is formed by combining the autoregressive model (AR), the moving average model (MA) and the difference method. The AR model is used to describe the relationship between current values and historical values, and MA represents the accumulation of error terms in the autoregressive model. The difference method is used to stabilize non-stationary time series. p represents the number of lag terms in the autoregressive part, d represents the number of differences, and q represents the number of lag terms in the moving average part. The formula of the ARIME model can be expressed as

$$\hat{p}^{(t)} = p_0 + \sum_{j=1}^{p} \gamma_j p^{\{t-j\}} + \sum_{j=1}^{q} \theta_j \varepsilon^{\{t-j\}} \tag{1}$$

Among them, p is the order of the autoregressive model (AR), q is the order of the MA model, $\varepsilon^{\{t-j\}}$is the error term between time λ and λ− 1, $\gamma_j$and θ_j are the fitting coefficients, andp_0 is Constant term.

Observing the collected historical extreme weather event data, we found that it conforms to the characteristics of a random walk, that is ARIMA (0, 1, 0).

It is a special form of the ARIMA model, often also called the "random walk model". The characteristic of this model is that it contains only a first-order difference with a unit root, no autoregressive part, and no moving average part.

In this model:

The parameter (p) of the AR part is 0, which means that there is no direct relationship between the current value and the past value;

The parameter (d) of the difference part is 1, which means that we only need to differentiate the original data once to make it a stationary sequence;

The parameter (q) of the MA part is 0, which means that the forecast error is not affected by past forecast errors.

2.1.2. Prediction results in the ARIMA model

Based on the above theory, we apply ARIMA to predict the frequency of extreme weather in China and the United States in the next 10 years. To meet the requirements of this problem, we used data from 2000 to 2023 to train our model and predict the number of extreme weather occurrences in the next ten years. Take the forecast results in China and the United States as examples, as shown in Figure 1 below:

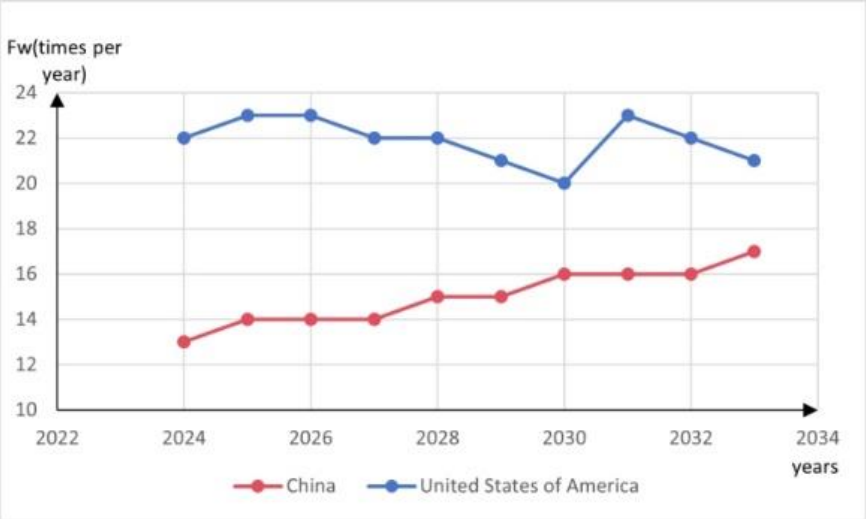


Figure 1. The frequency of extreme weather events in China and the United States in the next 10 years

2.2. Establishment of a comprehensive evaluation model based on local residents' paying ability

In order to achieve reasonable quantification of local policy sales volume, this article selects 5 indicators from the indicator system to measure the paying ability of local residents: population density, energy usage, energy loss, education expenditure, and greenhouse gas emissions. The evaluation system formed with the help of Entropy Weight method was reconstructed and analyzed.

These five indicators are important reference factors for measuring the paying ability of local residents. The following is the specific relationship between them and the paying ability of local residents.

Among the various indicators that affect residents' affordability, there may be high correlations between some indicators. Through the correlation between indicators, we can intuitively understand the connection between these indicators, and specifically select those indicators that are most relevant to residents' paying ability for model training, improving the explanatory power and prediction ability of the model. Therefore, we tested the Spearman correlation coefficient for five indicators.

In the following, we use the entropy weight method to establish a comprehensive evaluation model for the five indicators.

2.2.1. Data preprocessing

The various evaluation indicators in this article are incommensurable due to differences in their measurement units and orders of magnitude, which brings difficulties and problems to determining comprehensive evaluation indicators. Before starting, we ensure that all indicator data are on the same scale to eliminate the impact of differences in the original indicator data.

This article uses Z-score standardization to process each column to obtain new values:

$$x_{ij}{}' = \frac{x_{ij} - mean(x_j)}{std(x_j)} \tag{2}$$

In the formula, i represents the ordinal number of different regions, and j represents the ordinal number of different indicators. In this way, raw data of different scales can be normalized to data between 0 and 1. Assume that after normalization, the data matrix $R = (r_{ij})_{m\times n}$ can be formed

2.2.2. Calculate indicator entropy value

Based on the concept of information entropy, we calculate the entropy value of each indicator. The greater the entropy value, the greater the uncertainty of the information, and the importance of this indicator is relatively low. On the contrary, the smaller the entropy value, the lower the uncertainty of the information, and the higher the importance of the indicator.

First, we calculate the proportion p_{ij} of each country in the j-th indicator, expressed as:

$$p_{ij} = \frac{r_{ij}}{\sum_{i=1}^{m} r_{ij}} \tag{3}$$

- i represents the serial number of 11 countries
- j represents the sequence number of the five indicators
- r_{ij} represents the value of the corresponding indicator

Then, we calculate the entropy value E_j of the jth indicator as follows:

$$E_j = -\frac{1}{\ln n_0}\sum_{i=1}^{m} p_{ij}\ln p_{ij} \tag{4}$$

Finally, the weight of each indicator is calculated based on its degree of difference, that is, the ratio of the degree of difference to the sum of all degrees of difference. The greater the weight value, the greater the influence of the indicator in the entire evaluation. The greater the weight value, the greater the influence of the indicator in the entire evaluation.

Weight calculation formula:

$$w_j = \frac{(1-E_j)}{\sum_{j-1}^{n}(1-E_j)} \tag{5}$$

By calculating the corresponding weights of the five indicators: population density (0.3393), energy usage (0.1822), energy loss (0.0428), education expenditure (0.0407), and greenhouse gas emissions (0.3950), as shown in Figure 2 below:

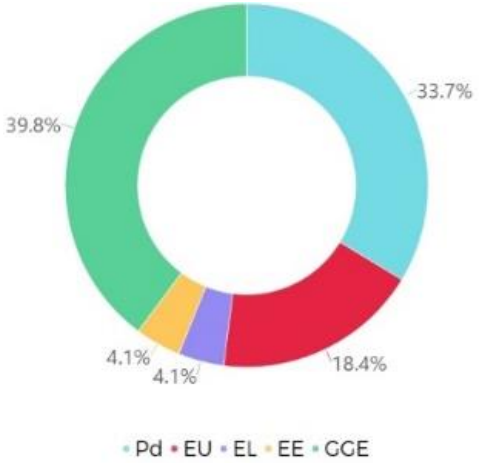


Figure 2. Weight Chart for Metrics

2.2.3. Comprehensive evaluation system of local residents' paying ability

Through the obtained weights, the formula for calculating the score of local residents' affordability is as follows:

$$index_{ij} = weight_1 \times Pd_{ij} + weight_2 \times EU_{ij} + weight_3 \times EL_{ij} + weight_4 \times EE_{ij} + weight_5 \times GGE_{ij} \tag{6}$$

Here, i is the country, j is the year, and $index_{ij}$ is the score of the i-th country under the j-th indicator.

3. A two-class model based on the SVM model

3.1. Quantification of insurance company profits considering the time value of costs

Whether an insurance company is suitable to insure in this region, the core of the question is whether the insurance company's profits can maintain the current operating status of the insurance company, that is, whether the insurance company's downside risk exposure is less than the current insurance company's profit.

An insurance company's profit is usually expressed as:

$$W = (1+k)^N \times P_0 \times P_s - P_c \times P_s \times C_A \tag{7}$$

In the formula, P represents the profit of the insurance company in a certain year, P_0 represents the policy pricing of the insurance company in a certain year, P_s represents the policy sales volume of the insurance company in a certain year, P_c represents the probability that the insurance company may pay compensation in the current year, which in this article is equivalent to the frequency of extreme weather, C_A the amount of compensation.

For an insurance company to operate normally, the premium must be at least equal to the total loss, that is

$$P_0 \geq C_A \times P_c \tag{8}$$

It is worth noting that the effectiveness of insurance often does not take effect immediately, and claims are likely to occur many years later. Therefore, we must consider the time value of money. 20 million in 10 years is not the same as 20 million now. Therefore, we need to calculate the value of 20 million discounted in 10 years to today. As for the problem studied in this article, it is necessary to calculate the predicted amount of compensation discounted to the current amount, and obtain the policy pricing based on the above formula.

Here, this article sets the insurance company's annual interest rate to 2.5%, based on Excel's discount formula:

$$P_v(2.5\%, N, C_A) = C_A \times \frac{1}{(1+2.5\%)^N} \tag{9}$$

In the formula, P_v represents the compensation amount discounted to the current corresponding value, that is, the present value. N represents the past time, and C_A represents the end-of-period cash flow, that is, future money, which is the compensation amount in this article.

In order to realize the decision-making ability of the two-classification model, this article chooses to take the profits of insurance companies in 2024 as an example. P_0 and P_s respectively represent

the insurance company's policy pricing and policy sales volume in 2024. C_A represents the compensation amount in 2024.

Combining equations (8) and (9), the policy pricing P_0 can be calculated. (8) (9) The probability of possible compensation Pc appearing in the two equations corresponds to the frequency of extreme weather when placed in the problem studied in this article. What needs to be considered when calculating the probability of extreme weather occurrence in each year is the total amount of data and the total number of bases during the analysis period, which respectively refer to the number of extreme weather occurrences in a future year and the number of days in the corresponding year mentioned in this article. Accordingly, the quantification of **Pc** can be achieved.

It can be seen from equation (8)(9) that based on the C_A set previously, whenC_A is fixed, the value of P_0 is determined. Policy sales are affected by many factors. This article only considers the paying ability of local residents. Observing and analyzing equation (7), if the average paying ability of local residents in the same area remains unchanged, P_S can be determined. From this, we can conclude that in the same area, P is negatively correlated with P_c . The question of whether an insurance company insures locally can also be converted into a problem of classifying the intervals between two different types of data: P_c and P_S.

In short, it is not difficult to find that the key to solving the problem lies in quantifying the difference between the frequency of extreme weather and the number of policy sales. From this, our team introduced the SVM model. Based on the essential characteristics of the SVM model, a hyperplane is found in the high-dimensional space so that this hyperplane can separate samples of different categories to the greatest extent. This hyperplane can find a balance between classification accuracy and maximization of the interval, thus Achieve the best classification effect.

3.2. Solution of SVM

Support vector machine (SVM) is a supervised learning algorithm mainly used for classification or regression problems. In classification problems, such as deciding whether an insurance company insures in a specific area, SVM classifies different samples by finding the optimal separation hyperplane (also known as the decision boundary) in high-dimensional space. In this article's insurance question, the data are often not linearly separable. To solve this problem, SVM introduces the kernel function. The function of the kernel function is to map (transform) the original feature space to a higher-dimensional space so that the data becomes linearly separable in this new space.

Step1: For the feasibility of machine learning, this article collected data on the number of extreme weather occurrences from 1980 to 2023 for a total of 51 countries in North America and Asia, and used the ARIMA model to predict the number of extreme weather occurrences in each region after 10 years, as shown in Figure 3 below:

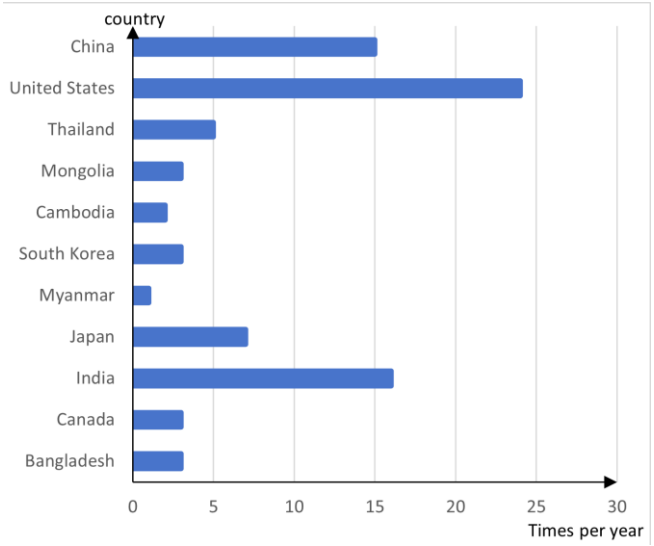


Figure 3. Projected Number of Extreme Weather Events in 11 Regions Over the Next Decade

Step2: The SVM model will output a decision as to whether the insurance company should insure in this area. SVM can regard the information of each region, such as risk level, compensation cost, historical data, etc., as features, and then input these features into the SVM model. We choose policy sales and the probability of extreme weather as a feature. Then set the classification labels. In this problem, the category labels are "insured" and "not insured", which are the label sets used in the SVM model.

Step3: Use the obtained features and labels to train the SVM model, and finally use the trained SVM model to distinguish new or unknown regional data and predict whether insurance companies should insure.

This article is a two-class classification model for deciding whether to insure or not based on SVM. Linear is used as the kernel function, which only performs linear transformation and does not change the dimensional space of the feature vector. The features are scaled before calculating the kernel function, and the final debugging model draws conclusions by predicting the insurance company's profits. The prediction results are shown in the following table 2:

Table 2. Predictions on whether to buy insurance in China and the United States

forecast result	area	affordability score	Probability of extreme weather
YES	China	100	15
NO	America	48.5488	24

Bringing the relevant feature data from China and the United States selected by our team into the SVM model, and considering the interests of insurance companies, we finally get:

For the China, the probability of extreme weather events in China is moderate, and the residents' ability to pay scores the highest among the selected regions. It is predicted that the profit of insurance companies here is more than 0, so insurance companies are suitable to provide insurance here.

For the United States, the probability of extreme weather is 60% higher than that of China, and the score of residents' affordability is 51.45% lower than that of China. It is predicted that the profits of insurance companies here are <0, and insurance companies are not suitable to insure here.

3.3. Model testing and evaluation:

We then make statistics on the accuracy, recall, and precision of the SVM model prediction results, as shown in the following table 3:

Table 3. Verification and evaluation of SVM model

	Accuracy	Recall	Precision	F1
Training set	0.917	0.917	0.93	0.917
cross validation set	0.625	0.625	0.623	0.599
test set	0.938	0.938	0.943	0.935

As can be seen from the above table, the test accuracy rate is 94.3%. It can be seen that the SVM model has good classification ability for the two-classification problem of whether to be insured or not.

4. Conclusion

We used ARIMA's time series forecasting model to predict the number and frequency of extreme weather in China and the United States in the next decade. At the same time, we selected five indicators and established a comprehensive evaluation model based on the local residents' ability to pay. Then, a dichotomous model based on the SVM model is established to determine whether the insurance company is insured, and under the consideration of the time value of cost, we use the interest rate, profit, premium, policy pricing and claim amount of the insurance company to determine the policy sales volume, and use the obtained data to train the SVM model to realize the reasonable quantification of the difference between the frequency of extreme weather and the policy sales volume.

In this paper, when the ARIMA model is used to predict extreme weather prediction in the next ten years, the actual situation and volatility of the data are not considered, the performance is limited, and the effect of dealing with complex nonlinear relationships is poor. Since the entropy weight method has too high data requirements, cannot deal with the correlation between indicators, and is sensitive to small changes in the data, which may be greatly affected by data errors, we can use the subjective and objective method to accurately evaluate the local residents' ability to pay, and then improve the binary classification model of whether insurance companies insure a certain place.

References

[1] Fantini, L., Blanchard, B., Rath, S., Removille, P., Schwemer, S., & Mayeres, F. (2023, December 4). An insurance risk framework for climate adaptation. BCG Global. https: //www.bcg.com/publications/2023/an-insurance-risk-framework-for-climate-adaptation.

[2] "Challenges Posed by Climate Anomalies to the Insurance Industry_"+ Bank of China Insurance Report Web. (n.d.). http: //www.cbimc.cn/content/2022-09/16/content_467963.html.

[3] Yao Peichang. Optimal allocation of investment assets of insurance companies: A case study based on Ping An of China [D]. Zhejiang University, 2016.

[4] Xia Qi. Application analysis of big data technology in investment and risk management of insurance companies [J]. Marketing Circle, 2022 (5): 167 - 169.)

[5] Duan Shenghui. Insurance Theory & Practice, 2023 (4): 1 - 31.

[6] Luo Xiangyang, Li Ruixiang, Sun Yuchen, et al. A Method for Obtaining Street-level Landmarks Based on SVM Classification Model.CN201910128483.6 [2024 - 03 - 10].

[7] Xu Jie. Research on the effectiveness of industrial investment based on SVM model: A case study of Chongqing [D]. Chongqing Technology and Business University, 2022.

[8] Wang Qiubing, Wu Jiaqian, Bian Zhenxing. Evaluation of village human settlements based on entropy weight method from the perspective of rural revitalization [J]. Agricultural Economics, 2020 (3): 2.)

[9] Wang Yan, Zhou Kai, Shen Shoufeng. Mining and cluster analysis of big data of educational affairs based on entropy weight method [J]. Journal of Zhejiang University of Technology, 2023, 51 (1): 84 - 87.)

[10] Guo Wenjing, Zhao Chengguo, Yuan Jianhui. Optimal Insurance Investment Decision in Leapfrog Diffusion Market [J]. 2021 (2011-4): 749 - 760.