ISSN: 2424-8460(O)
2251-2608(P)

Original Research Article

Analysis and statistical prediction of earthquake disaster impacts based on random forest algorithm

Yuxin Ji
University of Science and Technology Liaoning, Anshan, Liaoning, 114051, China

Abstract: Earthquakes are severe natural disasters. Accurate analysis and prediction of disaster impacts are vital for disaster prevention and emergency response. Based on domestic and international earthquake data over the past two decades, this paper builds a prediction model with the random forest algorithm. Using descriptive statistics and correlation analysis, key influencing factors are identified. Model performance is verified via actual disaster data. The results show that the random forest algorithm achieves high accuracy and stability in predicting earthquake casualties and economic losses. Compared with traditional statistical methods, it has strong anti-interference ability and good generalization. The model can provide technical support for earthquake early warning, disaster assessment and emergency resource allocation.
Keywords: random forest; earthquake disaster; disaster prediction; influencing factors; statistical analysis

1. Introduction

Earthquakes are extremely destructive natural disasters that often trigger a series of chain reactions including casualties, economic losses, social disorder and public panic [1]. Historical events such as the Wenchuan (2008), Yushu (2010), and Ludian (2014) earthquakes highlight the destructive potential of seismic activity, especially in densely populated regions with vulnerable infrastructure.

Traditional earthquake analysis methods rely heavily on empirical judgment and statistical summaries, which are limited in handling complex nonlinear relationships among influencing factors. With the development of machine learning, data-driven models have become increasingly important in disaster prediction tasks.

This study applies the Random Forest algorithm to multi-source earthquake datasets to identify key influencing factors and predict disaster outcomes, including casualties and economic losses. The model's performance is validated through quantitative evaluation and feature importance analysis.

2. Research methods and data processing

2.1. Overall research process

This study uses earthquake records from global and domestic sources over the past twenty years, including magnitude, epicenter location, population density, building characteristics, casualties, and economic loss.

Data preprocessing includes:

(1) Removal of missing and abnormal values.

(2) Data imputation using decision-tree-based estimation.

(3) Normalization to eliminate dimensional differences.

(4) 7:3 split for training and testing, with cross-validation for robustness.

2.2. Sample selection and data preprocessing

Descriptive statistics are used to analyze distributions of casualties and losses under different seismic conditions. Correlation analysis indicates that:

(1) Magnitude is strongly positively correlated with disaster severity.

(2) Population density increases casualty risk.

(3) Earthquake-resistant structures significantly reduce losses.

ISSN: 2424-8460(O)
2251-2608(P)

(4) Epicenter location influences spatial risk distribution.

These variables interact jointly rather than independently, forming a complex nonlinear system.

2.3. Descriptive statistical analysis

Raw Earthquake disaster assessment is a systematic and comprehensive analytical work involving multi-dimensional considerations. Subject to data limitations, this study mainly conducts statistical research from three core dimensions: Affected population, building damage degree and traffic facility damage, and obtains comprehensive assessment results accordingly [2]. The statistical results indicate that higher earthquake magnitudes and more complex seismic source conditions correspond to larger death tolls, injuries and direct economic losses. In regions with dense population and a high proportion of non-earthquake-resistant buildings, disaster losses will be further aggravated. Statistical indicators such as frequency, mean value and extreme values clearly demonstrate the distribution characteristics of disaster losses under different seismic conditions.

2.4. Correlation analysis

Four core indicators including earthquake magnitude, epicenter, population density and building type are analyzed for their correlations with disaster losses. Earthquake magnitude has a strong positive correlation with losses and is the dominant factor of damage severity. Areas in seismic belts suffer more frequent earthquakes and higher long-term risks. The proportion of earthquake-resistant buildings is negatively correlated with losses, as building seismic performance reduces earthquake damage. Higher population density results in more casualties and indirect economic losses. These factors interact jointly to affect overall disaster impacts.

2.5. Principle of random forest algorithm

With technological advances, machine learning has gradually replaced pattern recognition as the prevailing term in this field and evolved into an independent discipline. Its core lies in identifying underlying patterns by analyzing known data, then applying such patterns to make predictions and judgments on unknown data[3]. Random forest is an ensemble learning algorithm composed of multiple independent decision trees. It uses sample and feature random sampling to avoid overfitting, and adopts majority voting to produce final predictions.

The algorithm is robust and requires no strict data assumptions. It adapts well to highly random, multi-dimensional seismic data and provides feature importance to interpret influencing factors.

3. Establishment and solution of the model

3.1. Definition and theory of seismic disaster risk assessment

3.1.1. Definition of risk

Risk is defined as a state with uncertain consequences for things of human value[4]. Risk has three fundamental characteristics. First, it refers to an objectively existing state. Second, such a state may trigger uncertain outcomes. Third, risk must act upon entities that hold value to human beings, and uncertainty serves as the core element of the risk concept[5].

3.1.2. Seismic disaster risk assessment model

Define risk R as the product of occurrence probability P and loss consequence C, :

$$R=P\cdot C \tag{1}$$

We grade P and C to establish a risk matrix for quick assessment (**Table 1**).

Table 1. Risk matrix.

	LOW CONSEQUENCE	MEDIUM CONSEQUENCE	HIGH CONSEQUENCE
High Probability	Medium Risk	High Risk	Very High Risk
Medium Probability	Low Risk	Medium Risk	High Risk
Low Probability	Very Low Risk	Low Risk	Medium Risk

The random forest model, built on multiple independent decision trees with dual random sampling and majority voting, is trained on the dataset. It features strong robustness and generalization, achieving high prediction accuracy. Relevant figures visualize the contribution of each feature to disaster losses[6].

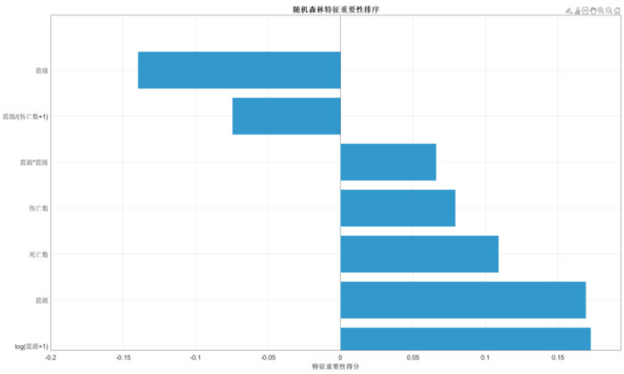

Figure 1. Feature importance.

The horizontal axis denotes feature importance scores, while the vertical axis lists indicators from top to bottom as follows: Magnitude, magnitude/(casualties+1), epicenter×magnitude, casualties, death toll, epicenter, log(epicenter+1). This illustrates the relative importance of each characteristic (independent variable), showing that the larger the proportion of the epicenter, the higher the likelihood of an event occurring.

4. Results and discussion

Statistical Result Analysis of Earthquake Disaster Prediction Model.

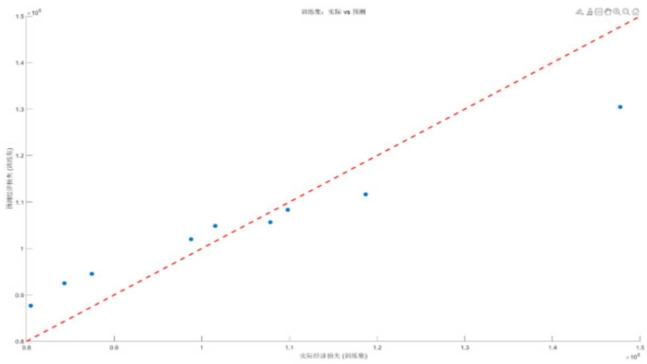

Figure 2. Comparison between actual values and predicted values.

The unit of the vertical coordinate is multiplied by 106, and this graph illustrates the correlation between actual values and predicted values on the training set. It can be seen from the figure that the predicted values are relatively close to the true values, which is consistent with our expected assumptions.

Table 2. Training data.

EPICENTER	MAGNITUDE	CASUALTIES	DEATH TOLL	TOTAL ECONOMIC LOSS
24	7	593	362	817200
25	9	422	226	852200
62	6.5	637	360	1180400
26	10	929	1800	1485800
50	6.6	636	293	1045100
65	6.5	413	244	1130800
41	7.1	3445	2244	2127200
25	6.2	845	331	828300
22	6.8	1598	656	1076400
25	6.4	1123	598	974000
21	6.7	1227	663	989300
28	8	1223	578	1098000
56	6.9	562	3411	2040700

ISSN: 2424-8460(O)
2251-2608(P)

5. Conclusion

This study constructs an earthquake disaster prediction model using the random forest algorithm. Combined with statistical analysis, the model captures nonlinear relationships between seismic factors and disaster losses.

Compared with traditional methods, it has better robustness, interpretability and prediction accuracy. Key influencing factors such as magnitude, population density, and structural resistance are quantitatively identified.

The model is limited in predicting extreme earthquakes and micro variations. Future research will use refined data and optimized models to improve prediction performance.

About the author

Yuxin Ji (2004.03-), female; Han Chinese; hometown: Yucheng County, Henan Province; educational level: Undergraduate student; research direction: Disaster big data analysis and intelligent prediction.

References

[1] Li D P, Li M, Yin J F, et al. Research on Earthquake Disaster Scenario Construction Strategy from the Perspective of Social Perception[J]. China Emergency Rescue, 2024(6): 16-22. DOI:10.19384/j.cnki.cn11-5524/p.2024.06.012.

[2] Ai R W. Research on Comprehensive Statistical Analysis Technology and Method of Earthquake Emergency Survey Disaster Information[D]. Fuxin: Liaoning Technical University, 2014.

[3] Zhang P. Research on Post-earthquake Building Damage Assessment Model and Ground Motion Characteristic Analysis Method Based on Machine Learning[D]. Guangzhou: Guangzhou University, 2023. DOI:10.27040/d.cnki.ggzdu.2023.002187.

[4] Yu X. Basic Theoretical Research on Risk Assessment of Major Geotechnical Engineering[D]. Beijing: Institute of Engineering Mechanics, China Earthquake Administration, 2018.

[5] Li X A. Research on Earthquake Disaster Risk Assessment in Gansu Province[D]. Langfang: Institute of Disaster Prevention, 2021.

[6] Lin H, Du J L, Xu G, et al. Application of Random Forest Algorithm in Casing Deformation Prediction of Hydraulic Fracturing[J/OL]. Lithologic Reservoirs:1-9[2025-04-09]. http://kns.cnki.net/kcms/detail/62.1195.TE.20250211.1049.002.html.