Research on the Sustainability of Property Insurance under the Background of Extreme Weather Events

Siyuan Wu[1, *, #], Ran Wang[2, #], Jiawei Hu[2, #]

[1]Department of Electronics&Information Engineering, Tiangong University, Tianjin, China, 300387

[2]School of Electronics&Information Engineering, Tiangong University, Tianjin, China, 300387

*Corresponding author: yueyingmimeng@gmail.com

#These authors contributed equally.

Abstract. In recent years, the frequency of extreme weather events has significantly increased, resulting in global losses exceeding $1 trillion. The profit crisis of insurance companies and the issue of homeowners' payment ability have become more apparent. This article focuses on the sustainability of property insurance under extreme weather events, establishes a comprehensive insurance and protection model, and provides recommendations for relevant departments.Firstly, a set of seven indicators including the proportion of household insurance expenditure was selected to establish an indicator system. After conducting Spearman correlation coefficient testing on the data and combining it with LSTM neural network to predict the frequency of extreme weather in the next decade, this paper uses the SVM binary classification model to classify high-risk areas as uninsurable areas and other areas as insurable areas. This decision-making process resulted in insurance companies covering areas such as Alabama and refusing to underwrite insurance in Arizona, while also making relevant recommendations.Secondly, by calculating indicators for communities and real estate developers, and considering the impact of population growth on housing demand, this article uses the ARIMA model to predict population growth in different regions over the next five years. After standardizing the data, the entropy weighting method is applied to consider the subjective importance of different real estate developers for different indicators and assign weights to the indicators. Weighted indicator data is used as input for the SVM model, dividing regions into two categories, with regions with better overall conditions being more favorable for real estate development. This has led people to prioritize the development of regions such as North Carolina and Indiana. In addition, this article introduces the TOPSIS regional ranking comprehensive evaluation index, which is beneficial for the development of higher ranked regions. Similar to our SVM classification results, Indiana and North Carolina rank second and third.

Keywords: ARIMA, TOPSIS, SVM Model, LSTM Neural Network.

1. Introduction

In recent years, extreme weather events have become increasingly frequent worldwide, resulting in global losses exceeding 1 trillion US dollars[1]. In 2022, due to natural disasters, insurance claims in the industry increased by 115% compared to the average level of the previous 30 years[2]. Anticipated losses from extreme weather events such as floods, hurricanes, cyclones, droughts, and wildfires are expected to further rise, leading to a rapid increase in insurance costs. By 2040, premiums triggered by climate change are estimated to increase by 30-60%[3][4].

Property insurance has not only become more expensive but also harder to obtain, as insurance companies alter their willingness to underwrite policies in terms of manner and location[5]. Weather-related events have driven up the costs of property insurance premiums, varying across different parts of the world[6]. Additionally, the insurance protection gap has reached an average of 57% globally and is on the rise. This highlights the challenges facing the insurance industry-a crisis in profit for insurance companies and concerns about the payment capabilities of property owners.

2. Establishment of the Evaluation System

This article develops a model to help insurance companies determine whether to insure policies in areas prone to extreme weather events. The model was demonstrated using two regions located on different continents that have experienced extreme weather events. For this issue, insurance companies have two types of decisions regarding specific regions: underwriting and non underwriting. 1 and 0 can be used to represent non underwriting and underwriting, respectively. The profits of insurance companies and the payment ability of owners can be reasonably quantified. Use these values as inputs and use the SVM binary classification model to find a solution.

By referring to relevant literature [7][8], the paper have selected indicators with higher occurrence frequencies as the criteria for evaluating whether insurance companies should underwrite policies in this study. The final established indicator system is as follows Fig 1:

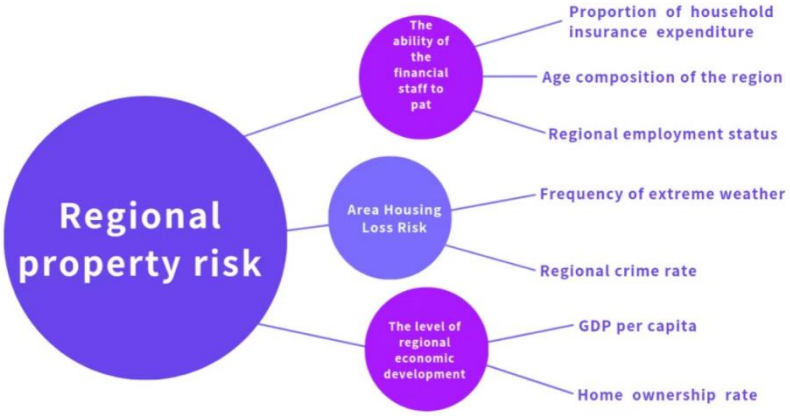


Fig 1 Indicator Evaluation System

Based on the indicator evaluation system we established, we collected relevant data for 51 states in the United States, including ALABAMA and ARIZONA, from 2010 to 2023. A partial display of the data is shown below Table 1:

Table 1: Indicator Evaluation System Data Table

STATE	The number of extreme weather events in the past decade	Per capita insurance expenditure(%)	Age composition (median age)(%)
ALABAMA	24	0.077207591	39.6
ALASKA	15	0.039891715	35.9
AMERICAN SAMOA	9	0.03240394	38.8
ARIZONA	18	0.043653424	38.9
ARKANSAS	22	0.009225336	37.9
ATLANTIC NORTH	20	0.049617702	37.7
ATLANTIC SOUTH	25	0.069761253	40.9
CALIFORNIA	30	0.075923758	41.5
COLORADO	17	0.001741663	34.9
CONNECTICUT	14	0.011844197	42.7

2.1. Correlation Analysis

Commonly used methods for correlation analysis include Pearson correlation coefficient and Spearman correlation coefficient. The distinction between them lies in whether the data satisfies a normal distribution. When the data satisfies a normal distribution, using the Pearson correlation coefficient can achieve higher accuracy. Therefore, the paper first conducted a normality test on the data in the Table 2.

Below are the results of the normality test:

Table 2: Normality Test Results

Variable Name	Shapiro-Wilk Test	Smirnov Test
The number of extreme weather events in the past decade	0.964(0.070*)	0.08(0.803)
Age composition (median age)(%)	0.963(0.059*)	0.112(0.397)
GDP per capita (US $10,000)	0.883(0.000***)	0.081(0.790)
Regional crime rate (%)	0.622(0.000***)	0.256(0.001***)
Employment status (employment rate)(%)	0.84(0.000***)	0.163(0.069*)
Per capita insurance expenditure(%)	0.848(0.000***)	0.175(0.042**)
Home ownership rate (%)	0.781(0.000***)	0.224(0.004***)

Tip: ***, **, * represent significance levels of 1%, 5%, and 10%, respectively.

In general, if it shows significance (P<0.05), it means rejecting the null hypothesis (data follows a normal distribution). Based on the results in the table, some of the data does not follow a normal distribution. Therefore, we use the Spearman correlation coefficient, which does not require the data to be normally distributed, for the correlation test.

The Spearman correlation coefficient[9]is defined as the Pearson correlation coefficient between ranked variables. For a sample with a capacity of n, the original data is transformed into ranked data, and the correlation coefficient.

$$\rho = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}} \quad (1)$$

For and is calculated based on their average descending positions in the overall data when arranged in descending order. They are assigned corresponding ranks, as shown in the Table 3 below:

Table 3: Ranking Table

Variables X_i	Descending Positions	Level X_i
0.8	5	5
1.2	4	3.5
1.2	3	3.5
2.3	2	2
18	1	1

Whereas the linkage between variables is not crucial, ρ can be calculated through a simple step. The observed difference in ranks between two variables and ρ is then:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \quad (2)$$

Where the denominator is a constant related to the length of the variable sequence, and the numerator is the sum of the sequence errors, reflecting the correlation differences between variables. ρ takes values in the range of [-1, 1], reflecting the direction and degree of the trend between two variables. A value of 0 indicates no correlation between variables, a negative value indicates a negative correlation, and a positive value indicates a positive correlation. The larger the absolute

value of ρ, the stronger the correlation between two variables. If the absolute value is less than 0.5, it indicates no correlation.

Based on the data, using MATLAB software for calculation and substituting the seven indicators, where 1-7 represent The number of extreme weather events in the past decade, Age composition (median age)(%), GDP per capita (US $10,000), Regional crime rate(%), Employment status (employment rate)(%), Per capita insurance expenditure(%), Home ownership rate (%), the paper obtained the following correlation coefficient table 4:

Table 4: Spearman Correlation Coefficients

	1	2	3	4	5	6	7
1	1	0.129	0.098	-0.247	0.161	-0.261	0.248
2	0.129	1	0.083	-0.164	-0.104	-0.01	-0.064
3	0.098	0.083	1	0.061	0.06	0.101	-0.007
4	-0.247	-0.164	0.061	1	-0.072	-0.044	0.144
5	0.161	-0.104	0.06	-0.072	1	0.141	-0.071
6	-0.261	-0.01	0.101	-0.044	0.141	1	-0.086
7	0.248	-0.064	-0.007	0.144	-0.071	-0.086	1

From this table, we can observe that there is virtually no correlation among the seven indicators we selected. This indicates that our chosen indicators are independent of each other, reflecting a comprehensive and complete representation of the situation. Further research can be conducted with confidence in the reliability of the selected indicators.

2.2. Establishment of the LSTM Neural Network Prediction Model

The paper need to predict the frequency of extreme weather occurrences in the next ten years based on historical data of extreme weather in various regions of the United States to support insurance company decision-making. LSTM is a variant of recurrent neural networks[10], specifically designed to handle and learn long-term dependencies in sequential data. In the LSTM model, specific memory units replace the hidden layer nodes in the chain structure of RNN, providing greater adaptability in the data inference and analysis process. It is suitable for analyzing and predicting the frequency of future long-term extreme weather occurrences.

LSTM's innovation lies in the introduction of a state for the information set. If the state is not modified, it will continue to be propagated backward. However, memories always change, so some modules for correcting memories need to be added, introducing the concept of 'gates.' The role of 'gates' is to filter information, allowing some information to pass through while filtering out others. It consists of several parts:

(1) Forget Gate: Decides which information in the previous time step's memory cell needs to be forgotten. Also using the Sigmoid function, it determines which information should be retained. The forget ratio function is given by:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{3}$$

(2) Input Gate: Determines the degree to which new information flows in. Using the Sigmoid function, it decides which information should be updated. The function expression for adding new information is:

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t} \tag{4}$$

(3) Output Gate: Decides the output at the current time step. By using the Sigmoid function and the Tanh function, it controls the output of the memory cell.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t * \tanh(C_t) \tag{6}$$

2.3. Solution of the LSTM Neural Network Prediction Model

The paper utilized MATLAB software, using historical data from various states in the United States as input. We specified the LSTM layer with 200 hidden units, selected 70% of the data for training, and reserved 30% for testing to evaluate the model's performance. The paper set the gradient threshold to 1, specified the initial learning rate as 0.005, and decreased the learning rate by multiplying it by a factor of 0.2 after 125 rounds of training. The model predicts the number of extreme weather events in the United States for the next 10 years across different regions, as shown in the following Fig 2:

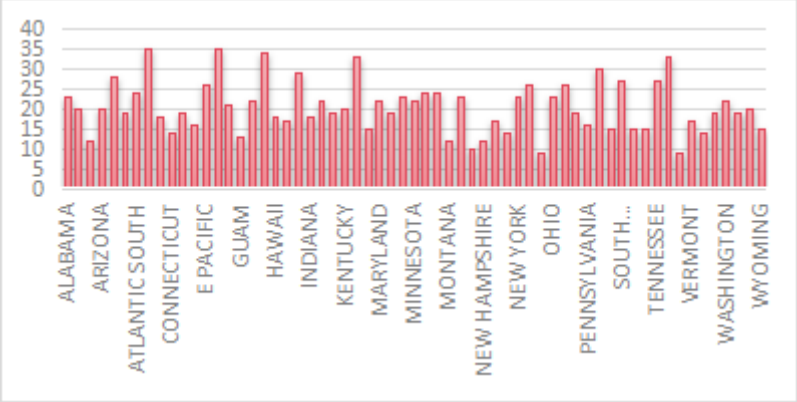


Fig 2 Future Extreme Weather Events in the United States for the Next Ten Years

Upon verification, our neural network simulation accuracy reached 89.3%. The model exhibits a good fit.

2.4. SVM Binary Classification Model

Support Vector Machine (SVM) is a machine learning method developed based on the principle of minimizing structural risk in the theory of statistical learning. The fundamental implementation of SVM involves constructing an optimal classification hyperplane in the original space or a high-dimensional space after projection. This hyperplane is then used as the classification decision boundary for data classification,as shown in Fig 3.

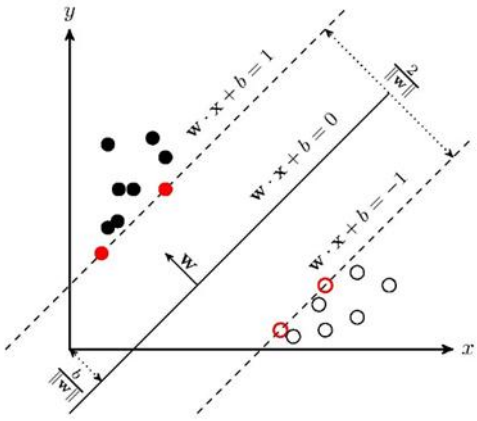


Fig 3 Binary classification model diagram

The basic theory of SVM is designed to address binary classification problems. The primary goal of SVM is to construct a linear optimal classification hyperplane that perfectly separates two classes of samples and maximizes the margin between them. For a given sample set , when the sample set is linearly separable, the general form of the linear discriminant function is: $g(x) = (w^T x + b)$.Normalizing the discriminant function, such that the closest sample to the classification hyperplane satisfies , makes the margin equal to . Maximizing the margin is equivalent to minimizing. Ensuring the classification hyperplane correctly classifies all samples is expressed as:$|g(x)| = 1$.Then, the

margin is equal to $2/(||w||)$ .Maximizing the margin is equivalent to minimizing $||\mathrm{w}||^2$ .The requirement that the classification hyperplane correctly classifies all samples is expressed as:

$$y_i(w \cdot x_i + b) \geq 1 \tag{7}$$

In summary, the problem of finding the optimal classification hyperplane is equivalent to solving:

$$\min\Phi(\mathrm{x}) = \frac{1}{2}\|w\|^2 \tag{8}$$

The solution to this equation can be obtained by defining the Lagrange function:

$$\mathcal{L}(w, \mathrm{b}, \mathrm{a}) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} a_i \{[(x_i \cdot w) + b]y_i - 1\} \tag{9}$$

Where ai> 0 is the Lagrange multiplier, then the problem is transformed into finding the minimum value of the Lagrange function with respect to w and b. And make

$$\frac{\partial L}{\partial w} = 0,\ \frac{\partial L}{\partial b} = 0 \tag{10}$$

So

$$\mathrm{w} = \sum_{i=1}^{N} a_i\, y_i x_i \tag{11}$$

$$\sum_{i=1}^{N} a_i\, y_i = 0 \tag{12}$$

Substituting the above equation, we get:

$$\mathcal{L}(\mathrm{w}, b, \mathrm{a}) = \frac{1}{2}\sum_{i,j=1}^{N} a_i\, a_j y_i y_j x_i x_j - \sum_{i,j=1}^{N} a_i\, a_j y_i y_j x_i x_j - b\sum_{i=1}^{N} a_i\, y_i + \sum_{i=1}^{N} a_i \tag{13}$$

$$\mathcal{L}(\mathrm{w}, b, \mathrm{a}) = \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i,j=1}^{N} a_i\, a_j y_i y_j x_i x_j \tag{14}$$

In this case, as the training set samples are linearly inseparable, we need to introduce a penalty factor C and a slack variable ξ_i. The problem of finding the optimal classification hyperplane can be described as:

$$\begin{cases} \min\limits_{w,b,\xi} \frac{1}{2}\|w\|^2 + \mathrm{C}\sum_{i=1}^{N} \xi_i \\ y_i(w \cdot x_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0, \mathrm{i} = 1,2,3 \ldots \mathrm{N} \end{cases} \tag{15}$$

The corresponding optimal classification function is:

$$f(x) = sgn(\sum_{i=1}^{N} a_i{}^* y_i k(x_i, x) + b^*) \tag{16}$$

When dealing with nonlinearly separable data, employing a kernel function is a common strategy to map the data from a lower-dimensional to a higher-dimensional space, enabling linear separability in the higher-dimensional space.

For solving classification problems with nonlinear data, support vector machine (SVM) algorithms utilize kernel functions to map the data and find the optimal hyperplane. The choice of the kernel function involves considerations of its characteristics and the features of the data. Sigmoid and polynomial kernels, for example, require tuning multiple parameters, introducing more uncertainty and instability to the model. In contrast, the radial basis function (RBF) kernel only requires the introduction of a single parameter, enhancing efficiency and precision. Both Sigmoid and RBF kernels perform well in handling large-scale data and convergence speed. In this particular case, where the data is relatively complex and nonlinear with a small dataset, the RBF kernel was chosen, and a penalty parameter C was set to 1.0 to achieve better fitting of the data.

The SVM model was solved using MATLAB, resulting in the following Table 5 and Table 6 classification conclusions.

Table 5 Confusion Matrix Plot

Reject Coverage	29	22
Accept Coverage	1	51
	Reject Coverage	Accept Coverage

Table 6: Model Parameters

Parameter Name	Parameter Value	Parameter Name	Parameter Value
Training Set Ratio	70%	Model Convergence Parameter	0.001
Penalty Coefficient for Error Term	1.0	Maximum Iteration Times	2000
Coefficient Value of Kernel Function	0.01	Accuracy	83.21%
Multi-class Decision Function	Ovr	f1-score	0.79

Taking a comprehensive view, we used a total of 7 indicators as independent variables, with the insurance company's decision to insure as the dependent variable. The data was split into 70% for training and 30% for testing. The results show that our model achieved an accuracy of 83.21% on the test set, with an overall f1-score of 0.79, indicating a good fit of our model.

In conclusion, the classification results for whether to insure policies in various regions are as follows Table 7:

Table 7 Classification Results for Various Regions (Partial)

State	Whether to Insure Policies	State	Whether to Insure Policies
ALABAMA	Insure	ATLANTIC NORTH	Insure
ALASKA	Insure	ATLANTIC SOUTH	Insure
AMERICAN SAMOA	Insure	CALIFORNIA	Reject
ARIZONA	Reject	COLORADO	Insure
ARKANSAS	Insure	CONNECTICUT	Reject

2.5. Application of the Insurance Model in Japan

By examining the relevant 7 indicators in Japan, we applied the SVM model for analysis.Through the LSTM neural network, the paper predicted the data forecast for the occurrence of extreme weather in the next ten years,as shown in Fig 4.

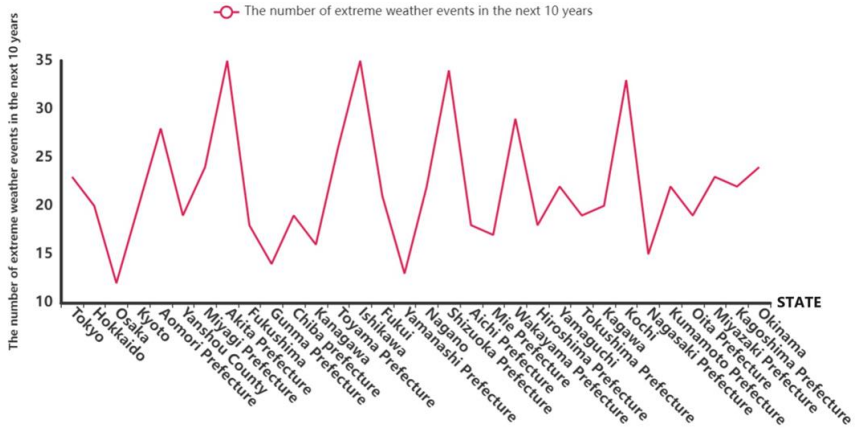


Fig 4 Japan's Forecasted Data for Extreme Weather Events in the Next Ten Years

Selected prefectures in Japan were used for classification modeling, and results obtained using the RBF kernel function for whether to underwrite insurance policies in each prefecture are as follows Table 8:

Table 8 Insurance Underwriting Results for Prefectures in Japan (Partial)

STATE	Insurance Status	STATE	Insurance Status
Tokyo	Insure	Toyama Prefecture	Insure
Hokkaido	Insure	Ishikawa	Insure
Osaka	Insure	Fukui	Reject
Kyoto	Reject	Yamanashi Prefecture	Insure
Aomori Prefecture	Insure	Nagano	Insure

3. Conclusions

The SVM binary classification model performs well in high-dimensional space and is suitable for processing datasets with large capacity. For the complex nonlinear data in this question, the SVM binary classification model is used to predict the effect with higher accuracy. The LSTM neural network prediction model is a model specially designed to process and learn the long-term dependencies of sequence data, and is suitable for analyzing and predicting the frequency of extreme weather over a longer period of time. The advantage is that it can filter out the anomalies, so that the extreme weather prediction results are more realistic. Based on SVM binary classification and K-Means clustering analysis models, we concluded that due to frequent extreme weather disasters such as thunderstorms and storms, insurance companies tend to reject coverage, similar to the first category studied earlier. Therefore, strong protective measures, including establishing protective measures and regular maintenance, should be implemented. Additionally, a letter to the local government of Agra was drafted, presenting a protection plan, progress timetable, and cost suggestions for the Taj Mahal.

References

[1] Zhou Ke, Spearman's study of the simple correlation coefficient[D].Guangdong:Guangdong University of Technology,2017

[2] Gao Zhenhai, Bao Mingxi, Gao Fei, et alAutomotive Engineering,2023,45(7):1145-1152,1162

[3] Xu Jianmin, Zang Peng, Shou Yanfang, Dynamic Programming Algorithm Based on ARIMA Model and K-means Clustering Analysis[J]Journal of Chongqing Jiaotong University (Natural Science Edition), 2022, 41(7) :9-13.

[4] Zhen Zhang, Yichen Zhang, Jiquan Zhang, et al., Urban resilience assessment based on entropy weight method and TOPSIS model---- taking Changchun City as an example [J].Journal of Catastrophology, 2023, 38(1): 213-219

[5] Cummins, J.D., Mahul, O., & Nzioki, E. Insurance and Climate Risk: Reflections on Selected Challenges and Opportunities. 2019.World Bank Group

[6] Mochizuki, Y., & Kunreuther, H. Risk Management of Natural Disasters in Japan and the Role of Earthquake Insurance. 2014.World Bank Group.

[7] Marlen, Zheng Yannan. Parameter Optimization of SVM Taxonomy: A Case Study of the Yellow River Estuary Wetland Application U1 Value Engineering,2022(1):203-204.205

[8] Zhen Zhang, Wei Hu, Weige Zheng, et al., Analysis and prediction method of oxygen partial pressure of manned spacecraft based on ARIMA model1. 2023,29(5):653-658.

[9] Yang Junchuang, Zhao Chao. A Review of K-Means Clustering Algorithms,Computer Engineering and Applications,2019,55(23):7-14.2019,55(23):7-14

[10] Yang Yanmei, Gao Ming, Zhao Jian, et al., Based on Intuitive Ambiguity-Entropy Weight-TOPSIS Research on Comprehensive Evaluation Methods of Agricultural Non-point Source Pollution Prevention and Control Technology, China Rural Water Resources and Hydropower, 2023(5): 18-25, 40