

Research on insurance policy formulation based on ARIMA-KMEANS algorithm

Zhufangyuan Ji, Yukai Wang, Jin Su, Jie Wu*

School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China, 214122

*Corresponding author: jjiewu20@jiangnan.edu.cn

Abstract. This paper is about the property insurance industry facing the challenge of extreme weather events. The research focuses on how to adapt and mitigate these challenges in order to maintain the sustainability of the insurance market and the financial security of property owners. This study focuses on the impact of extreme weather events on the property and casualty insurance industry, the risk assessment framework, and the use of open source risk assessment software and climate prediction models to analyze the impact of climate change on exposure to extreme climate events. This research fills a gap in the existing literature by developing more robust risk assessment methodologies, integrating climate change projections into insurance policies, and considering the role of socioeconomic factors in insurance decisions. A scientific disaster risk quantification framework is developed, which provides theoretical basis and methodological support for the formulation of disaster management strategies. The risk region is divided by K-means algorithm, and the Arima model is used to predict the potential future economic loss of the insurance industry.

Keywords: K-means Algorithm, Arima Model, Extreme Weather Insurance.

1. Introduction

Global climate change is altering our living environment at an unprecedented rate, and the frequency and intensity of extreme weather events such as floods, hurricanes, droughts, and wildfires are increasing. ^[1]This trend poses a significant risk to property owners and an unprecedented challenge to the insurance industry. In recent years, global losses due to these extreme weather events have exceeded \$1 trillion, resulting in a significant increase in insurance industry claims. ^[2]Insurers are faced with the reality that they must rethink their underwriting strategies, while the rapidly rising cost of insurance and the widening insurance coverage gap have highlighted a sustainability crisis for the industry as a whole. A substantial body of research exists concerning the impact of extreme weather events on the property insurance industry. Scholars have extensively explored the economic ramifications, risk assessment strategies, and underwriting policies in response to these challenges.

For many years, risk assessment frameworks, such as catastrophe models, have been instrumental within the insurance and reinsurance sectors for quantifying financial risks. Recently, the availability of publicly accessible, open-source risk assessment frameworks has significantly increased. Notable examples of these frameworks include CLIMADA ^[3] and the Oasis Loss Modelling Framework ^[4]. These open-source platforms provide enhanced flexibility in risk quantification, extending their utility beyond financial risk to include the assessment of non-financial risks. For instance, the Oasis Loss Modelling Framework was utilized by Salmanidou et al. (2021) to investigate the potential impact of future tsunamis originating from the Java Trench on household welfare ^[5]. Similarly, Kam et al. (2021) employed the CLIMADA framework to evaluate the increasing global risk of human displacement due to river floods ^[6]. These advancements in risk assessment tools offer a more comprehensive and adaptable approach to understanding and managing the complex nature of risks in a changing world.

Recent research has increasingly integrated open-source risk assessment software with ensembles of climate projections to analyze the consequences of climate change on exposure to extreme climatic events. Lange et al. (2020) examined the effects of climate change on global exposure to extreme climate impact events, such as floods, droughts, and wildfires, ^[7] using Probabilistic Event Attribution (PEA) derived from four Global Climate Models (GCMs) from the Fifth Phase of the Coupled Model

Intercomparison Project (CMIP5). Similarly, Kam et al. (2021) quantified future flood risks by combining data from six global hydrological models and four GCMs. Furthermore, Stalhandske et al. (2022) ^[8] assessed the future risk of heat stress in Switzerland employing the CH2018 Climate Scenarios ^[9], based on the EURO-CORDEX multi-model ensemble ^[10]. These studies highlight the importance of incorporating uncertainty quantification in risk assessments, accounting for the variability inherent in climate model projections.

While previous studies have provided valuable insights, there remain several gaps that warrant further exploration. These include the development of more robust risk assessment methodologies, the integration of climate change projections into underwriting policies, and the consideration of socio-economic factors in insurance decision-making. Our research extends the existing literature by addressing these gaps and offering novel perspectives on mitigating the challenges posed by extreme weather events. Building upon the existing body of research, our study focuses on examining how the property insurance industry is adapting to the increasing frequency of extreme weather events. We recognize the substantial losses incurred due to these events and the substantial global insurance protection gap, which now averages 57%. Our research aims to contribute to the field by proposing innovative models and approaches to balance underwriting policies effectively. By ensuring the long-term viability of insurers and affordability for property owners, our study seeks to address the sustainability crisis facing the property insurance industry. (Data Source: <https://www.kaggle.com/datasets/thedevastator/weather-disaster-costs-and-deaths?resource=download>)

2. Risk Assessment of Affected Areas

2.1. Quantification of impacts in disaster-affected areas

The study delves into the quantification of impacts in disaster-affected areas, conducting a comprehensive analysis of key elements in disaster risk assessment. This includes the identification of indicators and the determination of their respective weights. Initially, three primary indicators are identified: personal safety risk, property loss risk, and disaster intensity risk. Based on these, five critical secondary indicators are selected for a comprehensive assessment of disaster risk: the number of people affected by disasters, the number of casualties, direct economic losses, crop losses, and the duration of disasters. These indicators are deemed essential for a holistic evaluation of disaster risk.

To further refine the risk quantification model, this study introduces two new secondary indicators: Disaster Density (DD) and Per Capita Mortality Index (PC), thereby enhancing the model's predictive accuracy by incorporating considerations of area and population density. Specifically,

Disaster Density (DD) is defined as the ratio of the number of disasters to the state area, with the calculation formula being:

$$DD = \frac{DHN}{SA} \quad (1)$$

Here, DHN represents the total number of disasters in a specific area, and SA represents the area of the state.

Per Capita Mortality Index (PC) takes into account the impact of population density on casualty numbers, with the calculation formula being:

$$PC = \frac{DN}{\beta} \quad (2)$$

Here, DN represents the number of deaths caused by disasters, and β signifies population density.

To enhance the model's accuracy, the Principal Component Analysis (PCA) method was employed to ascertain the weights of each indicator. The process involved standardizing the dataset, computing the correlation matrix, analyzing eigenvalues and eigenvectors, selecting principal components that account for over 85% of the cumulative contribution rate, and ultimately constructing the weight

vector. This procedure allowed for a quantitative assessment of each indicator's contribution, as shown in Fig 1.

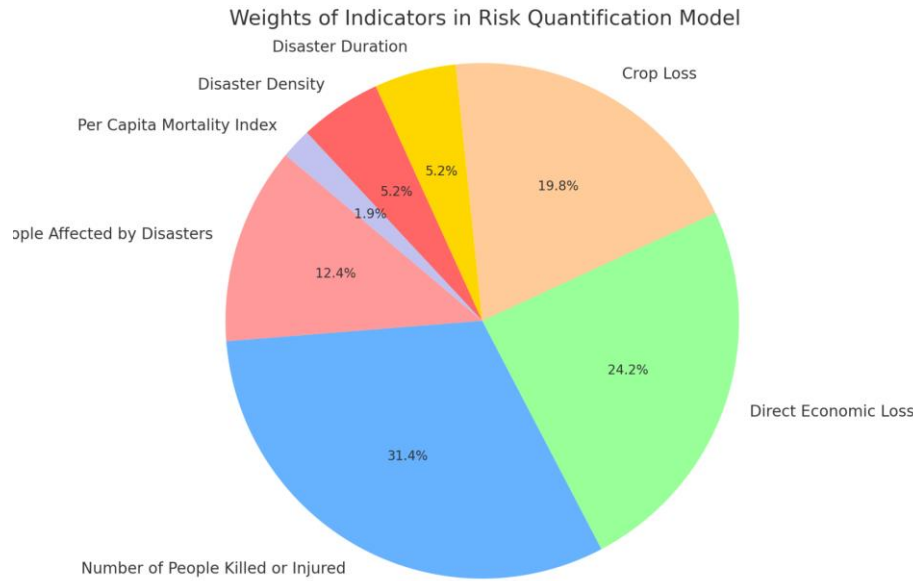


Fig 1: Weights of Indicator in Risk Quantification Model

The calculation of the Weight Vector (W_j) is defined as:

$$W_j = \frac{v_j}{\sum_{k=1}^m v_k} \quad (3)$$

Here, v_j represents the eigenvector associated with the j -th eigenvalue, and m denotes the number of principal components selected.

Through this detailed analysis, the study developed a scientific disaster risk quantification framework. It lays a solid theoretical foundation and offers methodological support for formulating disaster management strategies. This framework serves as an effective quantification tool and provides guidance for the insurance industry in evaluating potential risks from future extreme meteorological events.

2.2. Risk region division based on KMEANS algorithm

Risk determines the formulation of underwriting policies by insurance companies for the foreseeable future. Based on the disaster risk factors calculated for each state as mentioned previously, As shown in Fig 2, the K-MEANS algorithm is applied to classify the risk levels. (Data source: <https://www.kaggle.com/datasets/pompelmo/usa-states-geojson>)

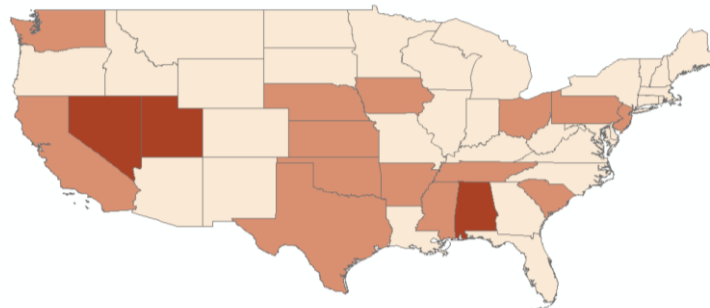


Fig 2: Risk Region in US

By clustering the risk scores into three categories: low risk, medium risk, and high risk, insurance companies can more precisely assess the potential risk of each state and accordingly adjust the pricing and terms of insurance products. This data-driven approach provides an objective framework for risk

assessment, aiding insurance companies in optimizing resource allocation, reducing potential future losses, and offering services that better meet customer needs.

3. The D-R-I Evaluation Model

3.1. Determination of Risk Factors

To delineate risk factors effectively, the model partitions geographic regions into areas of varying risk levels—low, medium, and high. Initial risk level parameters are established for these zones to construct a foundational framework that quantifies risk assessment benchmarks corresponding to each level.

The underwriting risk factor, denoted by M , is preliminarily defined as follows:

$$M = \alpha_{different\ risk} \quad (4)$$

Here, M constitutes a pivotal component in the risk evaluation process, providing a metric that insurance companies can utilize to ascertain the viability of underwriting risks in light of the divergent destructive capacities of various disaster types.

Historical data and risk level analyses have facilitated the integration of nuanced disaster evaluations into the underwriting risk factor assessment. This comprehensive approach has led to the enhancement of the underwriting risk factor M , reformulated as:

$$M = \alpha_{different\ risk} + \sum_i \alpha_i \times \omega_i \quad (5)$$

In the equations above, $\alpha_{different\ risk}$ signifies the baseline risk parameter. The variable ω corresponds to the weight attributed to each disaster type, and each α designates the risk coefficient for that specific disaster.

Through meticulous historical scrutiny and the analytical synthesis of risk levels, a revised risk quantification methodology emerges, encapsulating the intricate dynamics between various disaster types as shown in Table 1. Each risk type is assigned a risk factor which, when multiplied by its respective base risk parameter α and coefficient, yields a calibrated value indicative of its impact on the overall risk assessment.

Table 1 Calibrated Risk Values for Various Disaster Types

Event Type	Risk Factor
Hail	0.8256
Thunderstorm Wind	1.0000
Tornado	0.9540

3.2. Prediction of Insurance Claim Amount Based on ARIMA Model

Considering the inherent characteristics of time series data, this study opted for the ARIMA model for its analysis. This model is particularly suitable for analyzing and forecasting non-seasonal time series data that exhibits distinct trends or patterns. It predicts future values by integrating the historical values and error terms of the time series data, aligning with the study's goal of assessing potential losses to the insurance industry from future extreme climate events.

To ensure the applicability of the ARIMA model, the study initially conducted an ADF test to verify the stationarity of the data. Non-stationary data were subjected to a differencing process once to meet the basic conditions for time series analysis, with results presented in the following Fig 4. Subsequently, the study, based on determined optimal parameters, employed the ARIMA model to forecast future economic losses, as illustrated in Fig 3,4.

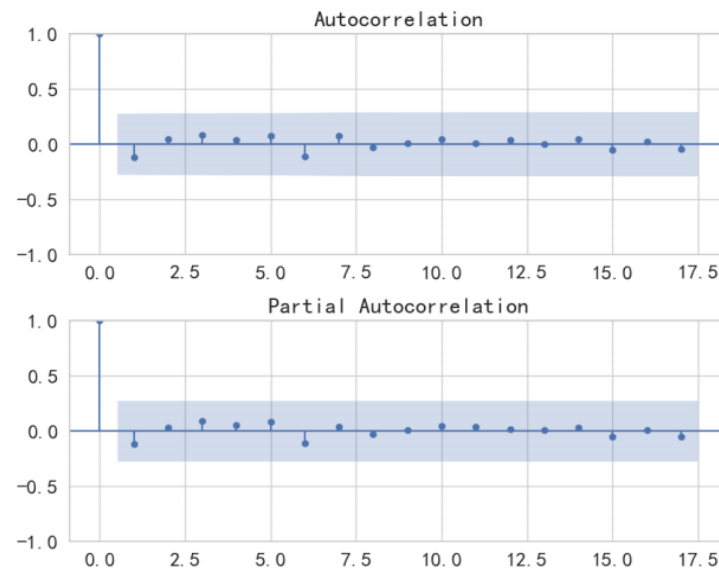


Fig 3 ACF and PACF Analysis for Stationarity Assessment

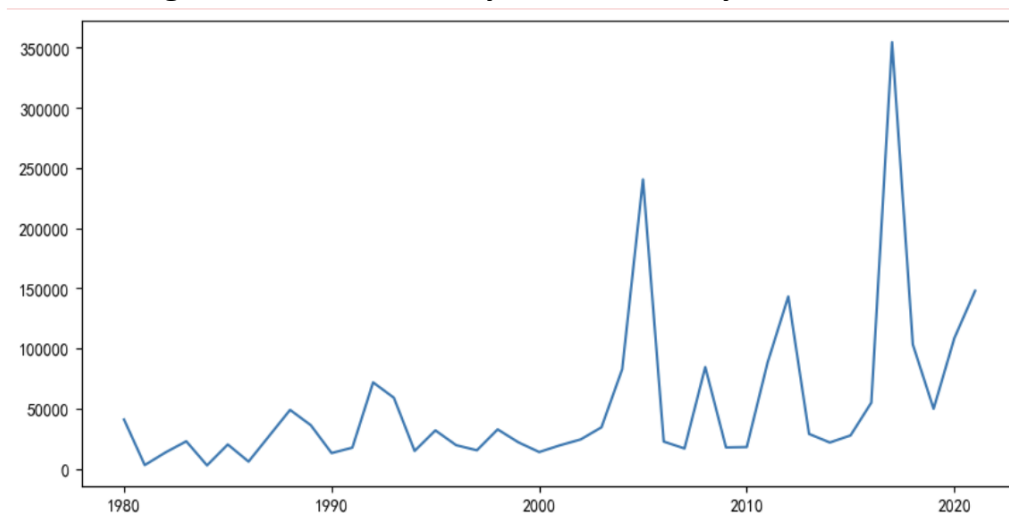


Fig 4 Historical Trends in Insurance Claim Amounts Over Time

3.3. Analysis of experimental results

From the comparison between prediction data and actual data, the BP neural network has better prediction performance and relatively small error, which can meet the demand completely, and has fast prediction speed and convenient operation.

The insurance cost participation for each individual is crucial in determining the underwriting policies of insurance companies for different risk areas. This paper describes the use of the underwriting risk factor (M) and the estimated total insurance claim amount (L) to establish a fundamental calculation for insurance cost participation. To encompass the spectrum of factors influencing the insurer's financial responsibility and target profit margins, four additional indicators are integrated: the number of insured individuals (N), the compensation cost (C), the expected profit (P), and the state's total population.

The comprehensive cost participation formula amalgamates these five indicators:

This formula comprises three distinct components: The base cost ($C \cdot M$): Allocated costs for different risk areas. The expected profit share (N/L): The profit designated per insured individual. The individual compensation amount (N/P): The expected compensation disbursed per insured individual.

An example calculation is provided for (2024), where the expected profit is (\$50,000), the number of insured individuals is (10,000), and the total US population is 140,000,000. With a base

participation cost of \$100 per person, as per industry standards, and a risk factor of (2.93), the calculated specific participation cost for the region is (\$300.21).

This formula and accompanying analysis serve as a cornerstone in formulating strategic insurance policies and reflect a tailored approach to managing financial risk in various geographic locations.

3.4. Application of the model in the United States

The risk level classification for various states in the United States is as shown in Fig 2. Due to space limitations, only the specific insurance costs for high-risk states are displayed in the table 2 below:

Table 2 Specific Insurance Costs for High-Risk States in US

STATE	RISK_SCORE	E
ALABAMA	2.93	300
NEVADA	2.47	250
UTAH	2.98	305

4. Conclusion

Through an in-depth analysis of extreme weather events faced by the property and casualty insurance industry, this study reveals the complexity and multidimensional nature of the impact of these events on the industry. By adopting the risk region division based on the K-MEANS algorithm, we successfully classified the risk level in the study area into three levels: low, medium and high. The results show that insurance claims are expected to increase by about 57% in high risk regions, a significant difference compared to medium and low risk regions. At the same time, the application of the D-R-I assessment model effectively identifies the key risk factors that lead to the increase in insurance costs, including the quality of home construction, the history of natural disasters in the region, and population density. Using ARIMA models to forecast future insurance claims, we find that total insurance claims due to extreme weather events are likely to increase by more than 40% over the next five years, especially in areas classified as high risk. The forecast highlights the urgency for the insurance industry to make major changes in risk management and pricing strategies.

References

- [1] CHENG Qiyun, SUN Caixin, ZHANG Xiaoxing, et al. Short-term load forecasting model and method for power system based on complementation of neural network and fuzzy logic [J]. Transactions of China Electrotechnical Society, 2023, 19(10): 53-58.
- [2] FANG Fang. Research on power load forecasting based on improved BP neural network [D]. Harbin Institute of Technology, 2021.
- [3] AZNAR-SIGUAN G, BRESCH D.N. CLIMADA v1: a global weather and climate risk assessment platform [J]. Geosci. Model Develop., 2019, 12: 3085-3097.
- [4] OASIS LMF. Oasis loss modelling framework [EB/OL]. <https://oasislmf.org/>, 2022. Accessed: 2022-05-24.
- [5] SALMANIDOU D.M., EHARA A., HIMAZ R., HEIDARZADEH M., GUILLAS S. Impact of future tsunamis from the Java trench on household welfare: merging geophysics and economics through catastrophe modelling [J]. Int. J. Disaster Risk Reduct., 2021, 61: 102291.
- [6] KAM P.M., et al. Global warming and population change both heighten future risk of human displacement due to river floods [J]. Environ. Res. Lett., 2021, 16: 044026.
- [7] LANGE S., et al. Projecting exposure to extreme climate impact events across six event categories and three spatial scales [J]. Earth's Future, 2020.
- [8] STALHANDSKE Z., et al. Projected impact of heat on mortality and labour productivity under climate change in Switzerland [J]. Natural Hazards Earth Syst. Sci., 2022, 22: 2531–2541.

- [9] NATIONAL CENTRE FOR CLIMATE SERVICES, Zurich. NCCS: CH2018 – climate scenarios for Switzerland, technical report [EB/OL]. <https://www.nccs.admin.ch/nccs/en/home/climate-change-and-impacts/swiss-climate-change-scenarios/technical-report.html>, 2018. Accessed: 15-08-2022.
- [10] JACOB D., et al. EURO-CORDEX: new high-resolution climate change projections for European impact research [J]. *Regional Environ. Change*, 2014, 14(2): 563–578.