Identification of PGA Outliers Using the IQR Method to Assess Potential Local Amplification

Annisa Aulia Putri[1], **Lailatul Husna Lubis**[1], & **Russell Ong**[1]
[1]Physics Study Program, Universitas Islam Negeri Sumatera Utara, Indonesia
*Corresponding Author: lailatulhusnalubis@uinsu.ac.id

Received: 12th August 2026; **Accepted**: 9th November 2026; **Published**: 10th November 2026
DOI: https://dx.doi.org/10.29303/jpft.v12i2.13119

Abstract - *Peak Ground Acceleration (PGA) is an important parameter in earthquake hazard analysis. However, its values may deviate from general trends due to local site conditions. This study aims to analyze the distribution of PGA recorded by BMKG stations in Northern Sumatra during the 2017–2025 period, detect statistical outliers in relative PGA residuals using the Interquartile Range (IQR) method and boxplots, and identify preliminary indications of potential local amplification based on Vs30 values obtained from USGS maps. The analysis used 5,360 records from 501 events recorded by at least three stations, following data cleaning and event selection from an initial dataset of 5,567 records. Horizontal amplitudes were calculated from the east–west and north–south components, transformed using the natural logarithm, and then compared with the median ln(PGA) of each event as a relative reference to obtain relative residuals. The initial PGA distribution was highly right-skewed (skewness = 19.71; kurtosis = 686.97) and became closer to a log-normal distribution after transformation. The IQR analysis yielded Q1 = 0.4782, Q3 = 2.1105, IQR = 1.6323, and an upper bound of 4.5590, identifying 639 records (11.92%) as outliers across 69 stations. ONSM, SKSN, LTSM, PSSN, and SNBI exhibited the highest outlier frequencies. Nine of the ten stations with the highest outlier frequencies belonged to NEHRP site classes C and D, indicating that stiff to soft soil conditions may contribute to elevated PGA responses. However, the relationship between relative residuals and Vs30 was not entirely consistent, suggesting that Vs30 alone is insufficient to explain the observed variations. Relative residuals do not explicitly account for the effects of source-to-station distance, wave propagation paths, or earthquake source characteristics. Therefore, the identified outliers are interpreted as potential anomalies in PGA response rather than confirmed evidence of local site amplification. Further research is recommended using attenuation models or distance and magnitude normalization, along with field-based geophysical investigations to confirm the underlying causes.*

Keywords: *Peak Ground Acceleration (PGA); outlier; Interquartile Range (IQR); relative residual; Vs30; local amplification.*

INTRODUCTION

Indonesia is located at the convergence of three major tectonic plates and lies along the Pacific Ring of Fire, making the country, including Northern Sumatra, highly susceptible to earthquakes (Oktarina, 2008). Recent studies on the seismicity of western Sumatra have further indicated that seismic activity in this region is not only controlled by the megathrust zone and the Sumatran Fault but may also be influenced by back-arc thrust faulting mechanisms, which have received relatively little attention in previous assessments (Widiyantoro et al., 2024). Such complexity in earthquake sources makes it difficult to generalize the ground shaking observed at the surface based on a single source mechanism.

One of the commonly used parameters to characterize the intensity of earthquake-induced ground shaking is Peak Ground Acceleration (PGA). PGA values are influenced by the earthquake source, wave propagation path, and local site conditions (Nurjannah et al., 2023). Under certain site conditions, seismic waves may be amplified, resulting in PGA values substantially higher than the general trends observed at surrounding locations. This phenomenon is known as local amplification.

Ideally, exceptionally high PGA values recorded at a particular station can serve as an initial indicator of potential local site amplification and can therefore support regional-scale disaster mitigation efforts (Boore, 2004). In practice, however, observed PGA data generally do not follow a normal distribution but are highly right-skewed due to the log-normal nature of ground motion. Consequently, outlier detection methods that require a normality assumption may be less appropriate for direct application. The Interquartile Range (IQR) method was selected as an alternative approach because it is robust to non-normal data distributions (Tukey, 1977) and is relatively insensitive to the extreme values themselves (Dovoedo & Chakraborti, 2015).

Furthermore, the average shear-wave velocity of the uppermost 30 meters of soil (Vs30) has been widely used as a proxy for site conditions and as the basis for the NEHRP site classification system (Borcherdt, 1994). Ideally, stations with low Vs30 values (soft soil) are expected to exhibit higher PGA responses than stations with high Vs30 values (hard soil). However, several recent studies have begun to question the adequacy of Vs30 as the sole predictor of amplification. For instance, a study conducted on the West Bohemia seismic network found that amplification values exhibited a wide range of variation even among sites belonging to the same site class, indicating the need for additional site parameters beyond Vs30 to explain variations in site response (Mazanec et al., 2024). Heterogeneous site conditions have also been documented across various regions of Sumatra. For example, seismic microzonation studies in Bengkulu City have revealed considerable variations in predominant frequency and amplification factors due to the dominance of soft alluvial deposits. These findings emphasize that site conditions across Sumatra are not uniform and may produce different ground-motion responses from one location to another.

Therefore, a gap exists between the ideal assumption that Vs30 can consistently explain variations in PGA and actual field conditions, which indicate that the relationship between these two parameters is not always linear. In addition, the application of the IQR method to PGA data in Indonesia, particularly to relative residuals for identifying potential local site amplification, remains limited. Most previous studies have primarily employed IQR as a data-cleaning technique to remove outliers rather than interpreting these outliers as potential physical indicators of site conditions.

Unlike previous studies that generally treat outliers as data points that should be discarded, this study retains the outliers detected using the IQR method for further analysis, as these values may reflect anomalous ground-motion responses associated with local site conditions. This study also combines relative PGA residuals with Vs30 data as a simple and objective preliminary approach to identifying potential indications of local site amplification in Northern Sumatra, while simultaneously examining the extent to which Vs30 can explain variations in PGA responses among stations based on observational data from 2017–2025.

This study aims to: (1) analyze the distribution characteristics of observed PGA values recorded by BMKG stations during the 2017–2025 period; (2) detect statistical outliers in relative PGA residuals using the IQR method and boxplot visualization; and (3) identify potential indications of local amplification at outlier stations based on their Vs30 values.

RESEARCH METHODS

Study Area

This study was conducted in Northern Sumatra (Sumbagut), Indonesia, geographically located between 2°N–6°S and 95°–109°E. This region was selected due to its high level of seismic activity. The distribution of 183 BMKG seismic stations across the study area was mapped using QGIS 3.42.2, as presented in Figure 1.

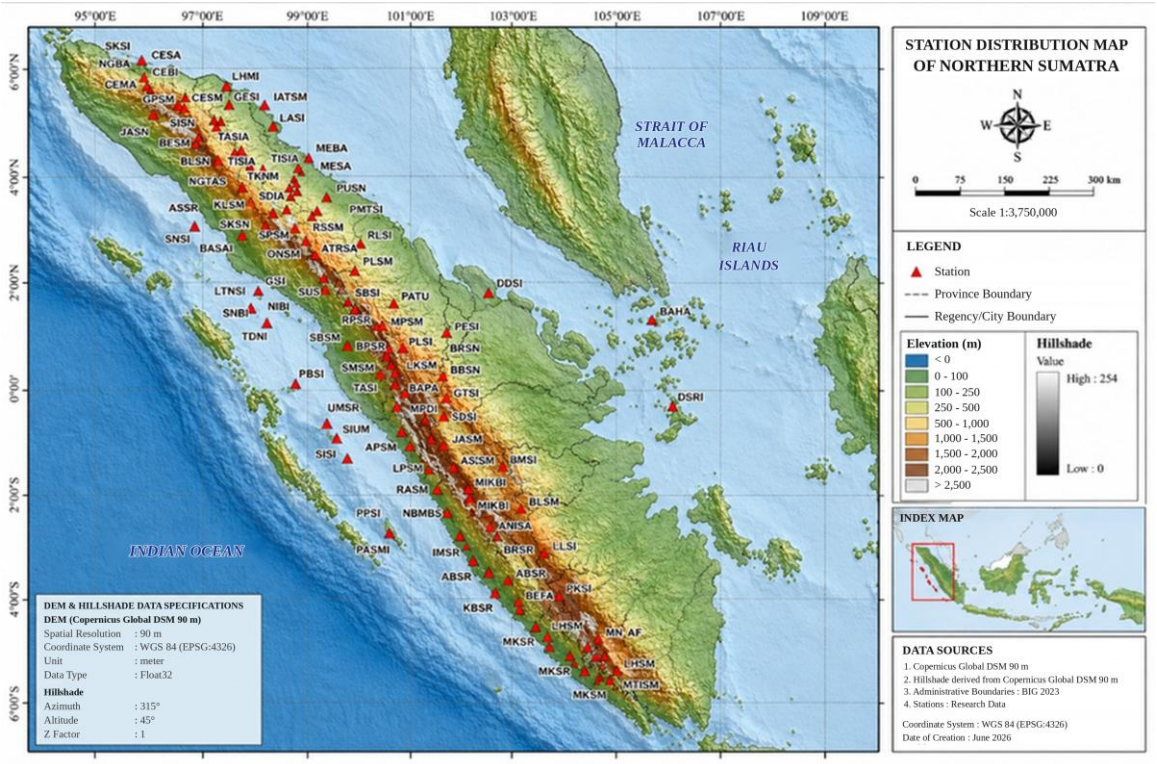


Figure 1. Distribution of BMKG Seismic Stations in Northern Sumatra

The data used in this study consist of ground acceleration records obtained from the BMKG seismic station network in Northern Sumatra and its surrounding areas during the 2017–2025 period. The dataset comprises 5,567 records from 183 stations associated with 628 earthquake events. The available data include event occurrence time, station codes, station coordinates, source-to-station distances, recorded amplitudes, magnitudes, focal depths, and epicentral coordinates. Following data cleaning and the selection of earthquake events recorded by at least three stations, a total of 5,360 records from 501 events were retained for analysis.

The horizontal amplitude (AmpH) was calculated from the east–west (AmpE) and north–south (AmpN) components using the Square Root of Sum of Squares approach (Salazar et al., 2005):

$$AmpH = \sqrt{AmpE^2 + AmpN^2} \quad (1)$$

The AmpH values were transformed using the natural logarithm to stabilize the data distribution. The general PGA trend for each earthquake event was determined from the median ln(PGA) across all stations recording that event, which was then converted back to the linear scale. The relative residual (R) was calculated as the ratio of the observed PGA to the general PGA trend for each earthquake event:

$$R = \frac{PGA_{obs}}{PGA_{pred\,(Median,event)}} \quad (2)$$

The median ln(PGA) for each earthquake event was used as the general

trend because the median is robust to extreme values (AlSalehy & Bailey, 2025). Consequently, it is not disproportionately influenced by stations recording exceptionally high PGA values. Furthermore, residuals calculated relative to an event-specific reference value are commonly used to compare responses among stations (Douglas & Edwards, 2016). All stations recording the same earthquake event share the same magnitude, focal depth, and source mechanism. Therefore, normalization against the event-specific median reduces the influence of source characteristics and allows residuals to be compared across different events without requiring an attenuation model. A minimum requirement of three stations per event was applied to ensure that the median was sufficiently representative (Douglas, 2003).

Nevertheless, this approach does not explicitly account for differences in source-to-station distance, wave propagation paths, or rupture directivity effects (Somerville et al., 1997). Consequently, stations located close to the earthquake source may exhibit high residuals that are not necessarily associated with local site conditions. Therefore, the identified outliers are treated as potential anomalies in PGA response rather than confirmed evidence of local site amplification. Analyses controlling for the effects of distance and other factors are beyond the scope of this study and are acknowledged as a limitation.

The site conditions at each station were classified based on Vs30 values using the NEHRP classification system (Borcherdt, 1994), which is comparable to SNI 1726:2019 (BSN, 2019), as presented in Table 1. The Vs30 values were obtained from the USGS global map based on a topographic slope proxy (Wald & Allen, 2007).

Table 1. Site Classification Based on Vs30 (Borcherdt, 1994 & BSN, 2019)

NEHRP Class	SNI Site Class	Site Description	Vs30 (m/s)
A	–	Hard Rock	> 1500
B	SA	Rock	760–1500
C	SB	Very Dense Soil and Soft Rock	360–760
D	SC	Stiff Soil	180–360
E	SD	Soft Soil	< 180
–	SE	Very Soft Soil	< 100

Outliers in the relative residual data were detected using the Interquartile Range (IQR) method (Tukey, 1977), which is robust to non-normal data distributions (Dovoedo & Chakraborti, 2015). Observations falling above the upper fence were classified as outliers and visualized using boxplots. Subsequently, the mean relative residuals at outlier stations were compared with their respective Vs30 values as supporting information to evaluate preliminary indications of potential local site amplification among the identified anomalous candidates. The relationship between the mean relative residuals and Vs30 was analyzed for the 20 stations with the highest outlier frequencies (at least 11 outliers per station) using Spearman's rank correlation.

RESULTS AND DISCUSSION

PGA Distribution Characteristics

Of the 5,567 records originating from 628 earthquake events and 183 stations, 23 records were removed because their AmpE and AmpN components were missing, making it impossible to calculate the horizontal amplitudes. This left 5,544 valid records, corresponding to a data loss of 0.41%. All valid records had positive PGA values, with a minimum value of 0.0009 gal. Subsequently, 184 records were excluded

because they originated from earthquake events recorded by fewer than three stations. Thus, a total of 207 records were excluded, and the relative residual and IQR analyses were conducted using 5,360 records from 501 earthquake events.

The minimum requirement of three stations was applied because the general trend median would not be representative if calculated from only one or two records (Douglas, 2003). For events recorded by only one station, the relative residual would always equal 1 and therefore provide no meaningful information. The PGA distribution discussed in this subsection was analyzed using the 5,360 records retained for the analysis.

The observed PGA values (N = 5,360) exhibited a highly right-skewed distribution, with a skewness of 19.71 and a kurtosis of 686.97. Following natural logarithmic transformation, the skewness decreased to 0.45 and the kurtosis to 0.19, resulting in a more symmetrical distribution that more closely approximated a normal distribution, as shown in Figure 2.

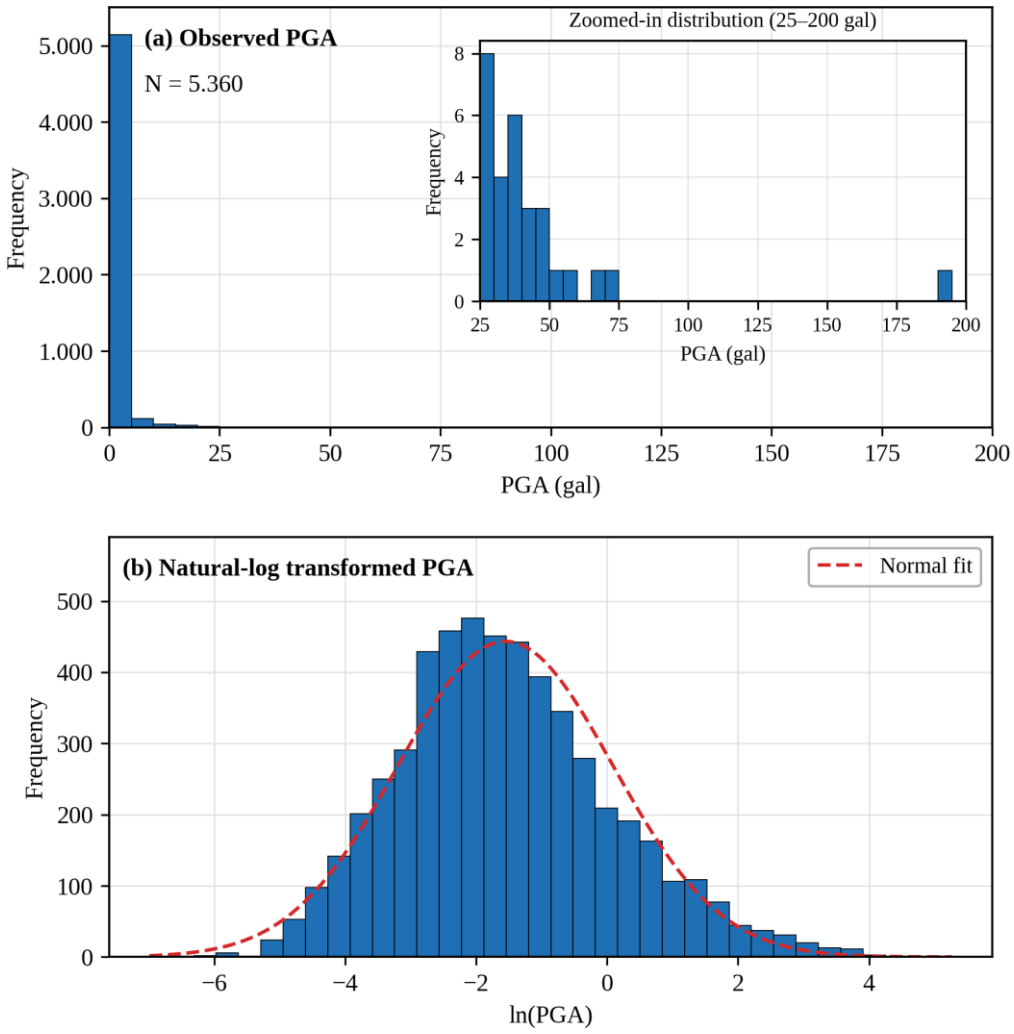


Figure 2. Distribution of PGA Values Before (a) and After (b) Natural Logarithmic Transformation (N = 5,360)

The highly skewed distribution of PGA data is commonly observed in ground-motion datasets, as most records exhibit relatively low ground-shaking values, whereas only a small proportion of records have exceptionally high values. Natural logarithmic transformation is applied to reduce this skewness, resulting in a more

balanced data distribution. This condition also supports the use of quartile-based methods such as IQR, as these methods do not require the data to follow a normal distribution (Tukey, 1977; Dovoedo & Chakraborti, 2015).

Outlier Detection Using the IQR Method

A total of 5,360 relative PGA residual records, originating from earthquake events recorded by at least three BMKG stations, were analyzed using the IQR method. The analysis yielded Q1 = 0.4782, Q3 = 2.1105, IQR = 1.6323, and an upper fence of 4.5590. The lower fence of −1.9703 was below the minimum possible value of the relative residuals; therefore, no lower outliers were identified. Based on the upper fence, 639 observations (11.92%) were identified as outliers, distributed across 69 BMKG stations. The ten stations with the highest outlier frequencies are presented in Table 2.

Table 2. Ten Stations with the Highest Outlier Frequencies

Station Code	Number of Outliers	Mean R	Maximum R
ONSM	52	16.43	90.95
SKSN	47	11.06	30.13
LTSM	37	17.29	68.90
PSSN	35	14.26	91.12
SNBI	29	15.61	68.11
MLSI	27	12.33	44.50
SASM	24	12.42	41.18
BPSM	24	16.23	45.63
SPSM	23	12.22	49.56
ASSR	21	51.42	359.85

ONSM recorded the highest number of outliers, with 52 occurrences, followed by SKSN with 47 occurrences, LTSM with 37 occurrences, and PSSN with 35 occurrences. However, outlier frequency does not necessarily indicate the magnitude of the residual values. For instance, ASSR had only 21 outlier occurrences but exhibited the highest mean relative residual of 51.42 and the highest maximum residual of 359.85 among the ten stations. This indicates that the frequency of outlier occurrence and the magnitude of residual values represent two distinct characteristics that should be considered together.

The overall distribution of relative residual data and the identified outliers is presented in Figure 3. Most observations fall within the interquartile range, whereas 639 observations lie above the upper fence and are classified as upper outliers, with no lower outliers identified.

The absence of lower outliers indicates that all deviations in relative residuals identified in this study correspond to values higher than the general trend, rather than lower values.

The repeated occurrence of outliers at several stations, such as ONSM and SKSN, indicates that high relative residual values do not occur merely as isolated events but tend to recur at the same locations. This condition suggests that elevated relative residuals may be associated with local site conditions or other physical factors influencing ground-motion responses at these locations.

It should be noted that the relative residuals used in this study do not explicitly account for the effects of source-to-station distance, wave propagation paths, or other earthquake source characteristics. Consequently, the identified outliers cannot be directly interpreted as evidence of site amplification, since stations located close to earthquake sources may exhibit high residuals even when their site conditions are not different. Therefore, the outliers identified in this study are treated as potential anomalies in PGA response that require further confirmation. Separating source, path, and site effects requires the use of a ground-motion prediction equation (GMPE) or normalization based on distance and magnitude in future research.

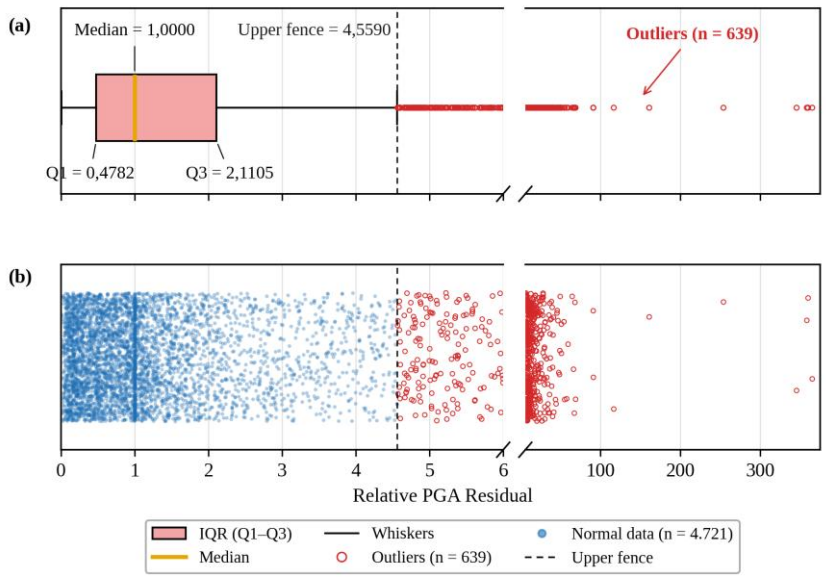


Figure 3. Boxplots of Relative PGA Residuals Identified Using the IQR Method: (a) boxplot, (b) distribution of individual data points. The horizontal axis uses two scales (0–6 and 6–375).

Relationship Between Relative Residuals and Vs30

A comparison of the mean relative residuals at the 20 stations with the highest outlier frequencies and their corresponding Vs30 values is presented in Figure 4. STSR exhibited the highest mean relative residual, at 93.85, with a Vs30 value of 282 m/s, corresponding to NEHRP Site Class D. ASSR had a mean relative residual of 51.42, with a Vs30 value of 411 m/s, corresponding to NEHRP Site Class C. Meanwhile, BPSM exhibited a mean relative residual of 16.23 despite having the highest Vs30 value of 761 m/s, corresponding to NEHRP Site Class B.

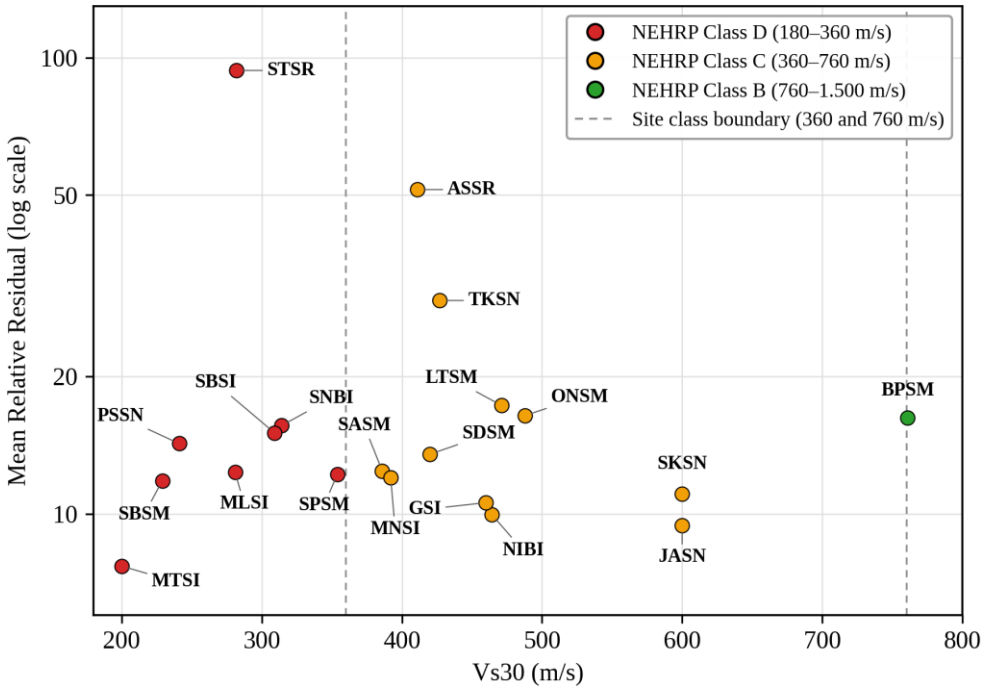


Figure 4. Relationship Between Mean Relative Residuals and Vs30 at the 20 Stations with the Highest Outlier Frequencies (the mean was calculated using only outlier records at each station; colors indicate NEHRP site classes)

In general, several stations with lower Vs30 values exhibited higher relative residuals. This observation is consistent with the understanding that soft to moderately stiff soils may experience greater earthquake ground-motion amplification than harder soil conditions (Borcherdt, 1994). This finding is also supported by a spectral analysis report of the Sinabang, Aceh, earthquake, which indicated that the STSR and ASSR locations are predominantly underlain by soft to moderately stiff soils with relatively thick sediment layers (BMKG, 2025). Nevertheless, the proximity of both stations to the earthquake source may also have contributed to the high ground-motion levels recorded.

On the other hand, this relationship was not observed consistently across all stations. BPSM represents one example of a contrasting condition. Based on its Vs30 value of 761 m/s, BPSM belongs to NEHRP Site Class B, indicating relatively hard ground conditions. However, the mean relative residual at this station remained relatively high at 16.23. This finding indicates that high Vs30 values are not necessarily associated with low relative residuals. Spearman's rank correlation test conducted on the 20 stations also revealed a very weak and statistically insignificant relationship between Vs30 and the mean relative residuals ($\rho = 0.0444$; $p = 0.8526$).

To further investigate the conditions at BPSM, predominant spectral period information from BMKG reports was examined. During the Mentawai Islands earthquake, BPSM was recorded as having a predominant spectral period of 0.71 seconds (BMKG, 2022). According to the Kanai–Omote-Nakajima soil classification described by Varecha et al. (2022), predominant periods exceeding 0.40 seconds correspond to Type IV or Type C, indicating very soft soil conditions. Thus, the predominant period recorded at BPSM suggests site conditions that differ from those indicated by its Vs30-based classification. This discrepancy indicates that the Vs30 values used in this study may not fully characterize all site conditions surrounding the BPSM station.

Similar findings have also been reported in a study of the West Bohemia seismic network in the Czech Republic. The study demonstrated that amplification levels may vary among stations belonging to the same Vs30 site class. This finding indicates that Vs30 alone may not fully capture differences in soil responses across individual locations (Mazanec et al., 2024).

Based on these findings, Vs30 is insufficient as the sole parameter for determining variations in PGA responses among stations, particularly those identified as outliers. Vs30 provides an indication of near-surface soil stiffness; however, recorded PGA values are also influenced by earthquake source characteristics (Boore, 2004), source-to-station distance, faulting mechanisms, wave propagation paths, and local geological conditions (Douglas & Edwards, 2016). Therefore, elevated relative residuals at outlier stations cannot always be explained solely by their Vs30 values. These factors need to be considered together with Vs30 to understand the causes of elevated PGA responses and evaluate the potential for local amplification.

CONCLUSION

This study applied the Interquartile Range (IQR) method to identify statistical outliers in relative Peak Ground Acceleration (PGA) residuals recorded by BMKG stations in Northern Sumatra during the 2017–2025 period. The initially

observed PGA values exhibited a highly right-skewed distribution, whereas natural logarithmic transformation produced a more symmetrical distribution suitable for statistical analysis. The IQR analysis identified 639 upper outliers (11.92%) out of 5,360 relative residual records, with no lower outliers detected. These outliers were distributed across 69 stations, indicating that elevated relative residuals were not confined to a single location.

Comparison with Vs30 values showed that most stations with the highest outlier frequencies belonged to NEHRP Site Classes C and D. This condition suggests that moderately stiff to soft soils may contribute to elevated PGA responses at some outlier stations and provides preliminary indications of potential local amplification at candidate stations that still require further confirmation. However, the high relative residual observed at BPSM, which belongs to NEHRP Site Class B, indicates that site conditions alone cannot fully explain the observed anomalies. Furthermore, the relative residuals used in this study do not explicitly account for source-to-station distance, wave propagation paths, or earthquake source characteristics. Consequently, these factors, together with local geological conditions, may also influence the relative residual values.

Therefore, the IQR method provides a simple statistical approach for the preliminary identification of stations exhibiting anomalous PGA responses, allowing these stations to be treated as potential anomaly candidates. Further investigations incorporating controls for distance, magnitude, and wave propagation paths, together with field-based geophysical studies, are necessary to confirm the underlying causes.

REFERENCES

AlSalehy, A. S., & Bailey, M. (2025). Improving Time Series Data Quality: Identifying Outliers and Handling Missing Values in a Multilocation Gas and Weather Dataset. *Smart Cities*, *8*(3). https://doi.org/10.3390/smartcities8030082

BMKG. (2022). *Ulasan Guncangan Tanah Akibat Gempa Kepulauan Mentawai Sumatera Barat*.

BMKG. (2025, November 28). *Analisis Spektral Gempabumi Sinabang Aceh 27 November 2025*. Badan Meteorologi, Klimatologi, dan Geofisika.

Boore, D. M. (2004). Can Site Response Be Predicted? *Journal of Earthquake Engineering*, *8*(1), 1–41. www.worldscientific.com

Borcherdt, R. D. (1994). Estimates of Site-Dependent Response Spectra for Design (Methodology and Justification). *Earthquake Spectra*, *10*(4), 617–653.

BSN. (2019). *Tata Cara Perencanaan Ketahanan Gempa untuk Struktur Bangunan Gedung dan Nongedung. SNI 1726:2019. Indonesia.*

Douglas, J. (2003). Earthquake Ground Motion Estimation Using Strong Motion Records: a Review of Equations for the Estimation of Peak Ground Acceleration and Response Spectral Ordinates. *Earth-Science Reviews*, 44–104. www.elsevier.com/locate/earscirev

Douglas, J., & Edwards, B. (2016). Recent and Future Developments in Earthquake Ground Motion Estimation. *Earth-Science Reviews*, *160*, 203–213.

https://doi.org/10.1016/j.earscirev.2016.07.005

Dovoedo, Y. H., & Chakraborti, S. (2015). Boxplot Based Outlier Detection for the Location Scale Family. *Communications in Statistics: Simulation and Computation*, *44*(6), 1492–1513. https://doi.org/10.1080/03610918.2013.813037

Mazanec, M., Valenta, J., & Málek, J. (2024). Does VS30 Reflect Seismic Amplification? Observations from the West Bohemia Seismic Network. *Natural Hazards*, *120*(13), 12181–12202. https://doi.org/10.1007/s11069-024-06679-x

Nurjannah, R., Iskandar, M., & Lubis, L. H. (2023). Identifikasi Percepatan Tanah Maksimum Di Wilayah Bengkulu Dengan Metode Kawashumi. *Komunikasi Fisika Indonesia*, *20*(3), 237–241.

Oktarina, R. (2008). Pemetaan Sistem Informasi Manajemen Logistik Dalam Penanggulangan Bencana Di Indonesia. Dalam *Seminar Nasional Aplikasi Teknologi Informasi*.

Salazar, A. R., Duarte, J. A. J., Barraza, A. L., Morán, J. de D. G., & Dimas, J. I. V. (2005). Evaluación De Las Reglas Más Comunes De Combinación De Los Efectos De Las Componentes Horizontales De Un Terremoto. *Revista de Ingeniería Sísmica*, *73*(73), 27–53.

Somerville, P. G., Smith, N. F., Graves, R. W., & Abrahamson, N. A. (1997). Modification of Empirical Strong Ground Motion Attenuation Relations to Include the Amplitude and Duration Effects of Rupture Directivity. *Seismological Research Letters*, *68*(1), 199–222.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley Publishing Company.

Varecha, P. V., Fadllan, A., & Hartono, H. (2022). Soil Classification Using Horizontal to Vertical Spectrum Ratio Methods on Scilab in Sendangmulyo, Semarang. *Physics Education Research Journal*, *4*(1), 71–78. https://doi.org/10.21580/perj.2022.4.2.12542

Wald, D. J., & Allen, T. I. (2007). *Topographic Slope as a Proxy for Seismic Site-Conditions (VS30) and Amplification Around the Globe*. http://www.usgs.gov/pubprod

Widiyantoro, S., Supendi, P., Rawlinson, N., Daryono, M. R., & Rosalia, S. (2024). A Note on The Seismicity of Sumatra, Western Sunda Arc, Indonesia, in Relation to The Potential for Back-Arc Thrusting. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-64076-7