Research on insurance model based on different mathematical models and algorithms

Cailin Liao [1, *], Nan Hu [2]

[1] School of Optical and Electronic Information, Huazhong University of Science and Technology, Wuhan, China, 430074

[2] College of Chemistry and Chemical Engineering, Hunan University, Changsha, China, 410082

* Corresponding author: 18284581946@163.com

Abstract. This article presents a risk assessment and prediction model for specific areas using the ARIMA model, entropy weight method, and AGNES algorithm. The model assists insurers in determining coverage and adjusting premiums based on economic losses predicted by the ARIMA model. It also aids real estate companies in deciding whether to develop in certain areas. The ERM model considers human, weather, and geographical factors to assess risks and predict potential property damage caused by extreme weather. The AGNES algorithm clusters similar samples to identify regions with higher risks. This helps real estate companies make informed decisions about development projects. The entropy weight method is used to adjust insurance premiums by calculating changes in regional risk coefficients. Continuous monitoring and analysis of risk coefficients improve risk assessment accuracy and enable reasonable premium adjustments.

Keywords: ARIMA, AGNES, Entropy Weight Method, Risk Assessment.

1. Introduction

The world's climate has changed significantly in recent decades as a result of human activities, especially the use of fossil fuels [1]. According to scientific studies, extreme weather events are becoming more frequent and more intense as a result of climate change [2].

The effects of global climate change are being felt around the world in various forms of extreme weather events. These events include, but are not limited to, droughts, floods, heat waves, cold snaps, hurricanes, tornadoes, earthquakes, and wildfires. The frequency and intensity of these events are increasing, causing huge economic, social and environmental losses to the world [3-5].

Given the escalating risks associated with extreme weather, insurers face the important task of effectively assessing and managing these risks. In order to determine insurance rates that accurately reflect potential risks in different regions, they need to develop a robust strategy. Russian scientists use the summary and analysis of insurance statistical data, synthesize and systematize the positive and negative factors of the development of the insurance market, and conduct research based on hypothesis reasoning models, statistical methods and modeling techniques [6]. They analyzed the risks and opportunities for the effective functioning of the entire insurance system and its individual components and assessed the prospects for the development of the Russian insurance market.

At present, scientists have considered in detail various influencing factors in the region, including regional economic structure, regional GDP scale, industrial production scale and residents' income, etc., and have conducted detailed analysis and adjustments to the insurance market in the region [7], and even regional climate models can be leveraged to provide index-based drought insurance for sovereign disaster risk transfer [8].

The models we build scientifically predict future risk losses by carefully evaluating historical data and assessing risks in specific areas to help insurance companies determine appropriate premium levels to provide adequate protection for policyholders while ensuring financial stability and insurance. long-term development of the industry. In addition, the model we built can also help real estate companies assess the risks in a specific area and decide whether to develop properties there to protect their interests.

2. Fundamentals of insurance models

2.1. ARIMA Model

The ARIMA model (Autoregressive Integrated Moving Average Model) is a commonly used time series analysis model used to predict future values. It combines the properties of autoregressive (AR), difference (I) and moving average (MA) models [9]. It works with many types of time series data, including data with linear trends, seasonal changes, and cyclical fluctuations. It automatically selects appropriate parameters based on the characteristics of the data and adapts to different data patterns. Moreover, it can be used for time series prediction. By modeling and analyzing historical data, the ARIMA model can capture trends and seasonality in the data to make future predictions. This makes the ARIMA model widely used in demand forecasting, stock price forecasting and other fields [10-11].

The general model of ARIMA consists of three basic elements, which are:

(1) The autoregressive (AR) model describes the correlation between the current value of the time series and several past values. The order p of the AR model represents the number of past observations. The expression of AR(p) can be written as:

$$Y(t) = \mathrm{C} + \sum_{i=1}^{p} \varphi_i \times Y(t-i) + \varepsilon(t) \tag{1}$$

In the above formula, Y (t) represents the current value of the time series, c is a constant, φ1, φ2, ..., φp are autoregressive coefficients, and ε (t) is the error term. The key to AR lies in the estimation of the autoregressive coefficients, which can be determined using least squares or maximum likelihood estimation.

(2) The difference (I) model is used to eliminate the non-stationarity of the time series. Stationary series are characterized by stable mean, variance or autocorrelation functions. The first difference can be expressed as:

$$\Delta Y(t) = Y(t) - Y(t-1) \tag{2}$$

If the sequence after first-order difference is still not stationary, multiple-order differences need to be continued until a stationary sequence is obtained. The mean, variance and autocorrelation function of a stationary sequence are all constant, so that a more consistent model can be established.

(3) The moving average (MA) model uses a linear combination of several past prediction errors to predict future observations. The order q of the MA model represents the number of past prediction errors. The expression of the MA(q) model can be written as:

$$\varepsilon(t) = \mathrm{C} + \sum_{i=1}^{q} \theta_i \times \varepsilon \tag{3}$$

The key to the MA model is the estimation of the moving average coefficient, which can also be determined using the least squares method or maximum likelihood estimation.

The selection of p and q can be determined by the images of the autocorrelation function (ACF) and the partial autocorrelation function (PACF), while the selection of d can be determined by data analysis after difference. The predictions of the ARIMA model rely on historical observations and estimates of model parameters. By fitting models to historical data, estimates of model parameters can be obtained, and these parameters can be used to predict future values.

2.2. AGNES algorithm

Agnes algorithm (Agglomerative Nesting) is an agglomerative hierarchical clustering algorithm, also known as a bottom-up clustering algorithm. The algorithm starts with each data point as a cluster and then merges the most similar clusters until a stopping condition is met. The following are the detailed steps of Agnes algorithm:

(1) Initialization: Treat each data point as an independent cluster.

(2) Calculate the similarity matrix: Calculate the similarity between each two clusters according to the selected similarity measurement method. The similarity measure used in this article is Euclidean distance, which can be calculated using the following formula:

$$d(A,B) = \sqrt[2]{\sum_{i=1}^{n}(a_i - b_i)^2} \tag{4}$$

(3) Merge the most similar clusters: Select the two clusters with the highest similarity, merge them into a new 33. cluster, and update the similarity matrix.

Repeat step 3 until the stopping condition is met.

(4) The stopping condition can be that the number of clusters reaches a predetermined threshold, or the similarity is lower than a certain threshold. Return the final clustering results.

The advantage of the Agnes algorithm is that it is easy to implement and understand, and can generate visual hierarchical clustering dendrograms. However, this algorithm has high computational complexity and needs to calculate and store the similarity matrix between all data points. In addition, the Agnes algorithm is sensitive to noise and outliers.

2.3. Entropy weight method

Entropy Weight Method is a commonly used multi-index weight determination method, used for index weight allocation in evaluation and decision-making problems [12].

It is based on the information entropy theory and takes into account the importance of different indicators by calculating the entropy value and weight of the indicators. It can comprehensively consider the differences between indicators by calculating the entropy value of indicators. The entropy values of different indicators reflect their degree of dispersion and uncertainty in the evaluation object. The greater the entropy, the greater the difference in indicators. Moreover, its calculation process is relatively simple and easy to understand and operate. The weight of the indicator can accurately reflect its importance in decision-making, which helps to formulate reasonable decision-making plans.

The general model of the entropy weight method consists of four basic steps, which are:

(1) Standardized indicator data: Standardize the collected indicator data to eliminate differences in dimension and magnitude between different indicators.

(2) Calculate information entropy: Calculate the information entropy of each indicator based on the standardized indicator data. The calculation formula of information entropy is:

$$H(i) = -\frac{1}{\ln n}\sum_{i=1}^{n} p_i \times \ln p_i \tag{5}$$

In the formula, different p is the relative frequency of indicator data, and H(i) is the information entropy of the i-th indicator.

(3) Calculate weight: Calculate the weight of each indicator based on the index entropy value calculated by information entropy. The calculation formula of weight is:

$$\varphi(i) = \frac{1-H(i)}{\sum 1-H(i)} \tag{6}$$

In the formula, φ(i) is the information entropy of the i-th indicator.

(4) Normalized weight: Normalize the calculated index weights so that the sum of all weights is 1. The calculation formula of normalized weight is:

$$\omega(i) = \frac{\varphi(i)}{\sum \varphi(i)} \tag{7}$$

Through the above steps, the weight of each indicator can be obtained. The larger the weight value, the higher the importance of the indicator in decision-making. The entropy weight method has high requirements on the initial data of the indicators, and the indicator data must have complete statistical information; it is easily affected by the disturbance and error of the indicator data; it cannot solve the problem of non-linear relationships between indicators.

3. Results

3.1. Future loss prediction

All data in this article comes from:
https://www.emdat.be/
https://www.undrr.org/data
https://www.gdacs.org/
https://www.weather.gov/
https://earthquake.usgs.gov/

We build models that help insurers determine whether to cover areas with increased extreme weather conditions. First, we built a risk assessment model. Considering that the property loss risk P is affected by multiple factors such as the number of extreme weather occurrences in the area and the valuation of all properties in the area, the model is subject to multi-objective constraints. Then, by observing claims loss data over the years, we use the ARIMA model to predict risk losses in future years. We offer the lowest premiums in the region based on the insurance company's ability to bear risk and remain profitable. Finally, in order to prevent companies from paying too much compensation, we compare the total compensation amount that the insurance company decides to invest in the area (without considering the investment income generated by the time value of money theory) with the predicted claim losses, and give insurance decision-making suggestions.

• Specifically, after predicting the probability of disaster occurrence, it is also necessary to predict the expected losses caused by the disaster. An intuitive approach is to directly predict the estimated number of local casualties or losses during that time period.

• Based on this, we can use the ARIMA time series model to predict compensation loss estimates for different regions each year.

We take the US disaster year losses from 1955 to 2015 as the data for ARIMA prediction; and we need to analyze the parameters and select appropriate parameters. We take $q = 2, i = 1, p = 3$; detect the PACF and ACF of this set of parameters. The results are shown in Figure 1.

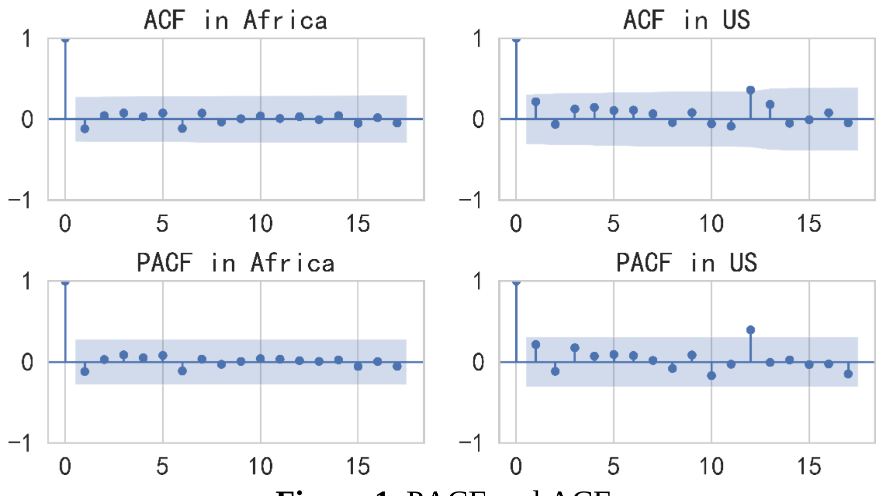


Figure 1. PACF and ACF

Figure 1. can illustrate that the ARIMA model of this set of parameters is suitable for future loss predictions in the United States and Africa. So we use the ARIMA model to predict the data; and compare it with the past.

As can be seen from Figure 2 and the forecast chart, the economic losses caused by disasters in the United States will increase in the future, while the losses caused by disasters in Africa will remain stable in the future.

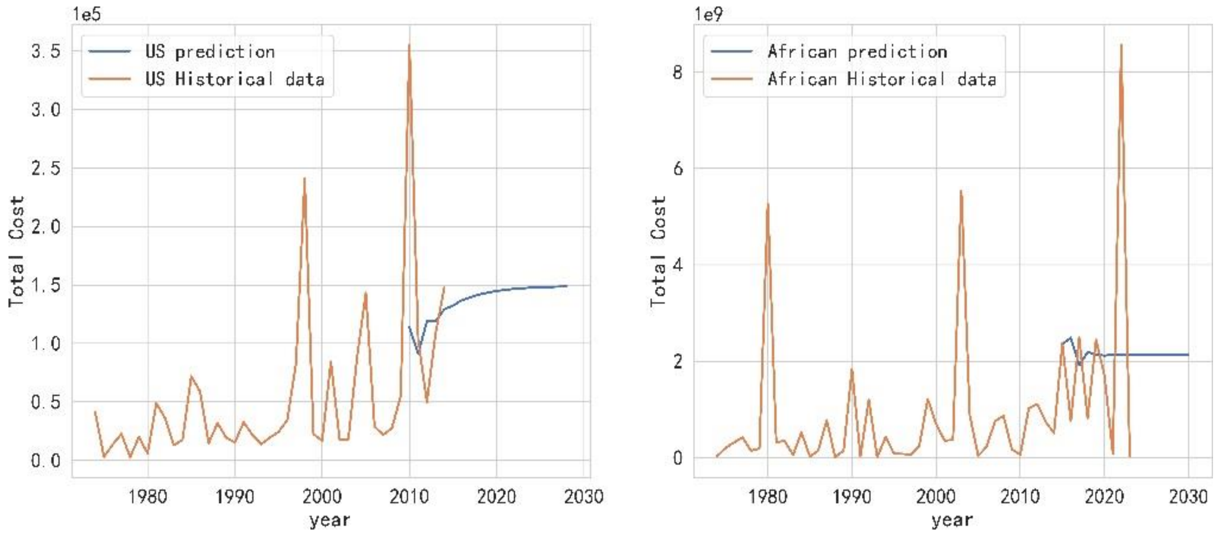


Figure 2. Forecast of economic losses from disasters in the United States and Africa

According to the following formula, we can determine whether an insurance company covers a certain area. If $pr < 0$ then do not insure.

$$Pr = N \times Ic - Ecc \tag{8}$$

$$Ecc = \left(\frac{Etl}{Pr}\right) \times N \tag{9}$$

Pr: profit, N: number of insured persons, Ecc: expected compensation cost, Etl: Expected total loss in a certain area, Pt: Total population in a certain area

The above formula means that if the risk assumed by a certain area is too high, that is, the expected compensation in that area exceeds the total premium collected by the insurance company, the insurance company will be deemed not to cover that area.

3.2. Regional risk assessment

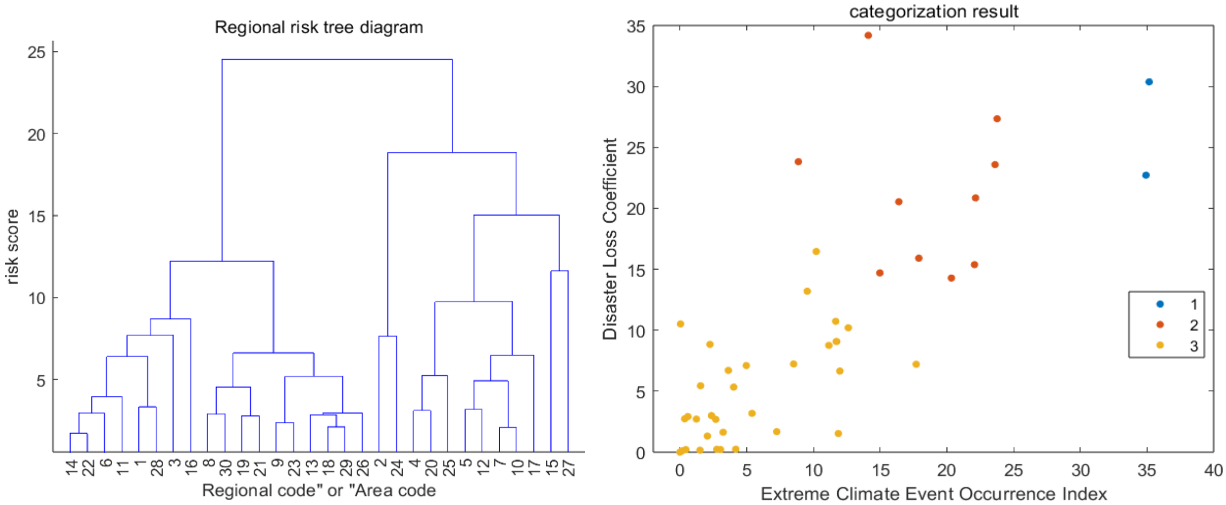
In order to intuitively select areas with smaller risks as underwriting areas; to this end, we have developed a risk assessment model to conduct risk assessments in different areas. We ignore secondary factors and mainly consider three factors: the number of extreme disasters, property compensation losses that are positively correlated with property valuations, and local casualties. Other factors can be left unchanged by default. List the following equations.

$$P = \sigma \sum_{i=1} \beta_i \times f_i + C \tag{10}$$

Among them, different β factors represent the different weights of this factor on the risk value; different f can represent the frequency of extreme weather events, disaster losses, geographical location, and artificial mitigation measures.

The number of extreme weather occurrences in a region is the sum of the number of occurrences of various disasters. Data are normalized by multiplying the number of property damage and extreme weather events by a reasonable constant for the property damage factor and hazard. Taking factors occurring in the United States as an example, the following figure can be obtained.

The number of extreme weather occurrences in a region is the sum of the number of occurrences of various disasters. Data are normalized by multiplying the number of property damage and extreme weather events by a reasonable constant for the property damage factor and hazard. Taking factors occurring in the United States as an example, the following Figure 3 can be obtained.



(a) Regional risk tree diagram (b) categorization result

Figure 3. Results of hierarchical clustering algorithm

In Figure 3, the sample in the upper right corner has more extreme weather occurrences and greater economic losses, which means that the risk in the area is greater; by calculating the similarity between different categories of data points, a nested clustering tree is obtained. We obtain the scatterplot distribution of high-risk areas by correcting values that are too far away from the dataset and analyzing the classification of high-risk areas. Through the above analysis and calculation, we concluded that "ARKANSAS" and "TEXAS" are two areas located in high-risk areas.

3.3. Premium changes under different circumstances

To help the company flexibly adjust insurance premiums in different areas based on risk profiles, because different types of disasters have different severity of economic impacts, we recommend that different types of disasters be weighed and evaluated based on the specific disaster risk profile faced by each region. Depending on the risk level of the disaster type, we can adjust the insurance premiums in the area accordingly. The impact of different types of disasters on the U.S. economic losses is shown in the Figure 4 below.

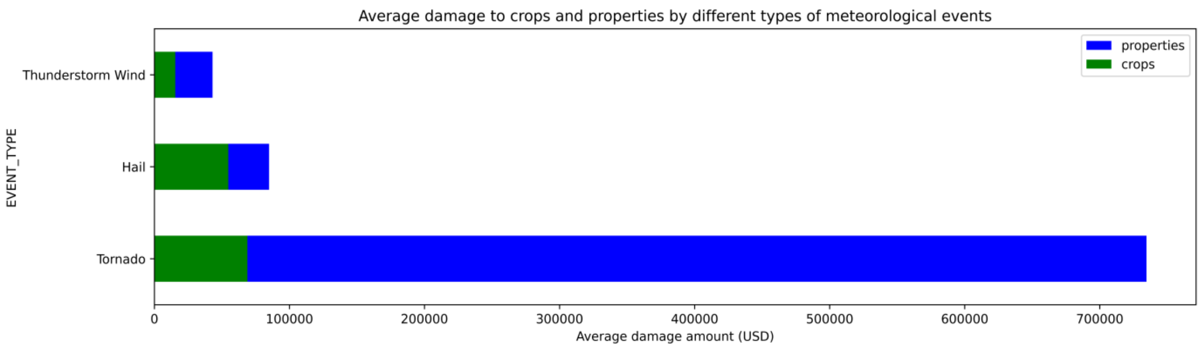


Figure 4. U.S. economic losses due to various disasters

Based on the Figure 4 above, we note that hurricanes cause the greatest economic damage. Therefore, in the risk assessment, we will focus on hurricanes and refer to hurricanes in other hazard types. Hurricanes are usually rated from F0 to F5. In order to calculate the economic losses caused by hurricanes of different levels in the United States within a year, we take Texas as an example. In this area, hurricane ratings range from F0 to F4.

In order to comprehensively consider the number of hurricanes of different levels and their impact on economic losses, we can use the entropy weight method to weigh and evalate. According to the steps of the entropy weight method, we first normalize the data and get the Phase Diagram:

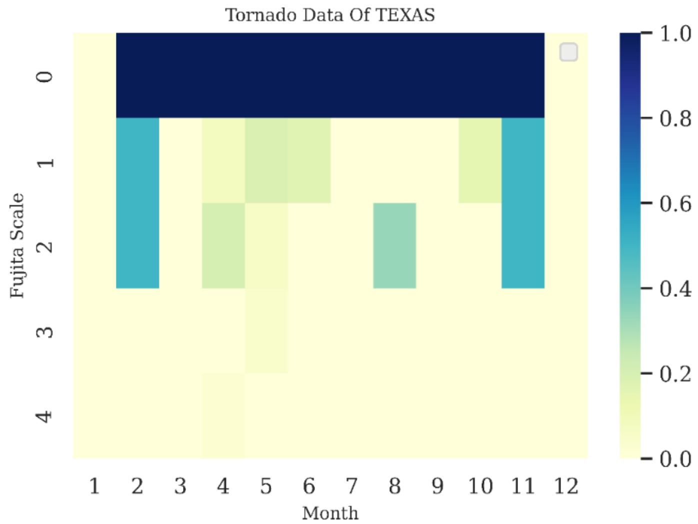


Figure 5. Phase Diagram of tornado data in Texas

From the Figure 5 above, we can get the weight of the number of occurrences of different levels of wind in each month , and the economic losses caused by one occurrence of different levels of wind. Through the following formula, we can get the risk levels caused by hurricanes in different months.

$$RH = \sum \omega_i \times L_i \tag{11}$$

According to different risk levels RH(Risk levels posed by hurricanes in different months), we can flexibly adjust the monthly insurance costs in Texas accordingly to ensure that our insurance model can enable the company's long- term healthy development. The monthly insurance cost adjustment trend chart is shown in the Figure 6.

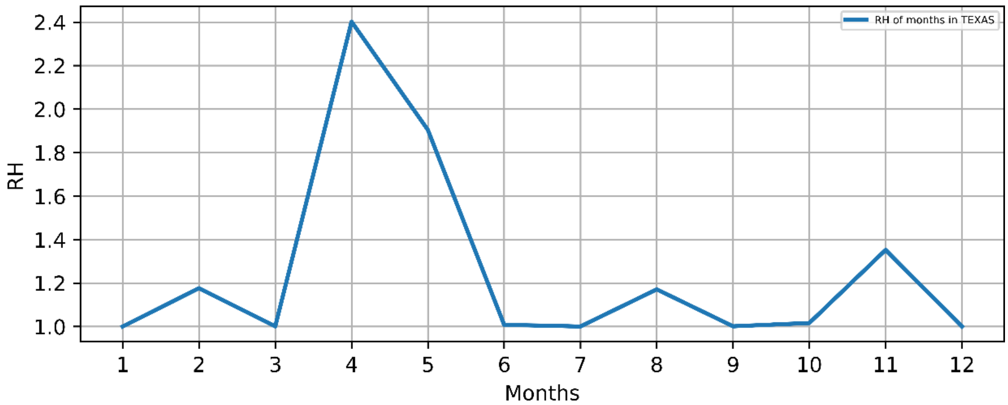


Figure 6. Insurance premium adjustment trends

The changing trend of insurance premiums is approximately consistent with that of RH in Figure 6.

4. Conclusion

In summary, we established a comprehensive model to conduct risk assessment and prediction in specific areas using methods such as the ARIMA model, the entropy weight method, and the AGNES algorithm. The model can help insurance companies determine insurance underwriting and premium adjustment strategies, and help real estate companies decide whether to pursue real estate development. By continuously adjusting and optimizing our models, we are able to more accurately assess risks, thereby providing reliable decision support for relevant industries.

References

[1] Clarke, Ben, et al. "Extreme weather impacts of climate change: an attribution perspective." Environmental Research: Climate 1.1 (2022): 012001.

[2] Nica, Aurelia, Adina Popescu, and Daniela-Cristina Ibanescu. "Human influence on the climate system." Curr Trends Nat Sci 8.15 (2019): 209 - 215.

[3] Bouwer L M. Observed and projected impacts from extreme weather events: implications for loss and damage [J]. Loss and damage from climate change: Concepts, methods and policy options, 2019: 63 - 82.

[4] Clarke B, Otto F, Stuart-Smith R, et al. Extreme weather impacts of climate change: an attribution perspective [J]. Environmental Research: Climate, 2022, 1 (1): 012001.

[5] Weil hammer V, Schmid J, Mittermeier I, et al. Extreme weather events in europe and their health consequences–A systematic review [J]. International Journal of Hygiene and Environmental Health, 2021, 233: 113688.

[6] Prokop’eva E L. The modern Russian insurance market: Problems and potential for development[J]. Finansy i kredit= Finance and Credit, 2019, 25 (1): 177 - 195.

[7] Morozova S A, Rostov A A. Development of economic and mathematical models of dynamics and analysis of insurance markets in the regions of the Volga federal district [J]. Вестник Самарского университета. Экономика и управление, 2023, 14 (1): 192 - 201.

[8] Hohl R, Jiang Z, Vu M T, et al. Using a regional climate model to develop index-based drought insurance for sovereign disaster risk transfer[J]. Agricultural Finance Review, 2020, 81 (1): 151 - 168.

[9] Awe O, Okeyinka A, Fatokun J O. An alternative algorithm for ARIMA model selection[C]//2020 international conference in mathematics, computer engineering and computer science (ICMCECS). IEEE, 2020: 1 - 4.

[10] Benvenuto D, Giovanetti M, Vassallo L, et al. Application of the ARIMA model on the COVID-2019 epidemic dataset [J]. Data in brief, 2020, 29: 105340.

[11] Khan S, Alghulaiakh H. ARIMA model for accurate time series stocks forecasting [J]. International Journal of Advanced Computer Science and Applications, 2020, 11 (7).

[12] Zhu Y, Tian D, Yan F. Effectiveness of entropy weight method in decision-making [J]. Mathematical Problems in Engineering, 2020, 2020: 1 - 5.