

Impacts of Urban Three-Dimensional Morphology on Seismic Vulnerability and its Driving Mechanisms: A XGBoost-SHAP Explainable Analysis in Jiangyou City, Southwestern China

Yaohui Liu

liuyaohui@aitech.edu.cn

Aerospace Information Technology University

Bingren Zhou

Shandong Jianzhu University

Hao Zheng

Beijing Normal University

Article

Keywords: Three-dimensional Morphology, Seismic Vulnerability, XGBoost, SHAP, Pearson Correlation

Posted Date: September 2nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10781691/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

1 **Impacts of Urban Three-Dimensional Morphology on Seismic**
2 **Vulnerability and its Driving Mechanisms: A XGBoost-SHAP**
3 **Explainable Analysis in Jiangyou City, Southwestern China**

4 Yaohui Liu^{a*}, Bingren Zhou^b and Hao Zheng^c

5 *^a School of Remote Sensing Science and Technology, Aerospace Information*
6 *Technology University, Jinan 250200, China; ^b School of Surveying and Geo-*
7 *Informatics, Shandong Jianzhu University, Jinan 250101, China; ^c Key Laboratory of*
8 *Environmental Change and Natural Disaster, Ministry of Education, Beijing Normal*
9 *University, Beijing 100875, China*

10 ● * Correspondence author: liuyaohui@aitech.edu.cn

11 ● ORCID

12 ● Yaohui Liu: 0000-0002-3041-3557

13 ● Hao Zheng: 0000-0003-1504-0120

14 ● Bingren Zhou: 0009-0002-5550-169X

15 ● **Funding**

16 This study was supported in part by the National Natural Science Foundation of
17 China [grant number 42201077]; the Shandong Provincial Natural Science Foundation
18 [grant number ZR2021QD074]; the China Postdoctoral Science Foundation [grant
19 number 2023M732105]; the Youth Innovation Team Project of Higher School in
20 Shandong Province, China [grant number 2024KJH087].

21

Impacts of Urban Three-Dimensional Morphology on Seismic Vulnerability and its Driving Mechanisms: A XGBoost-SHAP Explainable Analysis in Jiangyou City, Southwestern China

Abstract

Understanding the interaction between urban three-dimensional morphology (urban 3D morphology) and seismic vulnerability is crucial for optimizing urban spatial planning and disaster risk reduction. However, existing seismic vulnerability assessments have primarily focused on individual buildings or conventional structural and socioeconomic factors, with limited attention to the nonlinear relationships between urban-scale 3D morphological configurations and vulnerability patterns. To address this limitation, this study develops an integrated and explainable assessment framework that combines XGBoost, SHAP analysis, Pearson correlation-based feature selection, and Moran's I spatial clustering to quantify the contributions of urban 3D morphological characteristics and reveal their spatial heterogeneity in relation to building seismic vulnerability. Results show that among all feature categories, Spatial Crowding Degree is the most influential feature in the model, returning the widest range of SHAP value variations and the highest average influence. The influence of different urban 3D characteristics on seismic vulnerability exhibits diverse patterns, in which high-density and compact layouts can reduce vulnerability, while features such as building height variation and plot ratio significantly affect prediction contributions within specific ranges. Moran's I-based spatial clustering shows that High-Low and Low-High areas reflecting transitional patterns and spatial spillover effects. The results of this study provide an important basis for improving urban seismic resilience, improving disaster prevention capacity, and sustainable urban development.

Keywords: Three-dimensional Morphology; Seismic Vulnerability; XGBoost; SHAP; Pearson Correlation .

1 Introduction

With the rapid acceleration of global urbanization, the scale and density of urban construction have continued to increase, and urban spatial layouts and building morphologies have become

increasingly complex (Łabuz, Blazy and Hess, 2026; Liu et al., 2026). While these changes have promoted urban development, they have also intensified the potential risks cities face when natural disasters—particularly earthquakes occur (Yadim et al., 2026). As one of the most destructive natural hazards, earthquakes pose serious threats to urban building structures, infrastructure systems, and the safety of human lives and property, and therefore have attracted widespread attention from both academia and urban management sectors. In recent years, accurately assessing the seismic vulnerability of urban buildings through scientific approaches has become an important issue in urban planning and disaster risk management. In this context, urban three-dimensional morphological (3D) characteristics, such as building density, plot ratio, and building height, are considered key factors influencing seismic vulnerability (Chen et al., 2026; Chen, 2013; Maiorana, Conte and Briseghella, 2026; Moroz and Thieken, 2026). Investigating the relationships between these spatial morphological characteristics and seismic vulnerability is therefore essential for optimizing urban spatial structure, enhancing urban seismic resilience, and disaster risk reduction.

Seismic vulnerability refers to the intrinsic susceptibility of buildings or building groups to experience damage under a given level of seismic ground motion, as determined by their structural and morphological characteristics.(Wang et al., 2026) As a core component of seismic disaster risk assessment, seismic vulnerability analysis has been widely studied. Early studies primarily relied on post-earthquake damage surveys combined with expert judgment to assess building vulnerability, making this approach one of the most straightforward and practical methods in vulnerability analysis. It also allows for the reuse of findings from previous studies when analyzing similar building types, thereby ensuring a degree of continuity and comparability. This method can effectively capture actual damage patterns and reflect the true seismic performance of structures (He et al., 2025; Li and Xu, 2026). However, several limitations constrain its application. Damage data are often difficult to obtain across varying intensity levels and site conditions, resulting in relatively small sample sizes. In addition, data collected from remote areas may lack representativeness due to differences in construction practices and local customs, and therefore may not accurately reflect the seismic performance of buildings designed according to current codes. Furthermore, the data collection process is susceptible to subjective influences, and the method is generally applicable only to regions and structural types with abundant damage records, which limits its broader applicability.

As the limitations of damage survey-based empirical approaches become evident and the demand for accurate seismic damage prediction increases, numerical simulation-based methods are gaining prominence. Common approaches include regression analysis and the Analytic Hierarchy Process (AHP) (Malakar and Rai, 2023). Regression analysis quantifies the

effects of multiple factors by modeling relationships among variables and is characterized by a clear structure, computational efficiency, and good interpretability. In contrast, AHP decomposes complex problems into hierarchical structures and assigns weights to multiple criteria through expert-driven judgment, enabling systematic evaluation of multi-criteria decision problems. However, with increasing urban complexity, these methods show clear limitations. Regression models are typically based on linear assumptions and may fail to capture nonlinear relationships among variables, whereas AHP depends on expert judgment, which can introduce subjectivity and reduce the robustness of the results.

In recent years, the rapid development of machine learning techniques has enriched the methodological approaches for seismic vulnerability assessment and significantly improved prediction accuracy (Asadollahzadeh and Behnam, 2025; Zain, Dackermann and Prasittisopin, 2024). Machine learning models such as Support Vector Machine (SVM) (Wenyi et al., 2024; Yaohui et al., 2023), Random Forest (RF) (Sun et al., 2025), and Extreme Gradient Boosting (XGBoost) (Zhou et al., 2025) have become important tools in this field due to their ability to handle nonlinear relationships and complex high-dimensional data. Specifically, SVM demonstrates strong generalization ability in small-sample and high-dimensional datasets (Manikandan et al., 2025); RF enhances model stability and reduces overfitting through ensemble learning (Kallas and Napolitano, 2025); and XGBoost provides high computational efficiency and prediction accuracy, performing well in modeling complex feature interactions (Pinasco et al., 2025). Overall, machine learning approaches can uncover hidden patterns in data with fewer prior assumptions, thereby improving predictive performance (Aloisio et al., 2026; Chu et al., 2025; Gondaliya et al., 2024; Zhang et al., 2026). Nevertheless, these methods still present some challenges in practical applications, including dependence on high-quality data, complex parameter tuning processes, and limited interpretability of certain models (Chicco, Chieffo and Formisano, 2024; Eren and Şalgamcıoğlu, 2025; Yang et al., 2023; Zhang et al., 2026).

To further improve model interpretability, explainable artificial intelligence (XAI) techniques have recently been introduced into seismic vulnerability studies. Furthermore, spatial statistical methods provide important tools for identifying spatial distribution patterns of vulnerability (Luo, Zhao and Zeng, 2026; Pinasco et al., 2025; Weiyi and Limao, 2022; Yang et al., 2025). Among them, Moran's I, a widely used indicator of spatial autocorrelation, can effectively detect spatial clustering patterns of seismic vulnerability and reveal the distribution of high-value and low-value areas (Davtalab and Byčenkienė, 2026; Islam et al., 2021; Liu et al., 2022; Nishizaki, 2017; Papadopoulos and Karalis, 2023; Zhu et al., 2026). Furthermore, by integrating interpretability tools such as Shapley Additive Explanations (SHAP) and Partial

Dependence Plot (PDP), it becomes possible to analyze the influence mechanisms of urban 3D characteristics on seismic vulnerability from both global and local perspectives (Dong et al., 2022; Tian et al., 2023; Zhang et al., 2020; Zhi et al., 2017; Zhou et al., 2016).

In this study, we propose a comprehensive framework that combines spatial statistical analysis with explainable machine learning methods to investigate the relationship between urban 3D characteristics and seismic vulnerability. First, Variational Autoencoder (VAE) is introduced for data augmentation. By generating a large number of new samples based on the original dataset, the robustness and generalization ability of the model are significantly improved. Second, Pearson correlation analysis is employed to systematically explore the relationships between urban 3D characteristics—such as building density, plot ratio, and building height—and seismic vulnerability, thereby revealing the roles of these features in influencing seismic vulnerability. On this basis, XGBoost regression model is constructed to establish a more accurate and reliable seismic vulnerability assessment model. Meanwhile, SHAP value visualization techniques are applied to conduct an in-depth interpretation of the influence of different features in the seismic vulnerability assessment process and to further quantify the marginal contributions of each feature, thereby revealing the underlying influence mechanisms of urban morphological characteristics. In addition, Moran's I is used to identify the spatial clustering patterns of seismic vulnerability and to further reveal the spatial distribution characteristics of urban morphological features. The results of this study can provide a scientific basis and technical support for optimizing urban spatial planning, enhancing urban seismic resilience, and developing more effective disaster risk reduction strategies.

2 Study Area and Data Sources

2.1 Study Area

Jiangyou City (31°47'N, 104°44'E), located in Sichuan Province, southwestern China, is characterized by a predominantly residential building stock, where masonry structures are dominant, supplemented by a smaller proportion of reinforced concrete buildings and industrial frame structures. The urban form is mainly composed of low- and mid-rise buildings, with relatively few high-rise structures, resulting in a spatial pattern dominated by low- and mid-rise masonry buildings, which contrasts sharply with megacities where high-rise reinforced concrete buildings prevail (Figure. 1). Situated in the central segment of the North–South seismic belt and within the Longmenshan fault zone, the region is prone to frequent seismic activity. During the catastrophic Wenchuan earthquake on 12 May 2008 (Mw 8.0), the seismic intensity in

Jiangyou reached VIII on the local intensity scale. The earthquake caused more than 400 fatalities, the relocation of approximately 365,923 residents, and direct economic losses of about 59.2 billion RMB in Jiangyou. Therefore, assessing the seismic vulnerability of buildings is essential for improving seismic risk mitigation and providing a scientific basis for protecting human life and reducing economic losses.

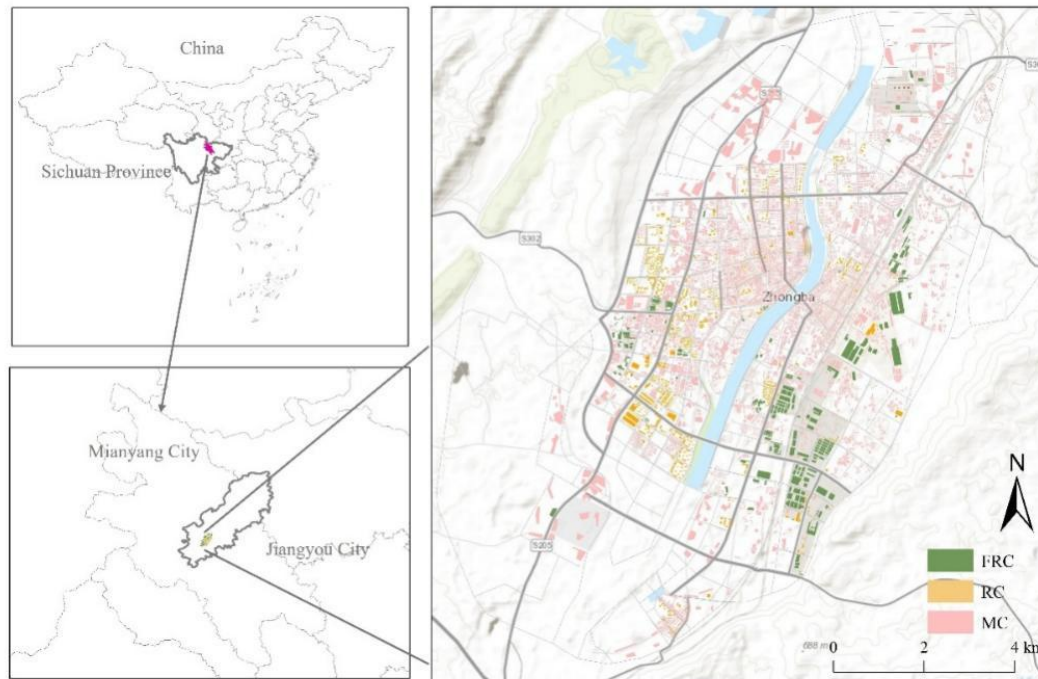


Figure 1. The study area of Jiangyou City (MC—masonry concrete structure, RC—reinforced concrete structure, FRC—factory buildings with reinforced concrete columns)

2.2 Data Sources

In this study, basic building data were obtained using tools such as Google Maps and ArcGIS Online, primarily including key attributes such as building height and number of floors. After organizing and classifying the relevant parameters, all validated data were integrated into a GIS database. The study further divided the study area into 281 spatial units based on block and community spatial patterns, and conducted all subsequent analyses at this scale. The spatial database of Jiangyou City was first established in ArcGIS by integrating neighbourhood boundaries and road network data derived from Google Maps. Building age and seismic design information were subsequently interpreted from historical remote sensing imagery, and structures were grouped into three categories based on the implementation periods of China's seismic design standards: before 2001, 2002–2010, and after 2010. The corresponding numbers

of buildings in these categories were 1,200, 800, and 500, respectively. Building structural characteristics and maintenance conditions were then obtained through a combination of Baidu Street View interpretation and field investigation records. Three dominant structural systems were identified, including reinforced concrete (RC), frame-reinforced concrete (FRC), and masonry construction (MC). The physical condition of buildings was evaluated and classified into four levels (good, slightly damaged, moderately deteriorated, and poor) based on observable indicators such as surface damage and moisture-related deterioration. In addition, circular buffer zones with radii of 50 m and 200 m were generated around mapped active faults and potential concealed faults to assess the potential influence of fault proximity on building vulnerability. All collected building-related attributes were integrated into a unified GIS database for subsequent analysis. To incorporate the effects of construction quality, structural configuration, material strength, and maintenance status on seismic damage assessment, the EMS-92 scale proposed a vulnerability classification table that links building types with corresponding vulnerability classes for intensity evaluation. The subsequent EMS-98 scale further refined this framework by improving the classification of building vulnerability and enhancing the reliability of earthquake intensity assessments.

The buildings in Jiangyou City are relatively concentrated, forming diverse three-dimensional spatial configurations and heterogeneous urban morphological characteristics. Significant variations exist among buildings in terms of physical attributes such as urban development intensity and building exposure, vertical urban morphology, variation in building volume, and building geometric morphology. Urban development intensity and building exposure characterize the concentration and spatial arrangement of buildings, which are potentially related to building-to-building interactions, population exposure levels, and the complexity of post-earthquake emergency responses. Vertical urban morphology captures variations in building height and vertical configuration, reflecting possible differences in seismic response behaviour among buildings. Building volume variability represents the diversity of building sizes within urban areas and may contribute to variations in structural complexity and interaction patterns of building clusters under seismic loading. Building geometric morphology describes the characteristics of building shape, scale, and spatial arrangement, which are associated with the distribution of seismic responses and the organization of urban space. Collectively, these morphological dimensions establish a conceptual framework for exploring the potential relationships between urban form

characteristics and seismic vulnerability. The graphical definitions of these indicators are provided in Table 1.

In recent years, a growing number of quantitative indicators have been developed to describe urban 3D building morphology and characterize urban spatial structure. Previous studies have demonstrated that these three-dimensional morphological metrics significantly influence a variety of geospatial processes, including the seismic vulnerability of urban buildings. With the advancement of researches, the critical role of building spatial form in urban earthquake disaster response has been increasingly recognized. Recent studies further indicate that these indicators have a substantial impact on regional building seismic performance and the spatial distribution of earthquake losses.

Table 1. Definitions of Urban 3D Morphological Indicators

Category	Indicators	Description	Formula
Urban development intensity and building exposure	Plot Ratio (PR)	Urban building development intensity and spatial utilization level	$PR = \frac{\sum A_i \times F_i}{A}$
Average Building Area (BA)	Average Building Area (BA)	Building footprint scale	$BA = \frac{\sum A_i}{N}$
Building Density (BD_1)	Building Density (BD_1)	Building volume scale level	$BD_1 = \frac{\sum V_i}{N}$
Spatial Crowding Degree (SCD)	Spatial Crowding Degree (SCD)	Degree of crowding of buildings in three-dimensional space	$SCD = \frac{\sum V_i}{H_{max} \times A}$
Average Height (AH)	Average Height (AH)	Overall building height level in the region	$AH = \frac{\sum H_i}{N}$
Maximum Height (H_max)	Maximum Height (H_max)	Upper limit of building height in	-

the region			
Variation in building volume	Standard Deviation of Building Height (H_std)	Dispersion of building heights	$H_std = \sqrt{\frac{1}{N} \sum (V_i - \bar{V})^2}$
	Coefficient of Height Variation (CH)	Uniformity of urban height distribution	$CH = \frac{\sqrt{\frac{1}{N} \sum (V_i - \bar{V})^2}}{\frac{\sum H_i}{N}}$
	Variation of Building Height (V_BH)	Differences in building scale	$V_BH = \frac{1}{N} \sum (H_i - \bar{H})^2$
	Variation of Building Volume (V_BD)	Differences in building volume	$V_BD = \frac{1}{N} \sum (V_i - \bar{V})^2$
	Building geometric morphology	Proportional characteristics of building morphology	$BSI = \frac{\sum A_i \times H_i}{N}$
	Building Structure Index (BSI)		

222

223 3 Methodology

224 The study explores the impact of the urban morphology on the variation of the seismic
225 vulnerability index. As illustrated in Figure.2, the comprehensive approach for
226 decoupling the mechanism between urban morphology and building vulnerability index
227 can be classified into three steps: (1) Employing the Variational Autoencoder (VAE)
228 data augmentation technique to expand the original dataset to 7,236 entries for model
229 training and validation (Liu et al., 2022). (2) Selecting urban morphological features
230 with high correlation to vulnerability by the Pearson correlation coefficient method
231 (Davtalab and Byčenkienė 2026). (3) Evaluating the robustness and accuracy of
232 XGBoost (Dong et al., 2022; Muhammad et al., 2021; Zhang et al., 2020) based on
233 comparative analysis with SVR (Research et al., 2019), RF (Adesina et al., 2022) and
234 DT(Abbas et al., 2023); (4) Applying the SHAP and PDP methods to explain the

influence degree of urban morphology indicators on building vulnerability and the interdependencies among them (Gawde et al., 2024; Gondaliya et al., 2024). (5) Using the Moran's I method to identify the optimal intervals and spatial heterogeneity of urban morphological characteristics under different scenarios (Chen, 2013; Jackson et al., 2010).

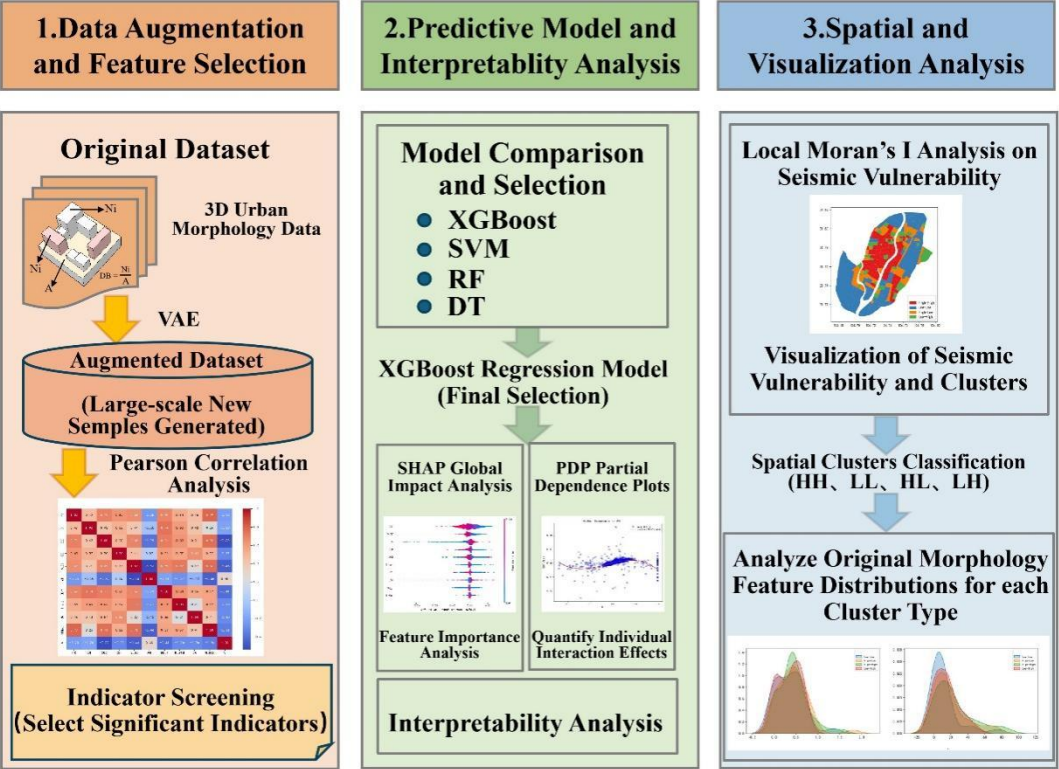


Figure 2. Framework of the study

3.1 The modified VIM

The VIM method is an important component of the RISK-UE project, which describes the degree of damage caused by seismic events to buildings through several important parameters that affect the seismic performance of buildings. In this study, the modified VIM is employed, which is divided into the following three main calculation steps:

Estimate the vulnerability index $V_l^{building}$ denotes the vulnerability index of the building structure, and its value ranges from 0 to 1. Reduced vulnerability is indicated by a smaller value. The VIM approach evaluates building seismic vulnerability based on the Building Type Matrix (BTM), which classifies buildings according to their structural systems and physical characteristics. By integrating extensive earthquake damage records, the method establishes relationships between different building types

and their corresponding damage responses under various seismic intensities. For Sichuan Province, where the study area is located and has experienced significant earthquake impacts, previous studies have compiled damage information from multiple historical earthquakes and developed damage matrices for different structural categories. These matrices were combined with the average loss ratios associated with each building type to derive the corresponding vulnerability indices V_I for the three major structural systems, as shown in the following formulation(1).

The value of the vulnerability index also depends on various modification factors. In the following formula, ΔM_R is the modification considering regional characteristics and construction age, while V_{m_j} is the modification including other aspects that affect the seismic performance of the building. Regional modification factors were incorporated to improve the applicability of the vulnerability assessment in the local seismic context. The selected factors included construction period, building–fault distance, building–fault orientation, number of floors, and location effects. The construction period factor followed the RISK-UE framework, considering the influence of seismic code evolution, construction practices, and material deterioration. The fault-related factors were determined with reference to Chinese seismic planning and geotechnical standards (GB 50413-2007 and GB 50021-2001), which emphasize the influence of near-fault zones (e.g., 50–200 m) on seismic response. Building–fault orientation was considered to account for potential directional ground-motion and torsional effects. The selected factors were further evaluated through sensitivity analysis, showing significant contributions to vulnerability variation (22–35% variance explained, ANOVA, $p < 0.01$). The modified parameter values are provided in Table 2.

$$V_I^{building} = V_I + \Delta M_R + \sum_{j=1}^n V_{m_j} \quad (1)$$

Table 2. Previous research results on the vulnerability of buildings in Sichuan Province.

Vulnerability factors			Code level	
V_{m_j}	1.Number of floors	MC	RC	FRC
	Low	−0.02	−0.04	−0.04

	Medium	0		0		0	
	High	+0.06		+0.06		+0.06	
	2.The angle	Parallel		Other		Perpendicular	
	between the	+0.06		+0.04		+0.01	
	building and the						
	fault						
	3.Location	1	2	3	4	5	6
	modifiers						
		+0.04	+0.04	−0.02	−0.02	0	+0.02
ΔM_R	4.Construction	Pre-2001		2001–2010		Post-2010	
	age						
		Low code		Medium code		High code	
		+0.16		0		−0.16	
	5.Distance						
	between fault						
	and building						
	(m)						
	Fault	200 (m)		200 (m)		200 (m)	
	Buried fault	50 (m)		50 (m)		50 (m)	

3.2 Data Augmentation Based on Variational Autoencoder

Data Augmentation Based on VAE is an unsupervised generative model based on variational Bayesian inference, which consists of an encoder and a decoder implemented by single-layer feedforward networks. By virtue of linear transformations combined with non-linear activation functions, it jointly learns the approximate mapping from raw data to the latent space distribution as well as the non-linear generative process from the latent space to reconstructed data. This approach is particularly useful in generating realistic synthetic data while preserving the underlying distribution and characteristics of the original dataset, which is crucial for training robust machine learning models in urban morphology indicators studies. VAE typically assumes that the input variable is generated by the random latent variable , and its

marginal likelihood function can be expressed as:

$$p(x) = \int p(z)p_{\alpha}(x|z)dz \quad (2)$$

Among them, the generative $p_{\alpha}(x|z)$ model represents the reconstruction of the latent variable z into the input variable x . The prior distribution $p(z)$ is usually assumed to follow a standard normal distribution. According to Bayes' theorem, the posterior distribution of the latent variable $p(z|x)$ can be expressed as:

$$p(z|x) = \frac{p_{\alpha}(x|z)p(z)}{p(x)} \quad (3)$$

However, since both the parameters α of the generative model and the latent variable z are unknown, the marginal likelihood of the observed data is difficult to calculate directly. Therefore, the VAE employs variational inference to obtain an approximate probability distribution $p(z|x)$ of the posterior distribution $q_{\varphi}(z|x)$. The distribution $q_{\varphi}(z|x)$ can be referred to as the inference model, which is described by a multivariate normal distribution where φ denotes the parameters of the inference model. The fundamental idea of the VAE is to utilize the inference model and the generative model to approximate the posterior distribution $p(z|x)$ of the latent variable via variational inference, thereby achieving effective modeling of complex data.

The model structure of the VAE is illustrated in Figure 3. First, the input variable x undergoes dimensionality reduction through a linear layer to yield the bottleneck layer h . To ensure the validity of latent variable sampling, each dimension in the bottleneck layer corresponds to a specific posterior distribution. This process is implemented by two parallel encoders, which are used to estimate the mean μ and variance σ^2 of the posterior distribution for each dimension, respectively. This encoding process can be described by $q_{\varphi}(z|x)$, whose essence lies in mapping high-dimensional input data to a low-dimensional latent space and completing the extraction of data features. The latent variable z , randomly sampled from the latent space, is fed into the decoder $p_{\alpha_x}(x|z)$ to reconstruct the input data, where z represents the feature representation and information transfer of the input data. During the training process, the VAE updates the neural network parameters by minimizing the following loss function:

$$J_{VAE} = - \int_Z q_\phi(z|x) \log(p_{\alpha_x}(x|z)) dz + KL(q_\phi(z|x)||p(z)) \quad (4)$$

Among these terms, the first one is the reconstruction term, which measures the discrepancy between the input data and the reconstructed data and ensures that the latent variables can effectively capture the features of the input data. The second term is the Kullback-Leibler (KL) divergence, which quantifies the difference between the encoder distribution and the prior Gaussian distribution.

Given the limited number of original spatial units, a variational autoencoder (VAE) was employed to augment the dataset and support subsequent machine-learning model training. The VAE used an encoder–decoder architecture with 32 neurons in the hidden layers and an 8-dimensional latent space. ReLU and Sigmoid activation functions were used in the hidden and output layers, respectively. The model was trained using the Adam optimizer with a learning rate of 0.001 for 50 epochs, using a loss function combining reconstruction loss and Kullback–Leibler divergence.

The generated samples were treated as augmented representations learned from the original 281 spatial units rather than as newly observed geographic units. Thus, the expansion to 7,236 observations does not represent an equivalent increase in independent spatial units or physical observations. The subsequent XGBoost-SHAP analysis was therefore interpreted in terms of the feature associations and contribution patterns learned from the urban morphological characteristics, with the synthetic samples serving as supplementary data for model learning rather than independent physical evidence.

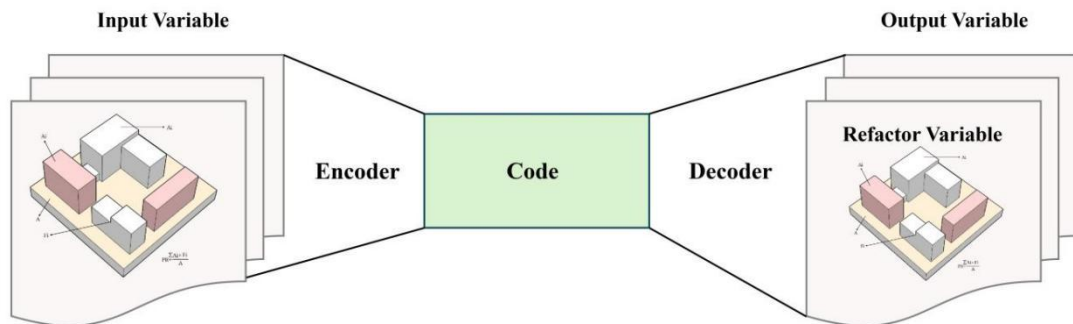


Figure 3. Framework of the VAE

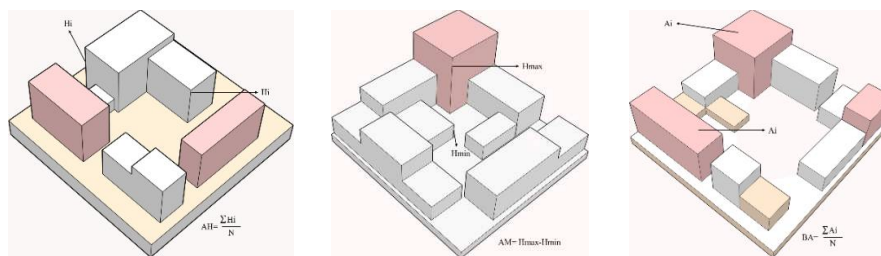
3.3 Feature filtration and hyperparameter tuning

The Pearson correlation coefficient is a non-parametric statistical method used to measure the direction and intensity of the trend relationship between two random variables. This coefficient has no strict requirements on data values, population distribution, sample size, or other aspects; as long as the data appear in pairs, it can be calculated using Formula (4).

$$r_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\sum_{j=1}^J (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{j=1}^J (x_i - \bar{x})^2} \sqrt{\sum_{j=1}^J (y_i - \bar{y})^2}} \quad (5)$$

In the formula: $x = [x_1, x_2, \dots, x_J]^T, y = [y_1, y_2, \dots, y_J]^T$ denote two J dimensional random variables; $\bar{x} = \frac{1}{J} \sum_{j=1}^J x_j, \bar{y} = \frac{1}{J} \sum_{j=1}^J y_j$ represent the mean values of x and y respectively.

The interaction among features holds significant importance in determining model accuracy. Features exhibiting high correlation are deemed redundant, thus selecting only those with minimal redundancy among continuous features can enhance both accuracy and stability of the model. The Pearson correlation coefficient, an established method for quantifying linear correlations between vectors, was chosen for its proven effectiveness in initial feature screening. A value exceeding 0.8 indicates the presence of multicollinearity between the respective features. Figure 4 depicted the Pearson correlation heatmap, illustrating the interplay of correlation among candidate features. The study identified features with correlation values surpassing 0.8 and further examined the coefficient values between influencing factors and Seismic vulnerability. By eliminating the variable with lower coefficient values with target value as well as higher correlation among two variables, the study achieved improved model performance.



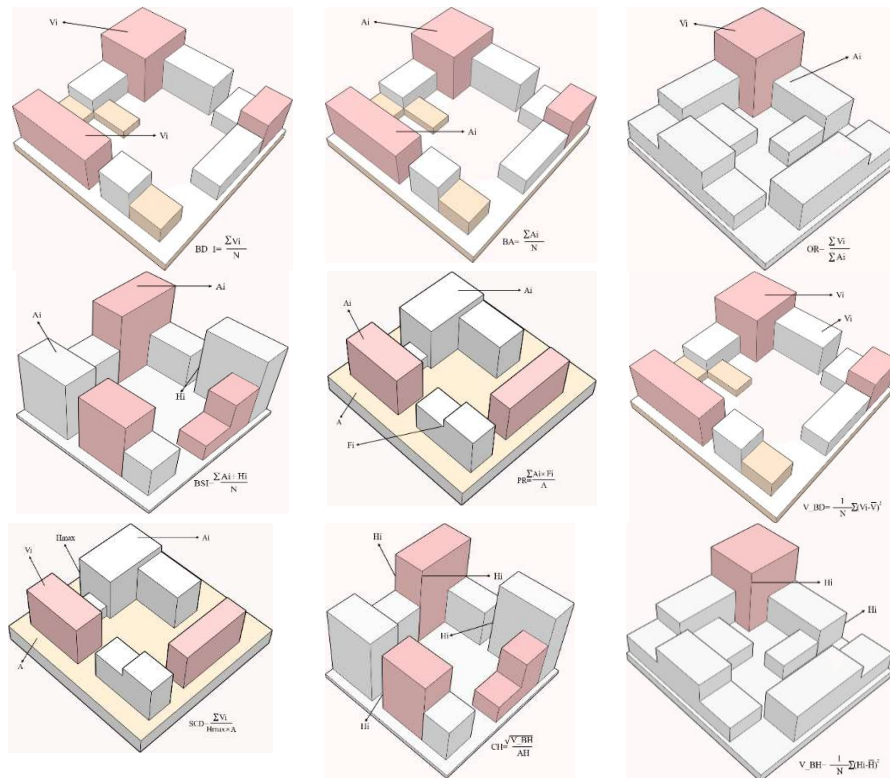


Figure 4. Illustrated graphical definition of urban 3D morphology

In this section, as shown in Figure 5, the study examined the association between building vulnerability (V) and ten other factors using Pearson correlation analysis.

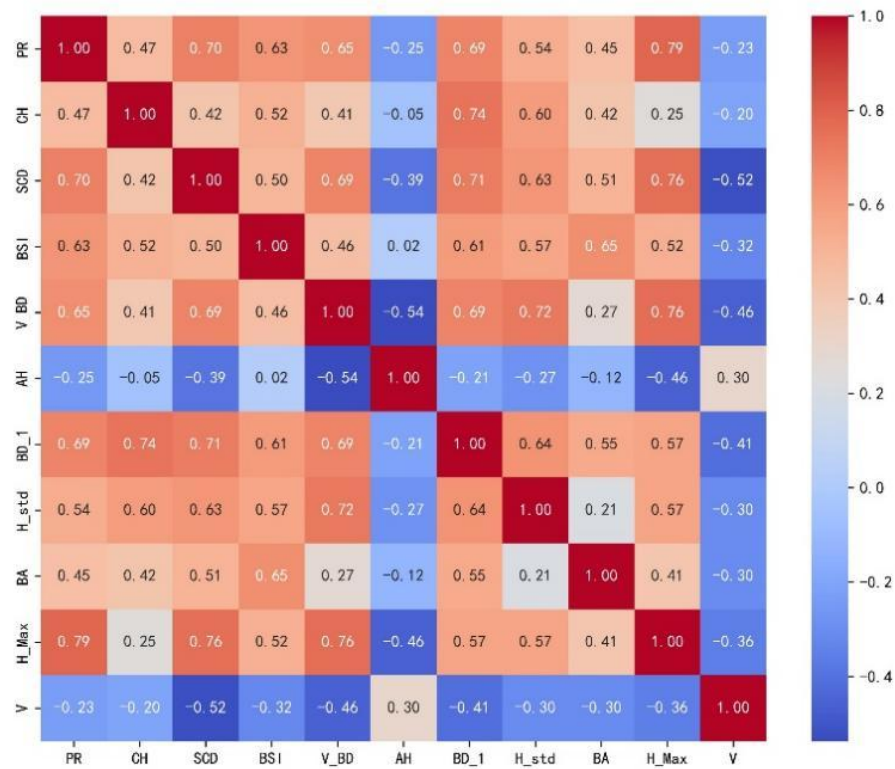


Figure 5. Heatmap of the Pearson correlation between influencing factors

As shown in Figure 5, for buildings with seismic vulnerability under intensity V conditions, most urban morphological indicators exhibited negative correlations with V, indicating that increases in building density, block density, and building volume may reduce vulnerability. PR (Pearson correlation coefficient=-0.23) and CH (Pearson correlation coefficient = -0.20) were weakly negatively correlated with vulnerability. An increase in floor area ratio provides more structural support for buildings, but the variations in individual building design limit its overall effect on vulnerability. Similarly, a staggered layout can help disperse seismic forces, distributing energy across buildings of different heights, which may reduce vulnerability at both the individual and collective levels, although this effect is relatively limited. SCD (Pearson correlation coefficient = -0.52) and BD_1 (Pearson correlation coefficient = -0.41) were strongly negatively correlated with vulnerability. Block density exhibited the strongest negative correlation with building vulnerability. Dense building clusters and compact urban blocks provide greater mutual structural support and protection, reducing building vulnerability and serving as a key factor in mitigating urban earthquake risk. Larger building volumes are associated with greater structural redundancy and stability, enhancing overall seismic performance and lowering vulnerability. It is also noteworthy that AH (Pearson correlation coefficient = 0.30) was positively correlated with vulnerability. Higher average building heights corresponded to increased vulnerability. High-rise buildings are more susceptible to damage during earthquakes or disasters, particularly in areas with insufficient overall density, indicating that urban high-rise development may increase risk and should be carefully planned.

3.4 Extreme gradient boosting algorithm

Extreme gradient boosting (XGBoost) algorithm is an ensemble learning algorithm and a variant of gradient-boosted decision trees. It enhances classification performance by integrating multiple decision tree models. Its basic principle is that in each iteration, a new decision tree is fitted by calculating the negative gradient of the loss function, and then the output of the new tree is added to the cumulative output of all previous trees. This process enables it to approximate the actual loss function with higher precision, thus rendering it can efficiently decouple complicated influencing mechanisms between seismic vulnerability and urban morphology.

The basic principles of XGBoost are elaborated as follows:

Taking the Classification and Regression Tree (CART), which is commonly

401 adopted as the base learner in XGBoost, as an example, the detailed description is
 402 provided below.

$$\hat{y}_i = \sum_{k=1}^K f_k(X_i), f_k \in R \quad (6)$$

403 In the above formula, X_i denotes the set of feature values corresponding to the i
 404 data sampling point; R represents the function space encompassing all CART trees;
 405 and $f_k(X_i)$ stands for the weight of the i data sampling point at the leaf node of the
 406 k tree.

407 Thus, the objective function of XGBoost can be expressed as follows:

$$obj(\theta) = \sum_{i=1}^n l(\hat{y}_i, y_i) + \sum \Omega(f_k) \quad (7)$$

408 Herein, $l(\hat{y}_i, y_i)$ denotes the loss error, and $\Omega(f_k)$ represents the regularization
 409 term, which controls the complexity of the tree model and prevents overfitting. $\hat{y}_i$ is the
 410 predicted value of the label for the i data sampling point output by the model, y_i is
 411 the true value of the label for the i data sampling point, and K is the number of trees.
 412 The expression of the regularization term is given as follows:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \omega^2 \quad (8)$$

413 Herein, γ and λ denote the penalty coefficients, T represents the number of
 414 leaf nodes, ω is the weight of the leaf nodes, and Const is the constant term. Thus, the
 415 objective function of XGBoost can be rewritten as follows:

$$obj(\theta) = \sum_{i=1}^n (y_i, \hat{y}_i + f_i(x_j)) + \Omega(f_i + Const) \quad (9)$$

416 **3.5 Shapley Additive Explanations**

417 The XGBoost model boasts high predictive performance, yet its internal decision-
 418 making process remains difficult to interpret. To address this issue, Shapley Additive
 419 Explanations (SHAP) is employed to realize both global and local interpretability of
 420 the model. Specifically, SHAP values are visualized to illustrate the contribution degree
 421 of each feature factor to the prediction results, and hierarchical clustering is utilized to

analyze the correlation among different feature factors. This integrated approach can be applied to the research on the driving factors of the spatiotemporal evolution of carbon storage. SHAP is capable of revealing the more complex logic underlying model predictions and helping users understand why specific factors exert an impact on the final results. The introduction of SHAP has brought new perspectives to machine learning research, enabling researchers to conduct in-depth exploration of the decision-making processes of models and to enhance the understanding and prediction of data.

Its calculation method is shown as presents:

$$y_i = y_{base} + F(\lambda_{i1}) + F(\lambda_{i2}) + \cdots F(\lambda_{ik}) \quad (10)$$

Where: λ_{ij} denotes the j feature of the i sample; y_i represents the predicted value of this sample obtained by the SHAP model; y_{base} is the mean of the predicted values of all samples. $F(\lambda_{ij})$ refers to the contribution value of the j feature in the i sample to y_i .

3.6 Partial dependence plot

Due to the relative opacity of data model training and construction, such models suffer from a "black-box" effect in fields with high interpretability requirements. This is because the large number of decision trees makes it difficult to understand and compare the differences and attributes between individual decision trees, thus necessitating the application of explanatory tools. Partial dependence plot (PDP), a commonly used analytical technique, illustrates the marginal effect of a single feature on the prediction results of a pre-fitted model. Specifically, the prediction function is fixed at several values of the selected feature, while averaging over all other features. Compared with traditional linear regression models, the PDP offers two distinct advantages. First, the marginal effects of influencing factors derived from this technique are generated based on models that account for the impacts of other covariates, rather than merely reflecting a bivariate relationship between the target variable and individual features. Second, this method is applicable to linear models, nonlinear models, and even more complex mathematical models. The specific calculation is follows:

$$\hat{f}_s(x_s) = E_{X_C} = [\hat{f}(x_s, X_C)] \int \hat{f}(x_s, X_C) dP(X_C) \quad (11)$$

Where x_s are features analyzed in the current section, X_C is the remaining

features, Partial dependence works by marginalizing the $\hat{f}$ output over the distribution of the features in set C. The study performed single feature and double feature PDP analysis, respectively.

3.7 Moran's I method

In local spatial correlation analysis, Local Moran's I is one of the most widely adopted statistics. It quantifies the spatial association between each individual spatial unit and its neighboring units and identifies localized spatial clustering or spatial outliers within the study area. In this study, this indicator is applied to quantitatively explore the spatial heterogeneity of urban building vulnerability and reveal the localized spatial distribution characteristics and differentiation rules of different urban morphological elements. The calculation formula is as follows:

$$I_i = \frac{(x_i - \bar{x}) \sum_{j=1}^n w_{ij} (x_j - \bar{x})}{\left(\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} \right) \sum_{j=1}^n w_{ij}} \quad (12)$$

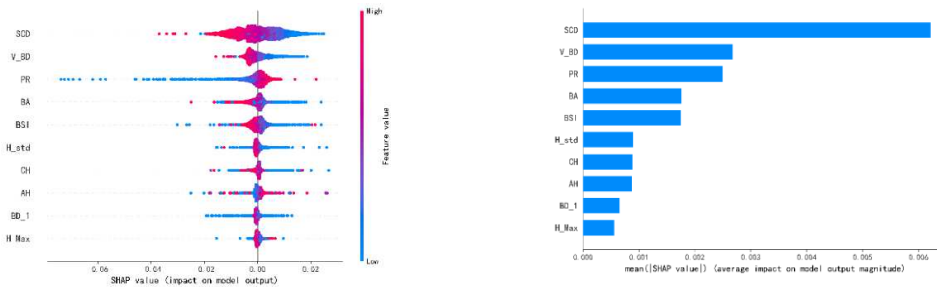
Among them, $S_0 = \sum_{i=1}^n \sum_{j=1}^n w_{ij}$, where n is the total number of spatial units, y_i and y_j represent the attribute values of the i -th spatial unit and the j -th spatial unit respectively, $\bar{y}$ is the mean value of the attribute values of all spatial units, and w_{ij} is the spatial weight value.

In this study, the spatial-weight matrix was constructed using the queen-contiguity criterion and further row-standardized. The weight was defined as $w_{ij} = 1$ if parcel i and parcel j shared at least one edge or vertex, otherwise $w_{ij} = 0$, and self-weight was set to zero. A permutation test with 999 random permutations was performed to obtain the p-values for Local Moran's I. The significance level was set at $\alpha = 0.05$; spatial units with ($p < 0.05$) were considered statistically significant. Non-significant units ($p \geq 0.05$) were not interpreted as local spatial clusters or outliers in the subsequent analysis. The spatial-weight matrix, neighbourhood criterion, distance threshold, row standardization, treatment of isolated spatial units, permutation testing, significance level, and multiple-testing correction were all specified and implemented in the Local Moran's I analysis. For conciseness and considering the overall scope of the manuscript, the detailed settings and procedural parameters are not presented individually in the main text.

4 Results

4.1 Overall Impact of urban 3D Morphological Features on Seismic Vulnerability

Figure 6 presents the SHAP value distributions (Figure 6a) and feature importance rankings (Figure 6b). All variables passed the Variance Inflation Factor (VIF) test, indicating no multicollinearity among features and ensuring the robustness of the model interpretation. Among all variables, SCD emerges as the most influential feature in the model. It shows the widest SHAP value range with clear positive and negative effects and also the highest mean SHAP value, indicating its dominant contribution to model predictions across samples.



(a) Results of the Theil–Sen median slope

(b) Results of Mann-Kendall Test

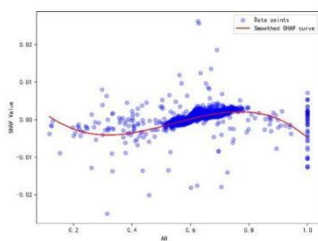
Figure 6. Results of Sen and MK

Specifically, high SCD values are associated with strong positive effects on model predictions, whereas low values lead to negative effects, indicating strong sensitivity and explanatory power. V_BD and PR also exhibit relatively wide SHAP distributions and clear bidirectional effects, indicating weaker influence than SCD but still stable contributions to model predictions. In contrast, BA, BSI, and H_std show SHAP values concentrated around zero, indicating weak contributions to model predictions. H_Max and BD_1 exhibit narrow SHAP ranges with negligible effects, indicating minimal influence. The feature importance ranking (right panel) further confirms these results, with SCD showing the highest importance, followed by V_BD and PR, while BA, BSI, and H_std are less influential, and H_Max and BD_1 contribute the least. Overall, the model is dominated by SCD, with V_BD and PR serving as secondary contributors, while the remaining variables have limited or negligible effects.

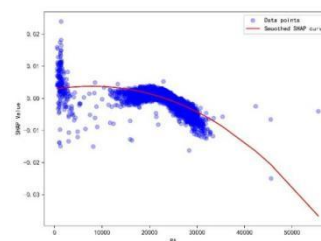
4.2 Feature Marginal Impact Analysis Based on PDPs

4.2.1 Impact of Single Variables on Seismic Vulnerability

To assess the marginal effects of urban three-dimensional morphological characteristics on seismic vulnerability, this study applies the PDP method (Figure 7), in which other variables are held at their mean values. AH (Figure 7 a) shows minimal variation, with values fluctuating around zero (within ± 0.005), indicating a negligible effect, with only a slight negative shift near 1.0. BA (Figure 7b) exhibits a clear transition pattern, changing from a weak positive effect at low values to a strong negative effect at high values (approximately -0.038), suggesting a decreasing association with vulnerability. BD_1 (Figure 7c), H_Max (Figure 7f), and H_std (Figure 7g) exhibit threshold-like behavior, remaining near zero at low values and increasing positively beyond certain thresholds (approximately 0.008 – 0.009), indicating higher vulnerability at larger values. BSI (Figure 7d) shows a nonlinear U-shaped relationship, with the strongest negative effect at moderate values (~ 250000), indicating a non-monotonic influence. CH (Figure 7e) shows a strong negative relationship, decreasing after approximately 0.6 and reaching about -0.035 at higher levels, suggesting reduced vulnerability with increasing compactness. PR (Figure 7h) exhibits a monotonic increasing trend, indicating a consistently positive contribution across its range. SCD (Figure 7i) exhibits a complex nonlinear relationship, characterized by an initial increase followed by a decrease and subsequent rebound, with peak positive effects near 10 and strongest negative effects near 55 . V_BD (Figure 7j) shows a non-monotonic trend, with positive effects at low values (0 – 40) and negative effects thereafter. Overall, urban morphological features exhibit heterogeneous nonlinear and threshold-dependent effects on seismic vulnerability.



(a) Partial Dependence for AH



(b) Partial Dependence for BA

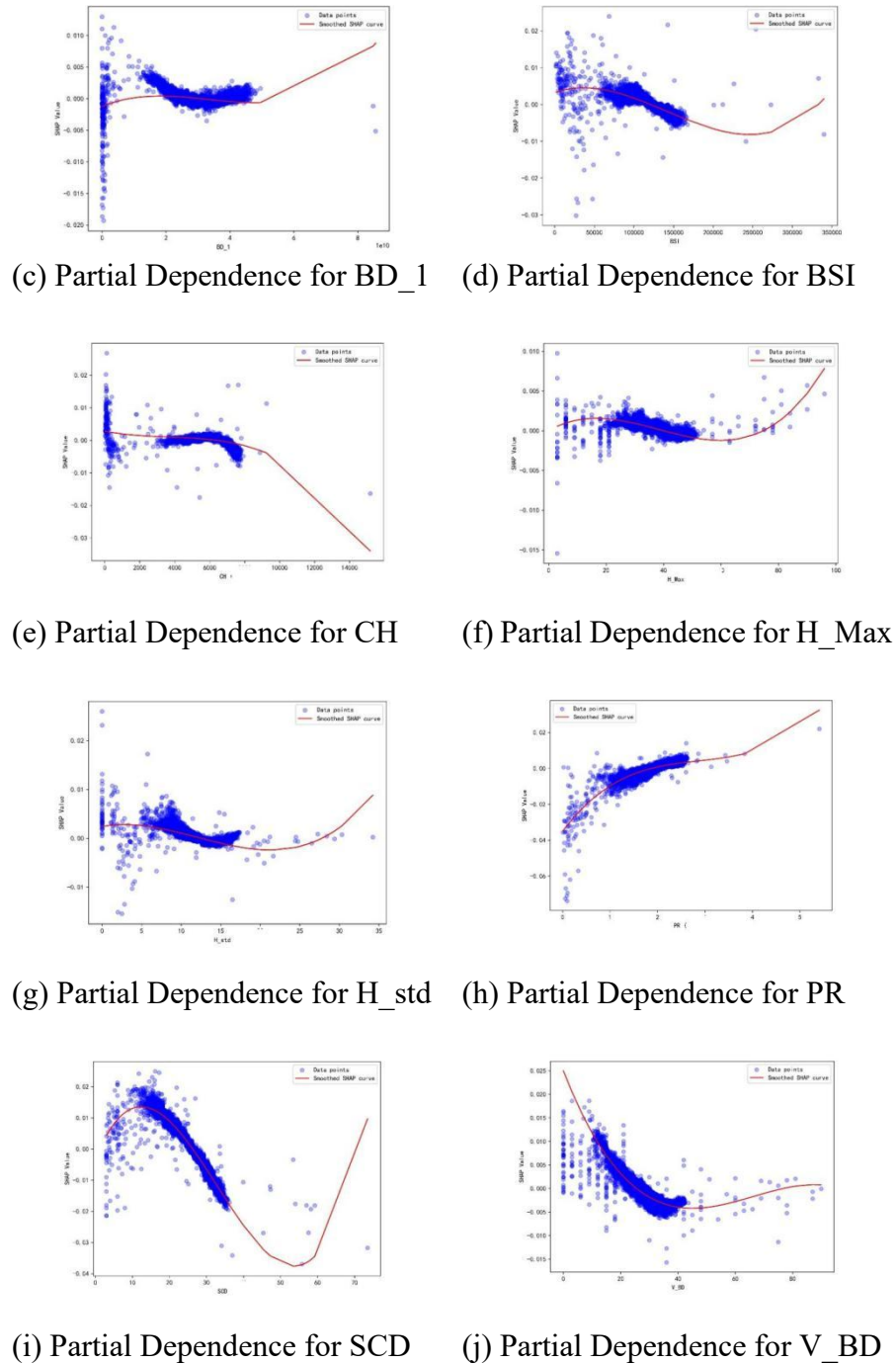
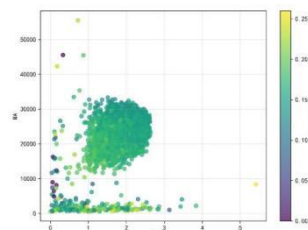


Figure 7. Nonlinear relationships between urban 3D morphological variables and seismic vulnerability

4.2.2 Feature Correlation and Distribution of Seismic Vulnerability Predictions

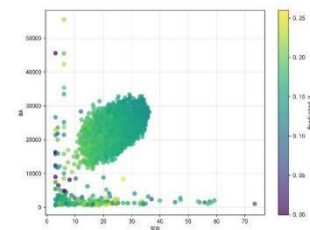
To investigate interactions among urban 3D morphological indicators (BA, BSI, PR, SCD, and V_BD) and their effects on seismic vulnerability under seismic intensity of V, this study constructs pairwise contour maps and scatter plots (Figure.8 and

Figure.9). Results indicate a negative interaction between BA and BSI, where low values of both correspond to higher V (~0.17–0.175), while higher values of either variable are associated with reduced V (~0.15). The scatter plots reveal that mid-range combinations show relatively stable patterns, whereas extreme combinations are associated with higher V values. PR shows a positive interaction with BA but an opposite relationship with BSI, with higher V observed under high PR–low BSI conditions or strong PR–BA imbalance, suggesting increased vulnerability under uncoordinated morphological configurations. SCD shows a generally negative interaction with BA, BSI, and PR, where low SCD combined with moderate BA or low BSI is associated with higher V (0.17–0.186), while high SCD corresponds to lower V (0.114–0.144), indicating a stabilizing effect within the 10–30 range. V_BD exhibits similar interaction patterns, with higher V under low V_BD conditions and reduced V under higher V_BD. Overall, most samples fall within V (0.15–0.20) indicating moderate seismic vulnerability, while extreme values are associated with extreme morphological configurations. The consistency between contour and scatter results suggests strong interaction effects, indicating that extreme or poorly balanced urban forms are associated with higher seismic vulnerability despite their low occurrence.



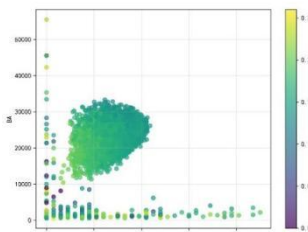
(a) BA and PR Interaction

Scatter Plot



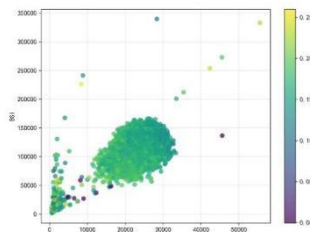
(b) BA and SCD Interaction

Scatter Plot



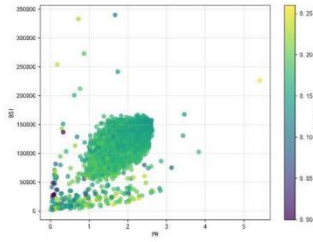
(c) BA and V_BD Interaction

Scatter Plot



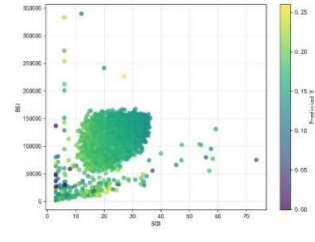
(d) BSI and BA Interaction

Scatter Plot



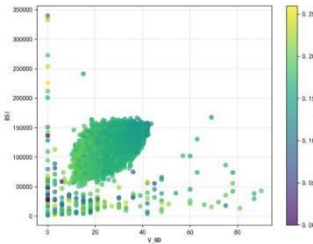
(e) BSI and PR Interaction

Scatter Plot



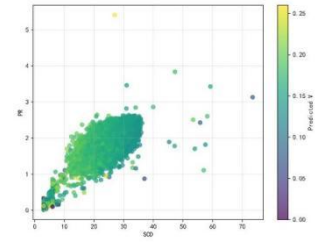
(f) BSI and SCD Interaction

Scatter Plot



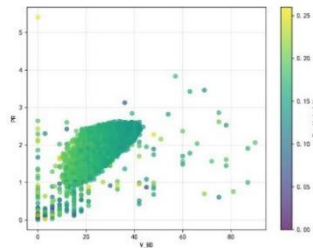
(g) BSI and V_BD Interaction

Scatter Plot



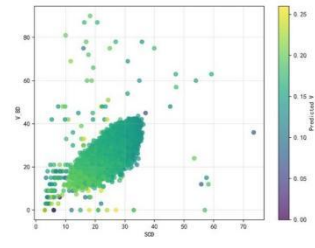
(h) PR and SCD Interaction

Scatter Plot



(i) PR and V_BD Interaction

Scatter Plot

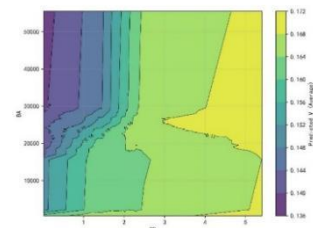


(j) V_BD and SCD Interaction

Scatter Plot

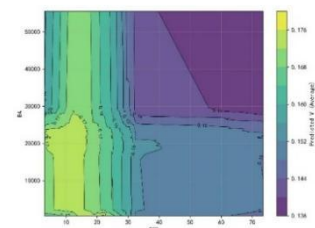
548 Figure 8. Interaction effects between urban 3D morphological variables and seismic vulnerability

549 (Scatter Plot)



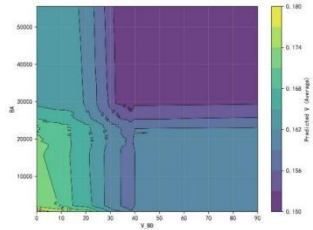
(a) BA and PR Interaction

Contour Plot



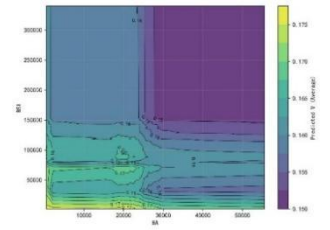
(b) BA and SCD Interaction

Contour Plot



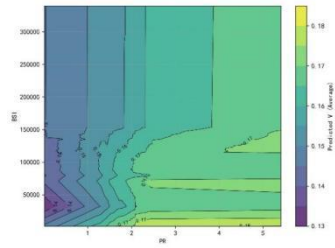
(c) BA and V_BD Interaction

Contour Plot



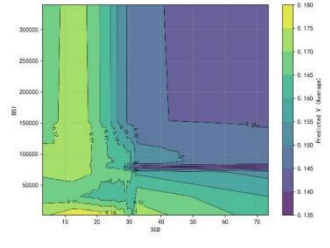
(d) BSI and BA Interaction

Contour Plot



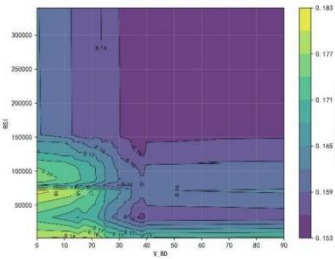
(e) BSI and PR Interaction

Contour Plot



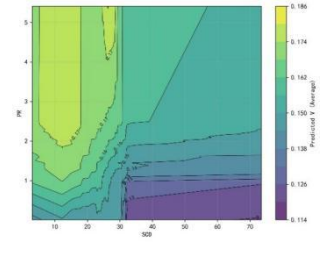
(f) BSI and SCD Interaction

Contour Plot



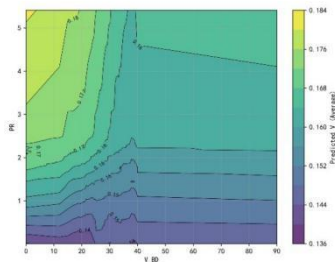
(g) BSI and V_BD Interaction

Contour Plot



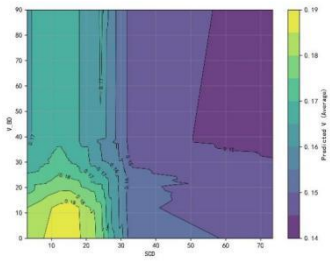
(h) PR and SCD Interaction

Contour Plot



(i) PR and V_BD Interaction

Contour Plot



(j) V_BD and SCD Interaction

Contour Plot

550 Figure 9. Interaction effects between urban 3D morphological variables and seismic vulnerability

551 (Contour Plot)

4.3 Spatial Distribution of Seismic Vulnerability Based on Local Moran's I

Figure 10 shows the spatial clustering results based on Moran's I index. Moran's I is a measure of spatial autocorrelation, indicating the degree to which a region's value is similar to the values of its neighboring regions for a particular variable. The map categorizes the regions into different Moran's classes as follows:

The High-High and Low-Low categories indicate regions where values are spatially clustered, which can help identify areas of high and low development, density, or infrastructure. The High-Low and Low-High categories highlight areas that stand out as outliers, where a high-value region is surrounded by low-value regions, or vice versa. These outliers might represent zones of uneven development or areas with unique characteristics. By analyzing multiple variables together, the kernel density plots help show how the distribution of each variable changes across the spatial categories. This can provide deeper insights into the spatial structure and urban dynamics of the study area.

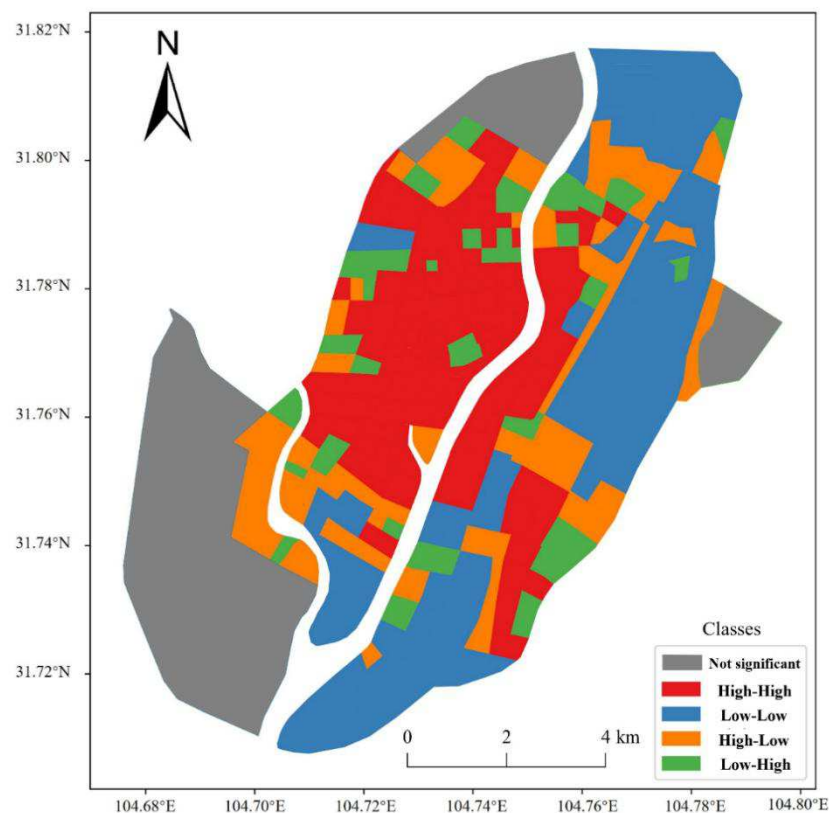


Figure 10. Moran's I distribution of seismic vulnerability

4.4 Feature Distribution under Different Spatial Cluster Types

From the perspective of the four Moran's I spatial classifications (High-Low,

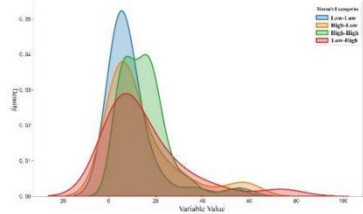
Low–High, Low–Low, and High–High), the kernel density distributions of urban three-dimensional morphological variables exhibit significant heterogeneity and distinct spatial differentiation patterns, as shown in Figure 11.

In High–Low category, variables generally display a transitional pattern from high values to low-value surroundings. PR and BD_1 show relatively high values locally, but are adjacent to low-value regions, resulting in more dispersed density distributions. This reflects a spatial mismatch between high-intensity developed areas and intensity surroundings. Similarly, the spatial SCD and AM exhibit wide variability, indicating substantial internal differences in structural complexity and building scale, highlighting strong spatial heterogeneity and boundary-transition characteristics.

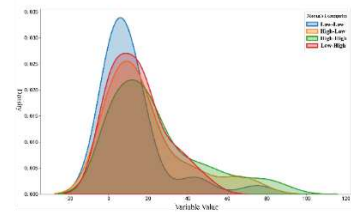
In Low–High category, variables exhibit the opposite transition pattern, where low-value are adjacent to high-value neighborhoods. PR and BD_1 are generally low but influenced by nearby high-density areas, leading to stretched and skewed distributions. This suggests a spatial spillover effect from highly developed regions. SCD and AM also show considerable dispersion, indicating coexistence of simple, small-scale buildings with influences from more complex and larger neighboring structures, thus forming clear spatial gradients.

In Low–Low category, all variables are predominantly concentrated in low-value ranges, with kernel density curves showing sharp and centralized peaks. PR and BD_1 indicate low development intensity and sparse building distribution; SCD remains low, reflecting simple spatial structures and limited morphological complexity; AM is also concentrated at smaller values, suggesting smaller and more dispersed building footprints. Overall, this category represents underdeveloped areas with strong spatial homogeneity.

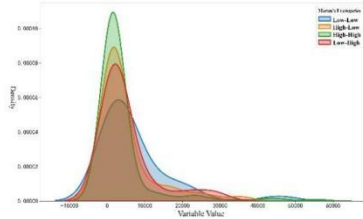
In High–High category, variables are clearly concentrated in high-value ranges, with density peaks shifted toward higher values. PR and BD_1 are both high, indicating intensive land use and compact urban structures; SCD is also elevated, reflecting complex spatial configurations and strong aggregation of buildings; AM shows larger values, suggesting larger building sizes or highly concentrated development patterns. Additionally, height-related variables such as H_Max and AH tend to be higher in this category, further indicating that these areas correspond to urban cores or highly developed zones with strong spatial clustering and structural complexity.



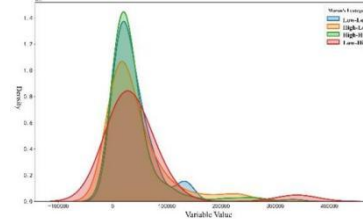
(a) AH KDE Plot



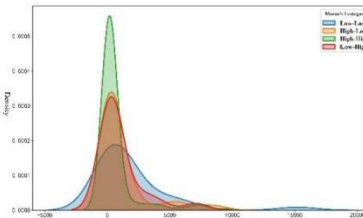
(b) AM KDE Plot



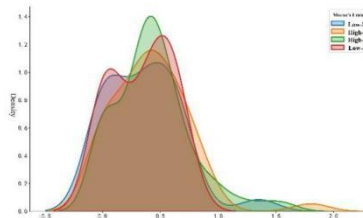
(c) BA KDE Plot



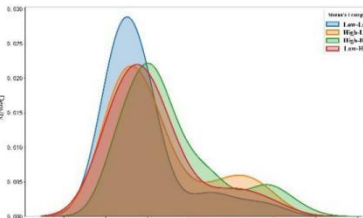
(d) BD_1 KDE Plot



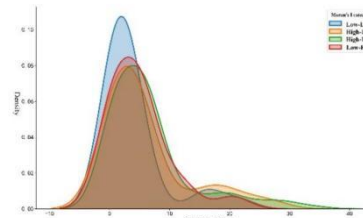
(e) BSI KDE Plot



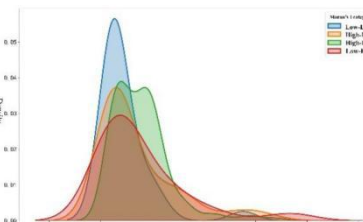
(f) CH KDE Plot



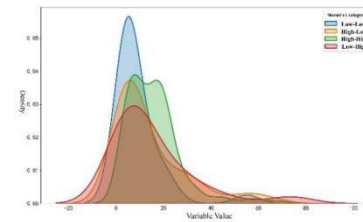
(g) H_Max KDE Plot



(h) H_std KDE Plot



(i) OR KDE Plot



(j) PR KDE Plot

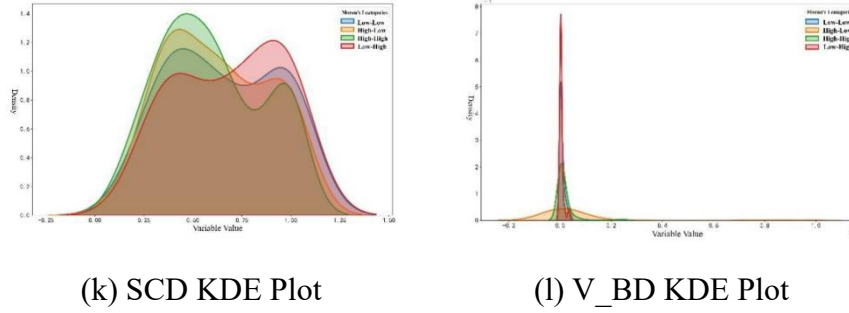


Figure 11. Spatial distribution of 3D morphological features

5 Discussion

5.1 Advantages of XGBoost Model

As shown in Figure.12, the visualization of the scatter plots between predicted and observed values reveals significant hierarchical differences in the fitting performance of the four regression models. XGBoost, RF, SVR, and DT were evaluated using the same dataset, experimental settings, and evaluation criteria to ensure a fair comparison of model performance. All models were consistently optimized under identical data conditions, and the hyperparameter settings, selected XGBoost parameters, random seed, and evaluation procedures were specified to improve reproducibility. The optimal hyperparameter configurations for each model are presented in Table 3. Specifically, the predictions of XGBoost are almost perfectly aligned with the ideal line of “Predicted = Observed,” with the error distribution predominantly concentrated in low-deviation regions. This is quantitatively supported by its superior performance ($R^2 = 0.92$, $RMSE = 0.0158$), indicating an excellent agreement between predicted and actual values. In comparison, RF model shows a relatively concentrated distribution around the ideal line, although with increased dispersion and a higher proportion of moderate errors, as reflected by a reduced R^2 of 0.78 and an increased RMSE of 0.0259. The SVR model exhibits a further decline in performance, with scatter points deviating more substantially from the ideal line and a noticeable increase in high-error regions, resulting in a lower R^2 of 0.51 and RMSE of 0.0300. DT model performs the worst, with predictions largely deviating from the ideal line and dominated by high-error distributions, yielding a very low R^2 of 0.20 and indicating poor predictive consistency.

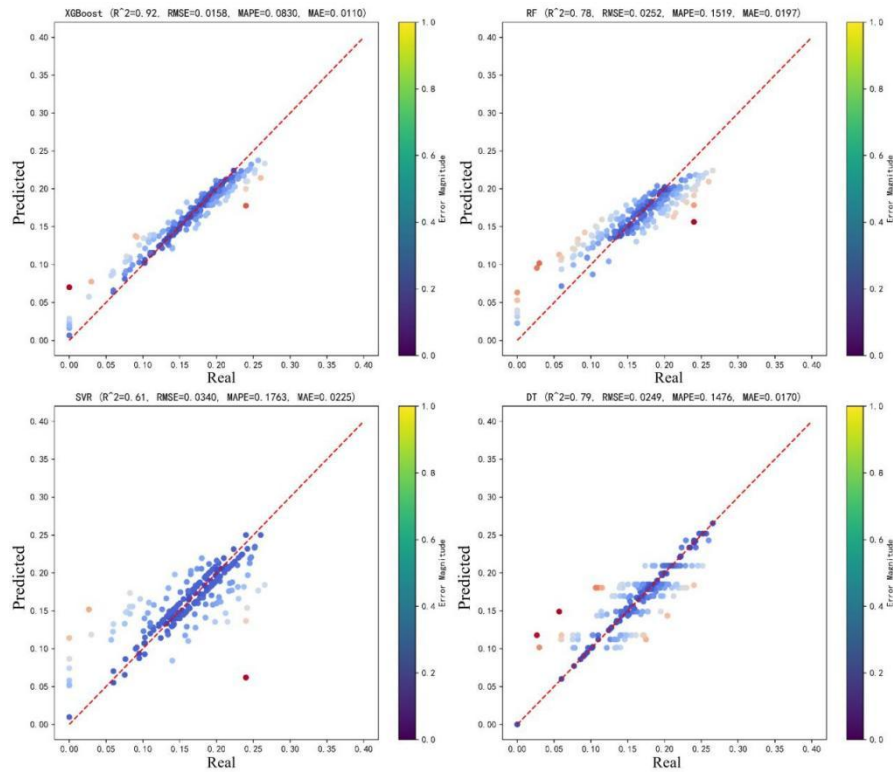


Figure. 12. Comparison between predicted depths and observed depths.

Table 3. Optimal parameters of each model

Model	Level	R2	RMSE	MAE	MAPE	Best_Params
XGBoost	V	0.9718	0.0027	0.0018	0.0111	'learning_rate': 0.05, 'max_depth': 7, 'n_estimators': 300
RandomForest	V	0.7631	0.0078	0.0049	0.0320	'max_depth': 7, 'n_estimators': 100
SVR	V	0.8583	0.0061	0.0038	0.0242	'C': 10.0, 'epsilon': 0.01, 'gamma': 'scale'
DecisionTree	V	0.7433	0.0082	0.0053	0.0345	'max_depth': 7, 'min_samples_leaf': 1, 'min_samples_split': 2

5.2 Mechanism of Statistical Relationships Between Features and Vulnerability

The predicted value of seismic vulnerability is not determined by a single urban 3D morphological indicator, but by the synergistic effects of 15 indicators, including

PR, BD_1, and V_BH. Essentially, these indicators collectively represent urban spatial structure, building aggregation patterns, and vertical morphological characteristics, which together influence seismic wave propagation, structural mechanical response, and damage severity. Based on the interaction patterns observed in contour maps and scatter plots, the mechanisms are as follows:

For PR, as a positive driving factor of seismic vulnerability, higher PR (>2) indicates concentrated vertical development and larger building volume. When combined with BA, BSI, SCD, and V_BD, it tends to form an “upper-heavy” unbalanced spatial structure, which enhances seismic wave transmission and amplification effects, significantly increasing structural damage probability and thus raising seismic vulnerability. In contrast, lower PR (<1) is often associated with compensatory high values of other indicators, leading to reduced seismic vulnerability.

BD_1 and SCD jointly reflect horizontal building aggregation and show a negative synergistic relationship with seismic vulnerability. A moderate range ($10 \leq \text{SCD} \leq 30$) forms a “buffer zone” that maintains urban functional order while avoiding stress concentration from excessive density or structural instability from excessive dispersion, corresponding to stable seismic vulnerability values. When SCD is too high (>40), buildings become extremely dense and form a rigid constraint system that increases lateral resistance, thereby reducing seismic vulnerability. However, when SCD is too low, spatial dispersion increases and seismic wave propagation becomes irregular, leading to higher seismic vulnerability.

Height-related indicators (V_BH, AH, CH, AM, H_std, H_Max, H_Min) mainly reflect vertical heterogeneity. Larger V_BH, CH, and AM indicate significant height differences and irregular structures, leading to torsional effects and local failure under earthquakes. Moderate AH and H_std support uniform stress distribution and stable seismic vulnerability values, while extreme H_Max or H_Min disrupt vertical balance and amplify seismic response.

Volume- and area-related indicators (V_BD, Average Volume, BA, Average Area) also show negative synergy with seismic vulnerability. Small V_BD (<20) combined with extreme or low BA leads to spatial imbalance and stress concentration, increasing seismic vulnerability. In contrast, large V_BD (>50) combined with moderate BA (10000–30000) improves structural balance and reduces seismic vulnerability.

BSI reflects façade-to-base relationships and shows nonlinear effects with

seismic vulnerability. Low BSI (<100000) generally indicates regular structure, but extreme combinations with BA or PR increase imbalance and raise seismic vulnerability. High BSI (>150000) increases complexity but strengthens overall structural constraint, reducing seismic vulnerability.

OR reflects occupancy rate. Excessively high OR causes congestion and wave superposition effects, while excessively low OR leads to spatial fragmentation and uneven force distribution; both increase seismic vulnerability. Moderate OR ensures balanced seismic response and lower vulnerability.

5.3 Interpretation of Spatial Distribution Patterns of Key Features

From the perspective of Moran's I spatial classification (High–Low, Low–High, Low–Low, and High–High), the relationships between urban 3D morphological variables and spatial clustering patterns exhibit pronounced heterogeneity and clear structural differentiation.

In High–Low category, regions are characterized by locally high values surrounded by low-value neighborhoods, reflecting strong spatial discontinuity. Variables such as PR, BD_1, SCD, and OR tend to show relatively high values, indicating locally intensive development, while adjacent areas remain at lower levels, resulting in sharp spatial contrasts. Morphological indicators including V_BH, V_BD, H_std, and CH generally display higher variability, suggesting heterogeneous building forms and mixed development patterns. These areas typically correspond to transitional or urban edge zones with pronounced spatial imbalance.

In Low–High category, the spatial configurations is reversed, where locally low-value regions are embedded within high-value surroundings. PR, BD_1, and SCD are generally low, indicating relatively sparse local development; however, surrounding high-density areas induce noticeable spatial gradients. Indicators such as AH, H_Max, and BSI may exhibit moderate increases due to spillover effects from adjacent clusters. The distributions of BA, average volume, and AM are often more dispersed, reflecting the coexistence of small-scale simple structures with influences from more complex neighboring forms. This category generally represents areas undergoing transformation or subject to development pressure.

In Low–Low category, all variables are predominantly concentrated at low levels, indicating strong spatial homogeneity and weak development intensity. PR, BD_1, SCD, and OR remain consistently low, reflecting sparse building distribution

and low land-use intensity. Morphological indicators such as V_BH, V_BD, CH, H_std, and BSI also show low values, suggesting uniform building heights, simple structural configurations, and limited morphological complexity. Building-scale indicators such as BA, AM, and average volume are generally small, while both H_Max and H_Min remain low. These areas are typically located in suburban or underdeveloped regions, where urbanization is limited and spatial structures are relatively simple, corresponding to generally lower seismic vulnerability due to reduced exposure and density.

In High-High category, variables cluster at high values, indicating strong positive spatial autocorrelation and intensive urban development. PR, BD_1, SCD, and OR are all significantly elevated, reflecting compact urban form, high land-use efficiency, and dense building distribution. Morphological variability indicators such as V_BH, V_BD, CH, and H_std are also high, suggesting diverse building heights and complex structural configurations. Additionally, AH, H_Max, BA, and average volume are typically large, indicating the presence of high-rise and large-scale buildings, while BSI is also elevated, reflecting more complex architectural forms. These regions generally correspond to urban core areas or highly developed zones, where both development intensity and structural complexity are maximized, and seismic vulnerability tends to be relatively high due to high exposure and concentration.

6 Conclusion

This study combines spatial statistics and explainable machine learning to explore the relationship between urban 3D morphology and seismic vulnerability. VAE is used for data augmentation, XGBoost for vulnerability assessment, SHAP for feature interpretation, and Moran's I for spatial pattern analysis. A case study was conducted in Jiangyou City, Southwestern China. The results indicate that urban 3D morphological characteristics are significantly associated with seismic vulnerability under Intensity V, showing clear and differentiated influence patterns. Most indicators are negatively correlated with vulnerability, with SCD having the strongest impact, followed by V_BD and PR, while AH shows a positive effect and H_Max contributes minimally. Different features exhibit distinct mechanisms: PR positively influences vulnerability, whereas BD_1, BA, and CH show negative effects; H_std and H_Max present threshold-type positive influences, and SCD and BSI display nonlinear relationships, consistent with seismic engineering principles. Feature interactions are generally weak, though BA and

BSI show some synergy, and extreme combinations (e.g., high PR, low SCD and V_{BD}) significantly amplify risk, especially under low V_{BD} conditions. Spatial patterns reveal clear core–periphery differentiation, with height- and density-related indicators decreasing outward, while BSI and BA show opposite trends. Among all models, XGBoost achieves the best performance ($R^2 = 0.9648$, MAPE = 0.99%), and VAE-based data augmentation improves model robustness. Overall, seismic vulnerability reflects the balance of urban 3D morphology, where coordinated configurations reduce seismic risk and extreme combinations increase vulnerability, providing a scientific basis for urban seismic risk mitigation and planning.

Acknowledgements

We sincerely thank the anonymous reviewers and associated editors for their valuable and constructive comments and suggestions that helped to improve this paper.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Disclosure statement

The authors report there are no competing interests to declare.

References

- Abbas, Yeganeh-Bakhtiary, Hossein, EyvazOghli, Naser, Shabakhty, and Soroush, Abolfathi. 2023. Machine learning prediction of wave characteristics: Comparison between semi-empirical approaches and DT model. *Ocean Engineering*, 286(P2). <https://doi.org/10.1016/j.Oceaneng.2023.115583>
- Adesina, Damilola, Hsieh, Chung-Chu, Sagduyu, Yalin E, and Qian, Lijun. 2022. Adversarial machine learning in wireless communications using RF data: A review. *IEEE Communications Surveys & Tutorials*, 25(1), 77-100.
- Aloisio, Angelo, Irti, Francesco, Rosso, Marco Martino, Quaranta, Giuseppe, Demartino, Cristoforo, and Fragiaco, Massimo. 2026. Fast-Track Code-Based Seismic Vulnerability Screening with Machine Learning: Evidence from 300 Italian Buildings. *Procedia Structural Integrity*, 78, 1-8. <https://doi.org/10.1016/j.Prostr.2025.12.001>
- Asadollahzadeh, Danesh, and Behnam, Behrouz. 2025. Machine learning approaches for seismic vulnerability assessment of urban buildings: A comparative study with analytic hierarchy process. *Progress in Disaster Science*, 25, 100398-100398.

<https://doi.org/10.1016/j.Pdisas.2024.100398>

- Chen, Jun, Chen, Nan, Shen, Jiaxu, Xiang, Mengjie, and Xu, Zekun. 2026. Scalable AI computational framework for seismic response simulation of million-scale urban building clusters. *Computer-Aided Civil and Infrastructure Engineering*, 41, 100016-100016. <https://doi.org/10.1016/j.Caiaie.2026.100016>
- Chen, Yanguang. 2013. New approaches for calculating Moran's index of spatial autocorrelation. *PloS one*, 8(7), e68336.
- Chicco, Roberta Di, Chieffo, Nicola, and Formisano, Antonio. 2024. Exposure and seismic vulnerability of masonry buildings grouped in aggregate of a typical historical centre in the Basilicata region of Italy. *Journal of Building Engineering*, 94, 109859-109859. <https://doi.org/10.1016/j.Jobbe.2024.109859>
- Chu, Yongbo, Wen, Haijia, Zhang, Chi, Liu, Bo, Li, Zhuohang, Di, Mingrui, and Qian, Long. 2025. Corrigendum to "Mountain environmental-oriented building vulnerability assessment and analysis of influencing factors based on the mRMR-stacking-SHAP method" [Structures 72 (2025) 108276]. *Structures*, 79, 109081-109081. <https://doi.org/10.1016/j.Istruc.2025.109081>
- Davtalab, Mehri, and Byčenkienė, Steigvilė. 2026. Synergistic effects of spatial urban form on PM2.5 concentration and urban heat islands: A multi-scale explanatory machine learning. *Journal of Cleaner Production*, 538, 147226-147226. <https://doi.org/10.1016/j.Jclepro.2025.147226>
- Dong, Jianwei, Chen, Yumin, Yao, Bingyu, Zhang, Xiao, and Zeng, Nianfeng. 2022. A neural network boosting regression model based on XGBoost. *Applied Soft Computing*, 125, 109067.
- Eren, Berfin, and Şalgamcıoğlu, Mehmet Emin. 2025. Scale, state and the city: Transformation of Diyarbakır, Suriçi region through the framework of spatial morphology and urban resilience. *City and Environment Interactions*, 100225.
- Gawde, Shreyas, Patil, Shruti, Kumar, Satish, Kamat, Pooja, Kotecha, Ketan, and Alfarhood, Sultan. 2024. Explainable predictive maintenance of rotating machines using LIME, SHAP, PDP, ICE. *IEEE Access*, 12, 29345-29361.
- Gondaliya, Kaushik M., Vasanwala, Sandip A., Desai, Atul K., Amin, Jignesh A., and Bhaiya, Vishisht. 2024. Machine learning-based approach for assessing the seismic vulnerability of reinforced concrete frame buildings. *Journal of Building Engineering*, 97, 110785-110785. <https://doi.org/10.1016/j.Jobbe.2024.110785>
- He, Zonghang, Guo, Hongmei, Zhang, Ying, Zhao, Zhen, Zhang, Can, Shi, Bingxin, Yang, Yuping, and Yang, Zhipeng. 2025. Seismic vulnerability model for typical buildings in Sichuan Province, China, based on empirical seismic damage statistics. *Structures*, 78, 109294-109294. <https://doi.org/10.1016/j.Istruc.2025.109294>
- Islam, Zubayer, Abdel-Aty, Mohamed, Cai, Qing, and Yuan, Jinghui. 2021. Crash data augmentation using variational autoencoder. *Accident Analysis & Prevention*, 151, 105950.
- Jackson, Monica C, Huang, Lan, Xie, Qian, and Tiwari, Ram C. 2010. A modified version of Moran's I. *International journal of health geographics*, 9(1), 33.
- Kallas, Joe, and Napolitano, Rebecca. 2025. Beyond field surveys: Understanding the role of 3D spatial attributes for data-driven blast vulnerability assessment of masonry buildings. *International Journal of Disaster Risk Reduction*, 128, 105672-105672. <https://doi.org/10.1016/j.Ijdr.2025.105672>
- Łabuz, Rita, Blazy, Rafał, and Hess, Daniel Baldwin. 2026. Using the weighted urban proliferation (WUP) approach to measure the effects of planning strategies on urban sprawl in Poland. *Cities*, 173, 106812.

- Li, Si-Qi, and Xu, Miao. 2026. Seismic failure and loss analysis of adobe and timber frame buildings considering statistical vulnerability and three-dimensional modelling strategies. *Engineering Failure Analysis*, 110724.
- Liu, Chang, Antypenko, Ruslan, Sushko, Iryna, and Zakharchenko, Oksana. 2022. Intrusion detection system after data augmentation schemes based on the VAE and CVAE. *IEEE Transactions on Reliability*, 71(2), 1000-1010.
- Liu, Chao, Li, Guoqing, Huang, Chengcheng, Herzog, Otthein, Ritter, Helge, Ma, Shengxin, and Ye, Yu. 2026. Generative AI for Urban Planning and Design: Progress Review and Future Perspectives. *Engineering*.
- Luo, Jun, Zhao, Ziyu, and Zeng, Rusi. 2026. Urban spatial structure and sustainable development: A bibliometric analysis. *Sustainable Cities and Society*, 140, 107260-107260. <https://doi.org/10.1016/j.Scs.2026.107260>
- Maiorana, Emanuele, Conte, Giorgio, and Briseghella, Bruno. 2026. Seismic vulnerability assessment of heritage-protected school buildings: experimental and numerical modeling approaches. *Procedia Structural Integrity*, 78, 57-64. <https://doi.org/10.1016/j.Prostr.2025.12.008>
- Malakar, Sukanta, and Rai, Abhishek K. 2023. Estimating seismic vulnerability in West Bengal by AHP-WSM and AHP-VIKOR. *Natural Hazards Research*, 3(3), 464-473.
- Manikandan, K., Nidhi, M., Micelli, Francesco, Cascardi, Alessio, and Sivasubramanian, Madappa V.R. 2025. Seismic vulnerability assessment of slender masonry structures: A machine learning approach with vibration-based assessment and nonlinear dynamic analysis. *Structures*, 79, 109272-109272. <https://doi.org/10.1016/j.Istruc.2025.109272>
- Moroz, Cassiano Bastos, and Thieken, Annegret H. 2026. Urban morphology as a proxy for housing and infrastructure inequality: A machine learning approach using open building footprint data. *Computers, Environment and Urban Systems*, 126, 102402-102402. <https://doi.org/10.1016/j.Compenvurbsys.2026.102402>
- Muhammad, Shahani Niaz, Xigui, Zheng, Cancan, Liu, Ul, Hassan Fawad, and Peng, Li. 2021. Developing an XGBoost Regression Model for Predicting Young's Modulus of Intact Sedimentary Rocks for the Stability of Surface and Subsurface Structures. *Frontiers in Earth Science*, 9. <https://doi.org/10.3389/feart.2021.761990>
- Nishizaki, Hiromitsu. 2017. Data augmentation and feature extraction using variational autoencoder for acoustic modeling. 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC),
- Papadopoulos, Dimitris, and Karalis, Vangelis D. 2023. Variational Autoencoders for Data Augmentation in Clinical Studies. *Applied Sciences*, 13(15). <https://doi.org/10.3390/app13158793>
- Pinasco, Silvia, Lagomarsino, Sergio, Oneto, Luca, Coraddu, Andrea, and Cattari, Serena. 2025. Training and validation of a machine learning model for seismic vulnerability of masonry buildings in aggregate and selection of key parameters. *Structures*, 80, 109773-109773. <https://doi.org/10.1016/j.Istruc.2025.109773>
- Research, Department of Electrical Engineering Institute of Infrastructure Technology, Management, Ahmedabad, 380026, India, Research, Department of Electrical Engineering Institute of Infrastructure Technology, Management, Ahmedabad, 380026, India, and Department of Energy Technology Aalborg University, Aalborg East, Denmark. 2019. Hybrid machine intelligent SVR variants for wind forecasting and ramp events. *Renewable and Sustainable Energy Reviews*, 108, 369-379. <https://doi.org/10.1016/j.rser.2019.04.002>
- Sun, Xiaoping, Zeng, Peng, Li, Tianbin, Xu, Qiang, Zhao, Hua, and Wang, Xin. 2025. Probabilistic

vulnerability evaluation of buildings under landslide runout impacts considering the 3D dynamic interaction processes. *Engineering Geology*, 354, 108164-108164. <https://doi.org/10.1016/j.Enggeo.2025.108164>

Tian, Chenxi, Ma, Yuliang, Cammon, Jared, Fang, Feng, Zhang, Yingchun, and Meng, Ming. 2023. Dual-encoder VAE-GAN with spatiotemporal features for emotional EEG data augmentation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31, 2018-2027.

Wang, Wei, Zhang, Xitang, Zhang, Xinyu, Liu, Wenyi, Zheng, Hao, and Liu, Yaohui. 2026. Integrated seismic damage scenario and resilience assessment using multimodal data and support vector machine, a case study of Jiangyou City, Southwestern China. *Geomatics, Natural Hazards and Risk*, 17(1), 2607461.

Wei, Chen, and Limao, Zhang. 2022. Building vulnerability assessment in seismic areas using ensemble learning: A Nepal case study. *Journal of Cleaner Production*, 350. <https://doi.org/10.1016/j.jclepro.2022.131418>

Wen, Liu, Jie, Zhou, Xiaoli, Li, Hao, Zheng, and Yaohui, Liu. 2024. Urban resilience assessment and its spatial correlation from the multidimensional perspective: A case study of four provinces in North-South Seismic Belt, China. *Sustainable Cities and Society*, 101, 105109-. <https://doi.org/10.1016/j.Scs.2023.105109>

Yadim, Yavuz, Yilmaz, Salih, Kaplan, Hasan, Corradi, Marco, and Castori, Giulio. 2026. Seismic vulnerability of deficient RC buildings: Insights from the February 6, 2023, Turkey earthquake. *Procedia Structural Integrity*, 78, 1895-1902.

Yang, Liu, Bo, Shu, Yang, Chen, and Hao, Zhang. 2023. Spatial vulnerability assessment of rural settlements in hilly areas using BP neural network algorithm. *Ecological Indicators*, 157. <https://doi.org/10.1016/j.Ecolind.2023.111278>

Yang, Xin, Hao, Yulu, Ding, Hanyue, Yu, Changshui, Liu, Jia, Li, Linyao, and Chen, Junmin. 2025. Explainable Artificial Intelligence (XAI) framework using XGBoost and SHAP for assessing urban fire risk based on spatial distribution features. *International Journal of Disaster Risk Reduction*, 129, 105798-105798. <https://doi.org/10.1016/j.Ijdr.2025.105798>

Yaohui, Liu, Xinyu, Zhang, Wenyi, Liu, Yu, Lin, Fei, Su, Jian, Cui, Benyong, Wei, Hao, Cheng, and Lutz, Gross. 2023. Seismic vulnerability and risk assessment at the urban scale using support vector machine and GIScience technology: a case study of the Lixia District in Jinan City, China. *Geomatics, Natural Hazards and Risk*, 14(1). <https://doi.org/10.1080/19475705.2023.2173663>

Zain, Muhammad, Dackermann, Ulrike, and Prasittisopin, Lapyote. 2024. Machine learning (ML) algorithms for seismic vulnerability assessment of school buildings in high-intensity seismic zones. *Structures*, 70, 107639-107639. <https://doi.org/10.1016/j.Istruc.2024.107639>

Zhang, Wanying, Wen, Haijia, Chu, Yongbo, Zhang, Chi, and Qian, Long. 2026. Deciphering decision-making mechanisms and obstacles for the building pressure-state resilience in mountain city: A case study on a SMOTE-XGBoost- SHAP hybrid model. *Journal of Building Engineering*, 117, 114685-114685. <https://doi.org/10.1016/j.Job.2025.114685>

Zhang, Xinmeng, Yan, Chao, Gao, Cheng, Malin, Bradley A, and Chen, You. 2020. Predicting missing values in medical data via XGBoost regression. *Journal of healthcare informatics research*, 4(4), 383-394.

Zhi, Xia, Yuexin, Song, Jin, Ma, Lujie, Zhou, and Zijian, Dong. 2017. Research on the Pearson correlation coefficient evaluation method of analog signal in the process of unit peak load regulation. 2017 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI), 中国江苏扬州.

- 949 Zhou, Haomiao, Deng, Zhihong, Xia, Yuanqing, and Fu, Mengyin. 2016. A new sampling method in
950 particle filter based on Pearson correlation coefficient. *Neurocomputing*, 216, 208-215.
951 <https://doi.org/10.1016/j.neucom.2016.07.036>
952
- 953 Zhou, Wenting, Li, Yunyan, Gou, Menghan, Wang, Binyan, and Fang, Chenhao. 2025. Urban spatial
954 vulnerability analysis based on urban systems using support vector machine. *Sustainable Cities*
955 *and Society*, 123, 106274-106274. <https://doi.org/10.1016/j.Scs.2025.106274>
956
- 957 Zhu, Wei, Ma, Ding, He, Biao, Lu, Xiaomin, Chen, Yebin, Ying, Jing, Li, Chengpeng, Li, Lin, and Guo,
958 Renzhong. 2026. Examining the spatial configuration of street junctions across cities: a fractal
959 approach on intra-urban clusters. *International Journal of Applied Earth Observation and*
960 *Geoinformation*, 146, 105021-105021. <https://doi.org/10.1016/j.Jag.2025.105021>
961
962