Advances in Computer, Signals and Systems (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2024.080412
ISSN 2371-8838 Vol. 8 Num. 4

A Study of Disaster Insurance in Extreme Weather Based on Logistic Regression and ARIMA Models

Shuting Wang[1,*]

[1]*Dept of Math, Qilu Normal University, Jinan, China*
[*]*Corresponding author: 2083628422@qq.com*

Keywords: Topsis, Extreme weather, Logistic model, ARIMA

Abstract: This paper examines the impact of extreme weather events on insurance under-writing decisions and real estate development. Using a combination of statistical mod-els, including Topsis entropy weight method and Logistic regression, we analyze the correlation between extreme weather indicators and insurance claims. Building upon this analysis, we investigate the factors influencing housing sales rates and forecast disaster losses using the ARIMA model. By treating the housing sales rate as a proxy for insurance compensation rates, we refine the insurance claims and profit model, providing insights for insurance underwriting decisions in different regions. Our findings offer new perspectives on mitigating risks and optimizing insurance policies in the face of changing environmental and social factors.

1. Introduction

Extreme weather events have a profound impact on social and economic stability, and the insurance industry plays a key role. With the intensification of climate change, insurance companies face more complex risk challenges. This study aims to explore the impact of extreme weather events on insurance underwriting decisions and real estate development, and to provide decision support for insurance companies. By combining extreme weather data and insurance claims data, we analyze the factors affecting the probability of insurance claims and predict the probability of future claims. We further explore the impact of social factors on home sales rates and use ARIMA models to predict disaster losses and improve insurance claims and profit models to provide new perspectives on insurance underwriting decisions and help optimize risk management strategies [1].

2. Insurance Claims vs. Profit Model

2.1 Data Processing

2.1.1 Data Visualization

We collected the annual maximum temperature difference, maximum temperature, minimum temperature, maximum precipitation, maximum wind speed and GDP of the corresponding region at different times in different regions. The histograms and KDE lines for each data are given below.

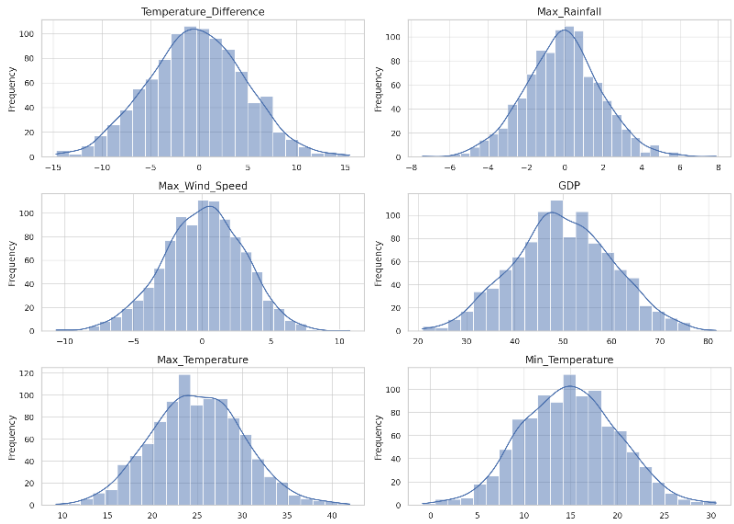

Figure 1: KDE images for each factor.

As can be seen in Figure 1, all data satisfy a normal distribution, and the distribution of the variable temperature difference is presented in a relatively symmetrical shape, implying that the data are roughly symmetrically distributed around the center value, suggesting that there was little change in the temperature difference during the observation period. Most of the data points are set around the center temperature value. The distribution of variable minimum temperatures is similar to the distribution of maximum temperatures, presenting a centrosymmetric bell shape, indicating that the temperature values are mainly concentrated within a certain range with relatively small variations.

2.1.2 Outlier handling

In order to analyze the data more effectively, the raw data is processed as follows:

Outlier judgment: due to the large number of measurements, the data is approximately regarded as a normal distribution, and the Laid principle is used to deal with the outliers of the data, that is, the 3∂ criterion, when the deviation is greater than 3∂, the data can be regarded as an outlier and can be eliminated, and the calculation formula for ∂ is:

$$\partial = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{1}$$

The criteria for determining outliers are as follows:

$$| x_i - \bar{x} | > 3\partial \tag{2}$$

Therefore, the range of normal data is $x_i > 3\partial + \bar{x}$ and $x_i < 3\partial - \bar{x}$, and the rest of the culling is the normal value.

Fill the missing items: For the missing items in the data, proximity interpolation and numeric interpolation are used to complete the character and numeric variables, respectively.

2.2 Establishment of insurance claims model

2.2.1 The importance of indicators is analyzed based on the Topsis entropy weight method

Through the statistics of global extreme weather and insurance claims, the local GDP, local

maximum temperature difference, local maximum rainfall, local maximum wind speed, local maximum temperature and minimum temperature are used as the criteria to measure whether extreme weather occurs and whether it is guaranteed [2].

The original matrix is forward-processed, and all indicator types are uniformly onverted into very large indicators.

Among the indicators, some data belong to very large indicators, that is, the larger the data, some belong to the range of indicators, that is, in a certain region, in order to eliminate the impact of different indicators, correctly reflect the comprehensive results brought by different reasons, the original data is positively processed.

The purpose of matrix standardization is to eliminate the influence of different.

Indicator dimensions, and the matrix composed of indicators with m samples is X.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \tag{3}$$

And X is normalized to obtain the normalization matrix Z, and the normalization formula for each element in Z is:

$$Z_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2}} \tag{4}$$

That is, each element divided by the sum of squares of all elements in the column s under the root number.

Calculate the weight of the value of option with under indicator j. For these indicators, the formula for calculating the information entropy is as follows:

$$e_j = -\frac{1}{\ln n}\sum_{i=1}^{n} p_{ij} \ln(p_{ij})(j = 1,2...m) \tag{5}$$

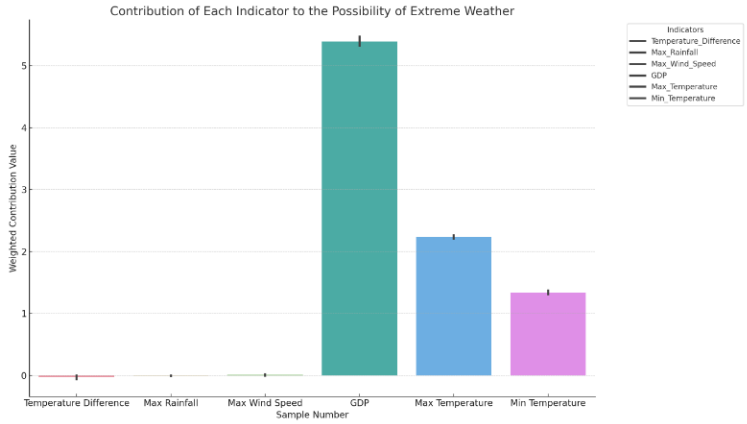

Figure 2: Contribution rate of each indicator.

It is clear from Figure 2 that GDP is the most important indicator, followed by temperature difference and maximum rainfall is the least important. From this, we can conclude that extreme weather can be judged on the basis of temperature difference, maximum precipitation, maximum wind speed, GDP, maximum and minimum temperature.

$$P = 0.1044 \times T_{diff} + 0.062 \times R_{\max} + 0.065 \times W_{\max} + 0.108 \times G + 0.089 \times T_{\max} + 0.09 \times T_{\min} \tag{6}$$

At the same time, we also know whether there will be claims in each region in the future, so we can use the logistic model to make more specific predictions about future claims, so the complete probabilistic model can be written as:

$$P = \frac{1}{1+e^{-(0.1044\times T_{diff} +0.062\times R_{\max} +0.065\times W_{\max} +0.108\times G+0.089\times T_{\max} +0.09\times T_{\min})}} \tag{7}$$

After fitting the function using logistic regression, the ROC curve was obtained to assess the predictive performance of the model (Figure 3).The AUC value of the area under the ROC curve was 0.82, and the closer the AUC value is to 1, the better the predictive performance of the model is, so the model has a strong ability to differentiate between the occurrence of an insurance claim and the occurrence of an insurance claim, and has a higher predictive accuracy of whether or not extreme weather occurs in the future [3].

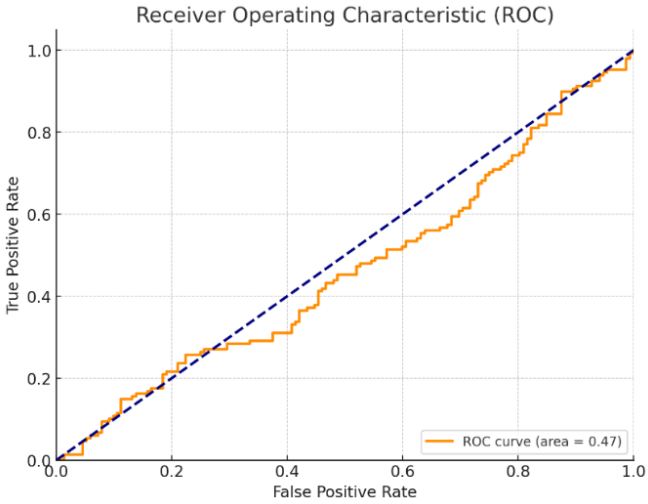

Figure 3: ROC plot of the logistic model.

Considering the most basic scenario, the equation for an insurance company's profitability is:

Profit = Principal + Policy Premium - Compensation

However, due to the increasing number of natural disasters due to climate change, insurers need to find a balance between whether to underwrite a policy. We will build a model that will determine whether an insurer should cover a catastrophe policy in the face of a catastrophe policy based on the number of extreme weather events in the insured area, the number of times the insurer has covered the insurance, the losses in the affected area, and the insurer's own assets [4].

The differential equation for the accumulated assets of the insurance company after T years is:

$$\frac{dZ}{dt} = G - E*P*W \tag{8}$$

$$P = \frac{1}{1+e^{-(0.1044\times T_{diff} +0.062\times R_{\max} +0.065\times W_{\max} +0.108\times G+0.089\times T_{\max} +0.09\times T_{\min})}} \tag{9}$$

$\frac{dZ}{dt}$ Indicates the rate of change in an insurance company's assets over time.

G It is the total amount of premiums received by the insurance company from the catastrophe policy.

$E*P*W$ is the expected annual total claim, E is the average loss per extreme weather event, P is the probability of extreme weather occurring, and W is the average compensation rate.

2.2.2 Insurance Strategy

A key concept when discussing insurers' strategies for running catastrophe insurance business is the "balance point". This concept states that in order to ensure that the company's assets remain stable, the equilibrium point is calculated to be in $G = E*P*W$. This means that to maintain neither increase nor decrease in assets (i.e., to achieve a balance between profit and loss), the difference between the total premiums received by the insurer from the catastrophe policy and the expected total annual payout should be equal. In this simplified model, if $G-E*P*W$ greater than 0, the insurance company will make a profit, and if $G-E*P*W$ less than 0, it will lose money.

Secondly, we assume that the initial assets of the insurance company are 0, the risk appetite is k, when $G-E*P*W$ less than 0, if $1-k > \frac{E*P*W}{Z}$, the insurance company is still willing to bear the catastrophe policy in the case of possible losses, and may accumulate a certain positive social impact, which is conducive to the future development of the insurance company, and when $1-k < \frac{E*P*W}{Z}$, the insurance company will definitely lose money if it still has to bear, which is not conducive to long-term stable development, then the insurance company should not bear the policy. At this point, homeowners can reduce the risk of potential disasters by investing in improved safety and disaster prevention measures for their homes. This may include installing fire alarm systems, firewalls, flood defenses, etc. By reducing risk, the property becomes more insurable and insurance companies may be more willing to cover it.

3. Real estate disaster prediction loss rate model

Firstly, based on the ARIMA model and the disaster loss statistics of the past ten years, the future disaster loss is predicted, then the linear regression is used to analyze several factors affecting real estate sales, and the linear relationship is found.

3.1 Based on ARIMA's projections of future losses.

The ARIMA model is a statistical model that is widely used for time series data analysis and forecasting [5]. The ARIMA model can describe the autocorrelation of univariate time series and is suitable for those with time series data, especially for non-seasonal data. The ARIMA model predicts future time series data points by combining three basic methods: autoregressive (AR), differential (I), and moving average (MA).

The ARIMA model is usually expressed as ARIMA(p,d,q), p denotes the order of the autoregressive term in the model, d denotes the number of differences required to smooth the sequence, and q denotes the order of the moving average term.

Autoregressive part: reflects the relationship between the current value and its own past value.

Differential part: Used to convert a non-stationary time series to a stationary time series.

Moving Average section: Describes the relationship between the current value and the past error term.

ARIMA is used to predict disaster losses in the region over the next 10 years.

Using ARIMA (2,0,2) to show that the data are predicted for the next ten years after two-order autoregression, zero difference and two-order shift, it can be observed that the actual data has a peak around 2010 and then fluctuates and decreases. Forecasts show that the number of losses will remain relatively flat from 2020 onwards (Figure 4).

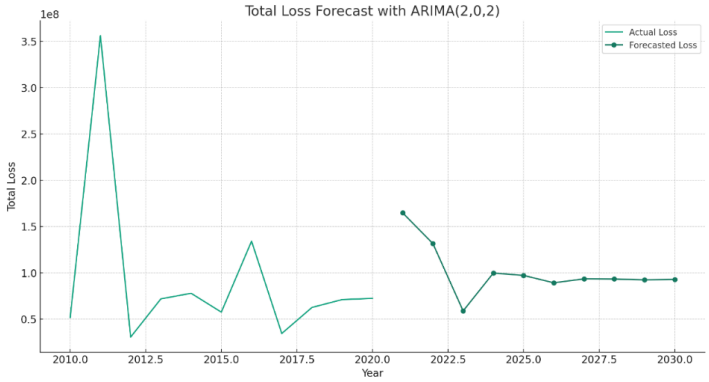

Figure 4: Time series predicts losses over the next decade.

3.2 Real estate claim model for risk areas based on linear regression.

We consider that due to the characteristics of extreme weather, there may be different extreme weather in different latitudes and climates. Due to the different types and probabilities of extreme weather events in different regions, and the impact on real estate is also different, we will comprehensively consider the impact of extreme weather on the past occurrence x1, future forecast occurrence x2, insurance compensation rate x3, age ratio x4, employment rate x5, crime rate x6, and GDP x7 on the house sales rate x8, and draw a heat map.

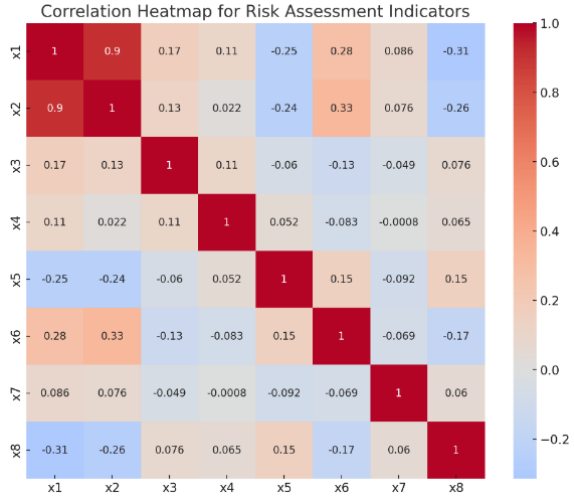

Figure 5: Diagram of the heat relationship between the factors.

In Figure 5, we can see the correlation coefficients between homeownership and other indicators. Key observations include the negative correlation between the number of extreme weather events (projected over the past and next ten years) and homeownership. The correlation between the percentage of insurance expenditures per capita and homeownership is weak at 0.08. The correlation coefficient between employment status (employment rate) and homeownership is 0.15, suggesting that areas with higher employment rates also have relatively higher homeownership rates. Area crime rate has a negative correlation with homeownership rate of -0.17, which may imply that areas with higher crime rates have lower homeownership rates. The correlation between GDP per capita and homeownership is 0.06, indicating a weak linear relationship. Relationships between homeownership and various indicators. Next, we plotted a scatterplot of several highly correlated factors as in Figure 6 and showed more specific effects on homeownership.

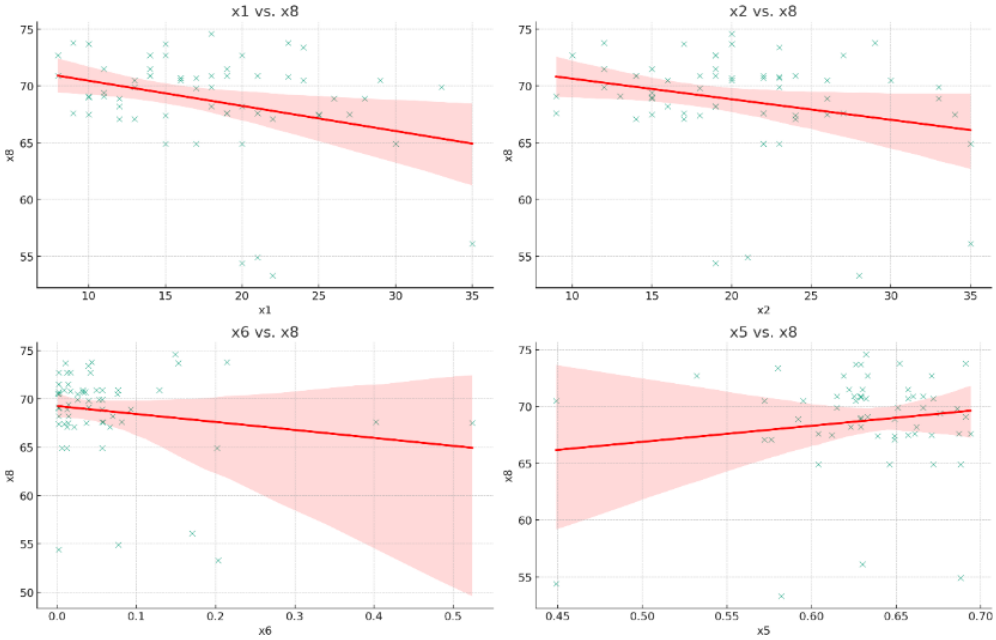

Figure 6: A scatterplot of the top four factors with the highest correlation

Expressing the impact of these four indicators on homeownership, we can create a multiple linear regression model in which homeownership (Y) is the dependent variable [6], and the number of extreme weather events in the past decade (X_1), the predicted number of extreme weather events in the next decade (X_2), the employment rate (X_3)and the regional crime rate are the independent variables (X_4).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \tag{10}$$

β_0 is the intercept, $\beta_1, \beta_2, \beta_3, \beta_4$ is the coefficient of the respective variable and ε the error term and the data are used for multiple linear regression analysis to obtain the coefficients of this comprehensive expression. A mathematical expression of the integrated model, combining the impact of four indicators on homeownership. It can be expressed as:

$$Y = 64.8 - 0.3X_1 + 0.14X_2 + 10.83X_3 - 6.56X_4 \tag{11}$$

The model shows that the number of extreme weather events in the past decade (X_1) has a negative impact on homeownership, the number of predicted extreme weather events in the next decade (X_2) has a slight positive impact on homeownership, employment (X_3) has a significant positive impact on homeownership, and regional crime (X_4) has a negative impact on homeownership.

4. Conclusions

This study provides new perspectives and decision support for the insurance industry through an in-depth analysis of the impact of extreme weather events on insurance underwriting decisions and real estate development. By combining extreme weather data and insurance claims data, we successfully explored the impact of different factors on the probability of insurance claims and utilized a model to predict future claim probabilities.

In addition, we further explored the key influences of social factors on the home sales rate and successfully predicted disaster losses using an ARIMA model. Considering the home sales rate as a proxy for the insurance claims rate, we refined the insurance claims and profit models to provide a more comprehensive consideration for insurance underwriting decisions. Through comprehensive

analysis, we suggest that insurance companies consider changes in social and environmental factors when formulating insurance policies and flexibly adjust risk management strategies to meet changing challenges.

Overall, this study provides an important reference for the insurance industry in facing the increasingly complex climate risks and social factors, and offers new ideas and methods for insurance underwriting decisions and risk management. We hope that these research results will contribute to the development and sustainability of the insurance industry, and promote the insurance industry to better cope with future challenges and opportunities.

References

[1] LI Yanbo, LIU Miaoyang, YANG Kai, et al. Optimal scheduling of mobile power vehicles with self-consistent energy system for highways under extreme weather [J/OL]. Journal of Jilin University (Engineering Edition), 1-10[2024-05-25]. https://doi.org/10.13229/j.cnki.jdxbgxb.20240224.

[2] Luo Qian, Li Yongmei, Wang Tenghua, et al. Constructing an evaluation system of rational medication use indexes in clinical departments based on improved entropy weight method combined with TOPSIS method [J]. Clinical rational drug use, 2024, 17(15): 170-172+ 177. DOI: 10. 15887/ j. cnki.13-1389/r.2024.15.049.

[3] Dai Daocheng,Yu Chenyang,Song Jihao,et al. Analysis of smartphone user monitoring data based on logistic regression [J]. Modern Information Technology, 2024, 8(08): 36-39. DOI: 10. 19850/ j. cnki. 2096-4706.2024.08.009.

[4] Tian Gengwen. Catastrophe insurance supports "umbrella" in response to disasters [N]. Rural Financial Times, 2024-05-06(A01).

[5] Xiang Junkun, Yu Jiaxing, Gao He, et al. Constructing SWECPX model based on ARIMA to solve the e-commerce demand forecasting problem [J]. China Business Journal, 2024, (08): 29-32. DOI: 10.19699/j.cnki.issn2096-0298.2024.08.029.

[6] WANG Weiqiang, LI Yongkang, SHENG Yali, et al. GF-1 PMS multispectral image reconstruction method based on multiple linear regression model [J]. Journal of Anhui Institute of Science and Technology, 2024, 38(03): 70-77. DOI: 10.19608/j.cnki.1673-8772.2024.0309.