

Nonlinear seismic site response classification using K-means clustering algorithm: Case study of the September 6, 2018 Mw6.6 Hokkaido Iburi-Tobu earthquake, Japan

Kun Ji^a, Ruizhi Wen^{a,*}, Yefei Ren^a, Yadab P. Dhakal^b

^a Key Laboratory of Earthquake Engineering and Engineering Vibration, Institute of Engineering Mechanics, China Earthquake Administration, No. 29 Xuefu Road, Harbin, Heilongjiang, 150080, China

^b National Research Institute for Earth Science and Disaster Resilience, 3-1, Tennodai, Tsukuba, Ibaraki, 305-0006, Japan

ARTICLE INFO

Keywords:

Nonlinear site response
Spectral ratio method
K-means clustering algorithm
2018 Hokkaido Iburi-Tobu earthquake

ABSTRACT

Spectral ratio methods have been widely used in evaluation of nonlinear seismic site response. Nevertheless, it remains inefficient and subjective to identify stations with nonlinear site response according to empirical threshold values of spectral ratio nonlinear degree indicators. This study, which was the first to apply the machine learning clustering algorithm to address this problem, used the September 6, 2018 M_w 6.6 Hokkaido Iburi-Tobu earthquake (Japan) as an example. First, we calculated the surface/borehole and horizontal/vertical spectral ratios using strong ground motion data recorded by KiK-net vertical array and K-NET stations, respectively. The degree of nonlinear site response (DNL) and percentage of nonlinear site response (PNL) were computed using the difference between the strong motion of the mainshock and weak aftershocks as the reference for linear site response. Then, the K-means clustering algorithm was incorporated in the identification of nonlinear site response using the DNL, PNL, strength of ground motion (PGA) and site condition (V_{S20} or V_{S30}) as explanatory variables. After careful multicollinearity diagnosis and confirmation of the optimum clustering number, we successfully classified the stations into two clusters with nonlinear and linear site responses. Overall, the clustering results were found in good agreement with the classification results based on empirical thresholds of several nonlinear indicators. For the stations identified with nonlinear site response, the reduction of amplification and frequency shift could be observed from the spectral ratio curves regarding the ground motions in the mainshock and the reference weak aftershocks, demonstrating typical nonlinearity response characteristics. Furthermore, a comprehensive indicator of nonlinear site response occurrence probability (NL_{score}) was obtained from a linear weighted combination of the normalized variables (PGA, V_{S30}/V_{S20} , DNL and PNL). The NL_{score} ranking of the top several stations was found consistent with the clustering identification results, irrespective of the choice of combination scheme. It was demonstrated that the performance of clustering algorithm in this application was satisfactory and that the identification results were convincing and robust. This work provides an enlightening example of using state-of-art machine learning technique to solve the traditional earthquake engineering problems.

1. Introduction

It is widely recognized that the seismic response characteristics of surface soft soil become nonlinear when the soil is struck by strong motion. Real strong-motion records of the 1994 Northridge Earthquake in the USA provide direct evidence of nonlinear site response [1-3], where the amplitude of near-field ground motions was observed to be lower than prediction value of attenuation relationship or simulation

results for “soft” soil condition. Further similar evidence had been observed for many subsequent strong earthquake events [4-6]. Laboratory-based soil dynamics tests have indicated that this phenomenon is attributed to the rise of the damping factors and the reduction in shear rigidity regarding the large strain level under strong input ground motions [7, 8]. Evidence from numerous earthquakes has shown that the nonlinear behavior of soil could be indicated and evaluated using comparative spectral ratio curves computed from weak and strong

* Corresponding author.

E-mail addresses: jikun@iem.ac.cn (K. Ji), ruizhi@iem.net.cn (R. Wen), renyefei@iem.net.cn (Y. Ren), ydhakal@bosai.go.jp (Y.P. Dhakal).

<https://doi.org/10.1016/j.soildyn.2019.105907>

Received 19 February 2019; Received in revised form 27 June 2019; Accepted 15 October 2019

Available online 26 October 2019

0267-7261/© 2019 Elsevier Ltd. All rights reserved.

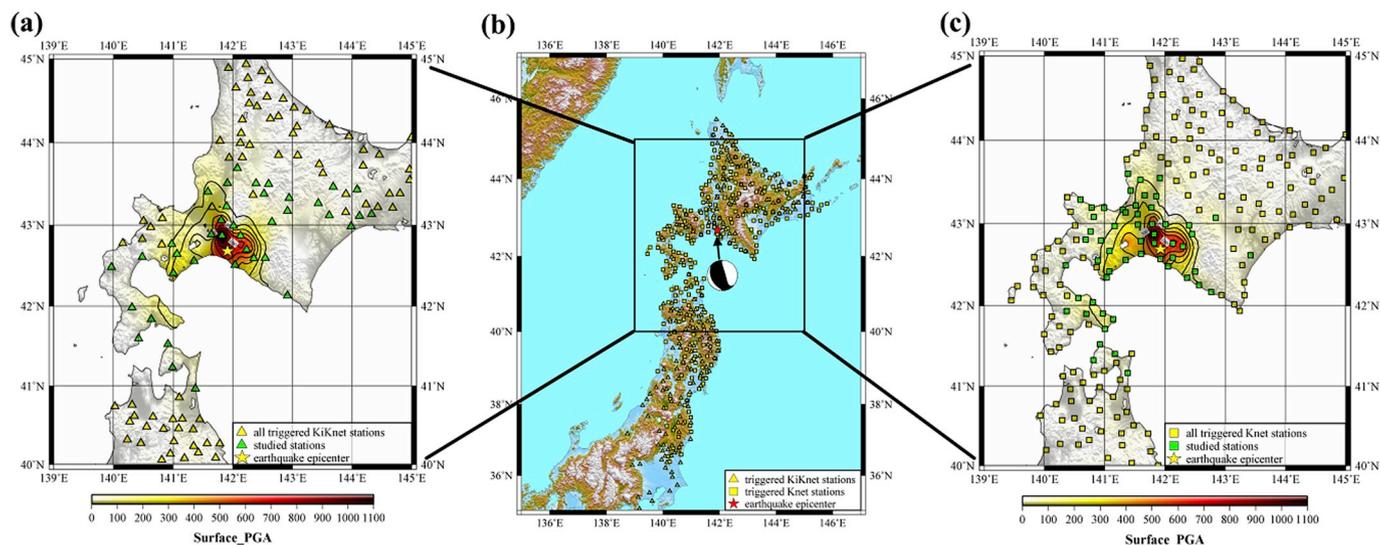


Fig. 1. A surface PGA contour map of the studied region regarding KiK-net and K-net stations were shown in Fig(a) and Fig(c) respectively. Locations of the studied KiK-net and K-NET stations and mainshock epicenter of the 2018 Hokkaido Iwate-Tobu earthquake were shown in Fig(b); the preliminary focal mechanism solution provided by USGS is also shown.

motions [9–14]. When the ground motion input exceeds a certain threshold, the shift of resonant frequencies toward lower values and a reduction in the associated amplification would be clearly observed in the spectral ratio curves.

Vertical arrays of accelerometers are the primary choice of data source for the estimation of seismic site response. This is because their use both avoids the difficulty of selecting reference rock sites and provides more stable spectral ratio curves. On the other hand, it is confirmed that the spectral ratios of horizontal-to-vertical (H/V) surface ground motion could be used as an alternative for sites without vertical borehole measurements or in areas where it might be difficult to establish usable reference sites [9,10]. To quantify the difference between linear and nonlinear site responses, some indicators have been proposed for the evaluation of the degree of nonlinearity. The parameters include the degree of nonlinearity (DNL) [9,10] and percentage of nonlinearity (PNL) [12]. We explain the above parameters in Section 3 of this manuscript.

Although both the DNL and PNL parameters have been used commonly in practice, it is difficult to distinguish definitively between sites with and without soil nonlinearity based only on their values. This is partly because the empirical thresholds of the DNL and the PNL are subjective and change with specific earthquake events and site conditions. In addition, identification results based on empirical fixed thresholds of DNL and PNL are not always the same given the variability of the spectral ratio curves. Moreover, to obtain reliable classification results, both the ground motion input level and the site condition need to be considered. However, the corresponding classification thresholds for both the ground motion strength and the soil condition are also variable and unfixed. In practice, the assessment of nonlinear site response depends on existing knowledge and experience. Therefore, a manual classification process would require considerable time, especially when the number of records is large.

In view of these shortcomings, this study proposes to incorporate the clustering algorithm of machine learning techniques in the identification of seismic nonlinear site response. Clustering analysis is one of the principal tools of exploratory data mining. It is actually not a single specific algorithm but an unsupervised machine learning process that classifies unlabeled similar objects into the same group or cluster [15]. There are many clustering applications available for use in the field of seismology and earthquake engineering. For example, Chen [16] developed an approach to pick seismic wave arrival times using fuzzy clustering, which is based on the idea that the amplitudes of the seismic

data before and after arrival can be treated as separate clusters. Riahi and Gerstoft [17] used a graphical clustering method to locate earthquake sources within a dense sensor array. The clustering analysis is usually considered as a useful tool for solving multi-objective optimization (MP) problems [18] that require simultaneous optimization of more than one objective function. Stations with nonlinear site response often have similar recorded large ground motions, similar soft site conditions and similar features in their spectral ratio curves that could be measured using the DNL or the PNL. These inherent similarities provide solid foundations for the use of clustering analysis in classifying stations with and without nonlinear site response, which could be regarded as a typical MP problem that needs to consider all these factors as variables to obtain comprehensive results.

In this study, the September 6, 2018 M_w 6.6 Hokkaido Iwate-Tobu earthquake [19,20] was taken as an example for analysis of nonlinear seismic site response. Surface/borehole and horizontal/vertical (H/V) spectral ratios were calculated using strong ground motion data recorded by the KiK-net vertical array and K-NET stations, respectively. Then, the DNL and PNL values were computed using weak motions recorded during aftershocks as the reference for linear site response. Finally, the K-means clustering algorithm was introduced and applied in the nonlinear site response identification by considering the DNL, PNL, strength of ground motion (PGA) and site condition (V_{S30} or V_{S20}) as clustering variables.

2. Strong ground motion data and data processing

The September 6, 2018 M_w 6.6 (M_{JMA} , Japan Meteorological Agency (JMA) magnitude, 6.7) Hokkaido Iwate-Tobu earthquake occurred following a shallow reverse rupture (according to the preliminary focal mechanism solution provided by the USGS, <https://earthquake.usgs.gov>). The hypocenter depth of the earthquake was located 37.0 km beneath the island of Hokkaido which were determined by JMA (<https://www.jma.go.jp/jma/indexe.html>). The maximum registered intensity is 7.0 by JMA. This earthquake reportedly killed 41 persons, including 36 dead by nearly 6000 landslides [19]. According to research in Ref. [20], the observed peak ground accelerations (PGAs) and peak ground velocities (PGVs) generally followed the median GMPE values at fault distances ≥ 50 km. However, at smaller distances, the PGA and PGV prediction equations significantly underestimated the observations, which may be induced by nonlinearity site response in the near-field region. Ref [20] proposed an ad hoc equation to correct the nonlinear

Table 1
The Pearson correlation coefficients and VIF values for the clustering variables regarding K-NET observations.

	Surface-PGA	DNL	PNL	V_{S20}	VIF
Surface-PGA	1.00	0.80	0.85	-0.14	4.8
DNL	0.80	1.00	0.74	-0.11	3.0
PNL	0.85	0.74	1.00	-0.18	3.8
V_{S20}	-0.14	-0.11	-0.18	1.00	1.0

Table 2
Same as Table 3 but for KiK-net observations.

case A					
	Surface-PGA	DNL	PNL	V_{S30}	VIF
Surface-PGA	1.00	0.74	0.95	-0.19	10.3
DNL	0.74	1.00	0.80	-0.37	3.3
PNL	0.95	0.80	1.00	-0.20	13.3
V_{S30}	-0.19	-0.37	-0.20	1.00	1.2
case B					
	Borehole-PGA	DNL	PNL	V_{S30}	VIF
Borehole-PGA	1.00	0.76	0.87	-0.15	4.5
DNL	0.76	1.00	0.80	-0.37	3.4
PNL	0.87	0.80	1.00	-0.20	5.2
V_{S30}	-0.15	-0.37	-0.20	1.00	1.2

site amplification in predicting horizontal PGVs with respect to one of the most widely used attenuation models in Japan.

Overall, 208 KiK-net and 268 K-NET stations were triggered by the mainshock. We used recordings from the stations with epicenter distance less than 200 km to investigate the possible nonlinear seismic site response, as shown in Fig. 1. Weak motions were selected from 29

aftershocks for computation of reference linear site responses based on the following two criteria. (1) To remove recordings potentially affected by soil nonlinearity, the geometric mean of the surface PGA of two horizontal components should be less than 30 cm/s^2 (2) At least five records matching criterion (1) must be recorded at each station to ensure a reasonably low level of scattering of the spectral ratio curves. Accordingly, 39 KiK-net stations and 56 K-NET stations were selected for the evaluation of their nonlinear site response during the mainshock. The characteristics of nonlinear site response at station IBUH03 would be detailed analyzed in Reference [20] which would not be further discussed in this paper. It is worth noted that the seismic site response of aftershock weak motions may have been affected by the strong motions during the mainshock and that may have resulted in underestimation of the degree of nonlinearity. But using the aftershock has the advantage of limiting the incident angles nearly similar between the mainshock and aftershocks, thus letting the discrepancies mainly relate to the strength of ground motions. Thus, we think that our results should be applicable to the classification of sites for nonlinearity irrespective of the reference motions such as those during pre-shocks or aftershocks of reasonably weak-motions.

The locations of the studied stations and a PGA contour map are shown in Fig. 1. The station code, PGA (recorded at surface and borehole), and V_{S30} (V_{S20} was used for K-NET) of each station were listed in Table 1 and Table 2 for the KiK-net and K-NET stations, respectively. The surface PGAs correlating with the epicenter distance of the selected records, shown in Fig. 2(a) and (b), clearly illustrate that the selected records have uniform distribution within the range of 200 km and that half of them are higher than 50 gal. Histograms of V_{S30} and V_{S20} values with respect to KiK-net sites and K-NET sites are shown in Fig. 2(c) and (d), respectively. For some K-NET stations with shallow shear-wave velocity profiles <20 m, the value of V_{S20} was computed using an extrapolation

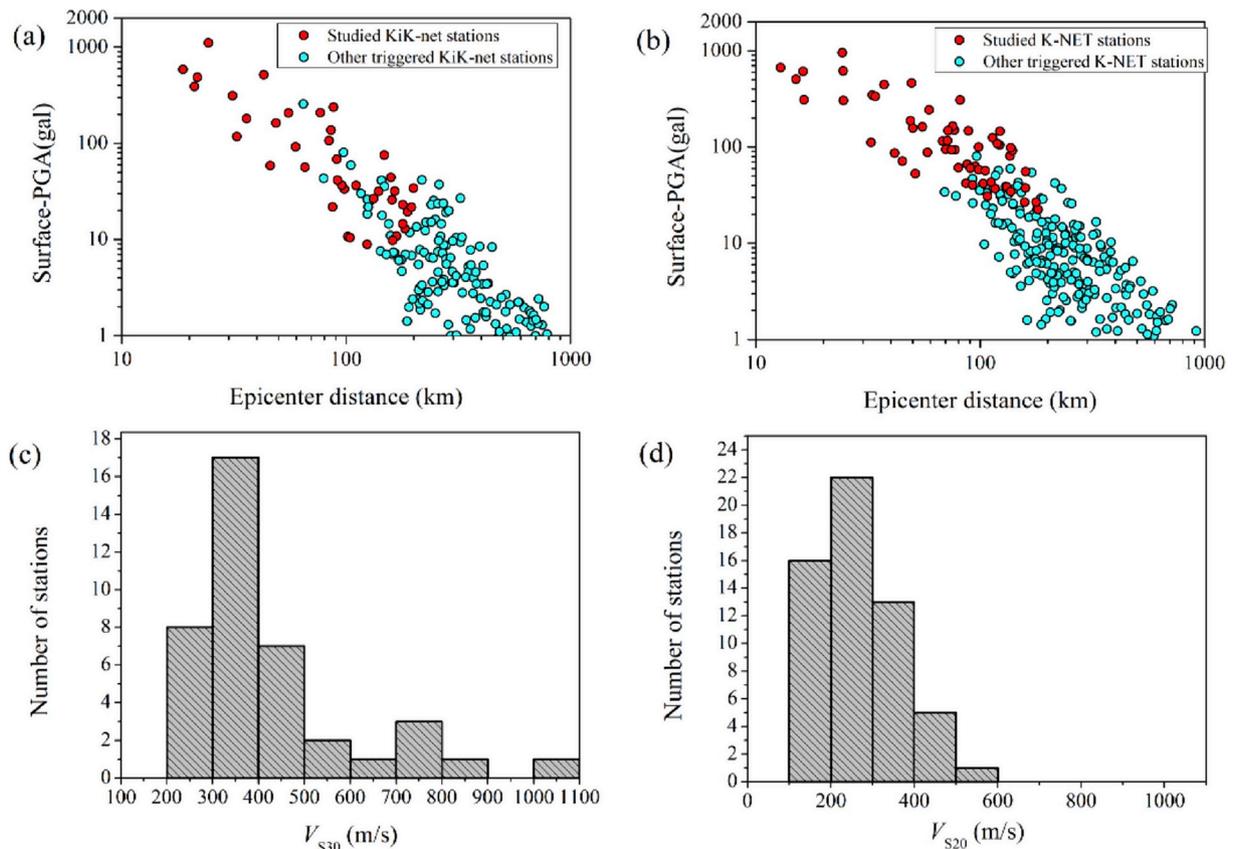


Fig. 2. Surface PGA distribution with epicenter distance for the selected stations: (a) KiK-net and (b) K-NET. Histograms of (c) V_{S30} values of KiK-net stations and (d) V_{S20} values of K-NET stations.

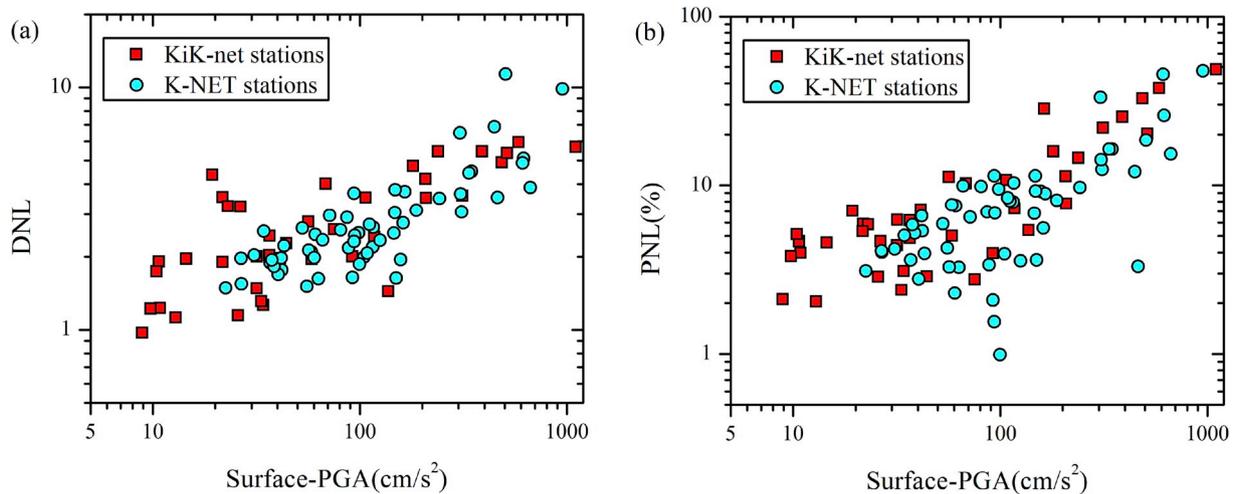


Fig. 3. (a) DNL and (b) PNL values versus observed horizontal PGA at the ground surface based on records at KiK-net and K-NET stations.

method proposed by Wang and Wang [22]. Two-thirds of stations exhibit V_{S30} and V_{S20} values smaller than 500 and 400 m/s, respectively. Potential nonlinear site responses might be observed in this region because of the reasonably strong ground motions recorded and the soft surface soil.

A Butterworth filter with a bandwidth of 0.2–20.0 Hz was applied to each record. The Fourier amplitude spectrum for each component was calculated using the fast Fourier transform method, and a Pazen window with width of 0.5 Hz was used to smooth the spectrum. The spectral ratio of the horizontal component at the ground surface to that of the borehole (surface/borehole) was computed for the KiK-net sites. The H/V spectral ratios at the ground surface were utilized for the K-NET sites because of the lack of borehole measurements. The horizontal amplitude spectra were derived from the square root of the products of the two horizontal components. The linear site response reference curve was calculated as the average of the spectral ratio curves derived from all the selected weak ground motions. Using the S-wave portion of the record need to hand pick the onset of S-waves and determine the S-wave time window which depends on the users' experience and practice. To avoid signal-processing bias, we used the entire duration of the seismic record in the computation of the spectral ratio curves rather than just the S-wave portion like in Ref. [20]. The nonlinearity indicator results were overall similar between using S-wave portion and the entire duration of recordings, which would have limited impact on our following clustering results but significantly improve the efficiency.

3. Quantification of nonlinear site response

3.1. Nonlinear site response indicator parameters

The DNL parameter used in this study was proposed in Ref. [9,14], as shown in Eq. (1):

$$DNL = \sum_{i=N_1}^{N_2} \left| \log \left[\frac{R_{strong}(i)}{R_{weak}(i)} \right] \right| (f_{i+1} - f_i) \quad (1)$$

where R_{strong} is the spectral ratio value for strong ground motion in the mainshock, R_{weak} is the average spectral ratio value computed using the weak aftershock records, f_i is the i th frequency (calculated in the frequency range 0.5–20 Hz), N_1 is the first index of the frequency above 0.5 Hz, and N_2 is the final index of the frequency 20.0 Hz.

To account for the variability of the linear reference site response curve, the PNL indicator was proposed by Ref. [12]:

$$A = \sum_{i=N_1}^{N_2} \begin{cases} (R_{strong}(i) - R_{weak}^+(i)) \log_{10} \left(\frac{f_{i+1}}{f_i} \right) & \text{if } R_{strong}(i) \geq R_{weak}^+(i) \\ (R_{weak}^-(i) - R_{strong}(i)) \log_{10} \left(\frac{f_{i+1}}{f_i} \right) & \text{if } R_{strong}(i) \leq R_{weak}^-(i) \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

$$PNL = 100 \frac{A}{\sum_{i=N_1}^{N_2} |R_{weak}(i)| \log_{10} \left(\frac{f_{i+1}}{f_i} \right)} \% \quad (3)$$

where $R_{weak}^-(i)$ and $R_{weak}^+(i)$ represent the values of the average linear spectral ratio curve minus and plus one standard deviation at the i th frequency f_i , respectively; N_1 and N_2 are the same indices defined for Eq. (1). The value obtained from Eq. (2) is normalized by the linear site response spectral ratio curve to obtain an absolute estimation of the nonlinear soil behavior independent of the linear site response amplitude as shown in Eq. (3).

3.2. DNL and PNL computation results

The DNL and PNL were calculated following the method explained above. Although there is an apparent positive correlation relationship between the DNL, PNL and surface PGA, as illustrated in Fig. 3, the data are clearly scattered because of the variability of the spectral ratio curves. Moreover, site condition might also have an impact on the DNL and PNL values. Sites such as SRCH10 and HKD123 that have relatively high PGA and low DNL values show higher values of V_{S30} or V_{S20} , as indicated in Tables A1 and A2, respectively.

The contour maps of DNL and PNL values, plotted in Fig. 4 using hermit interpolation, indicate the general spatial region of the occurrence of nonlinear site response. The station ID number in Table A1 and A2 were labeled aside the corresponding stations. It is difficult to ascertain the exact region or stations with nonlinear site response solely by the empirical threshold of the DNL or the PNL. The job of classification becomes much more complex if we consider the difference between the K-NET and KiK-net identification results depicted in Fig. 4. If we were to manually compare and check the spectral ratio curves of each station, it would take considerable time and lead to loss of objectivity in the results. Therefore, the use of the K-means clustering algorithm to derive a comprehensive explanation of the observed data is described in the next section.

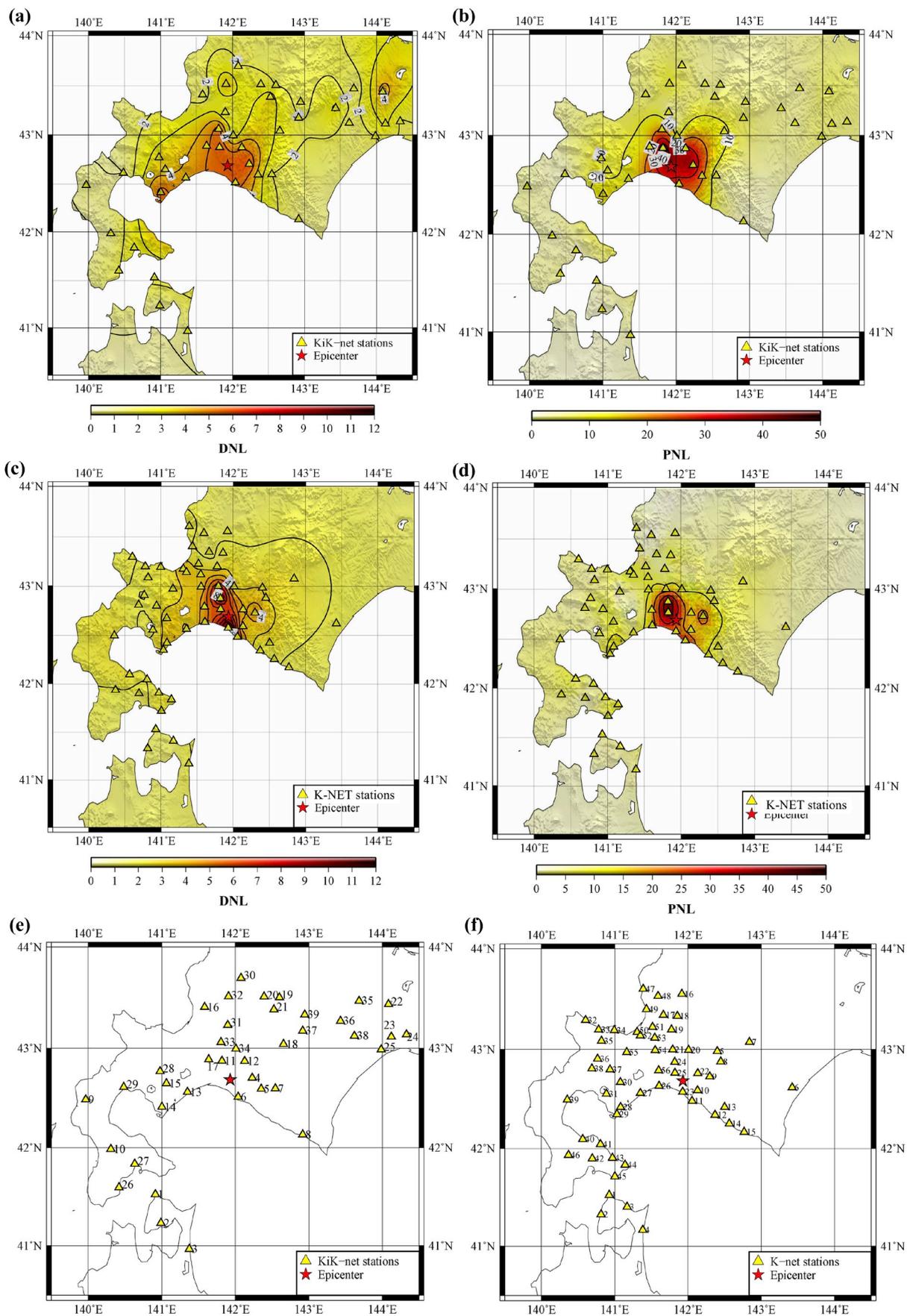


Fig. 4. Contour maps of DNL and PNL values calculated using ground motions recorded at (a) and (b) KiK-net stations and at (c) and (d) K-NET stations during the September 6, 2018 M_w 6.6 Hokkaido Iwate-Tobu earthquake, Japan. The station ID number in Table A1 and A2 were labeled beside KiK-net (e) and K-NET stations (f).

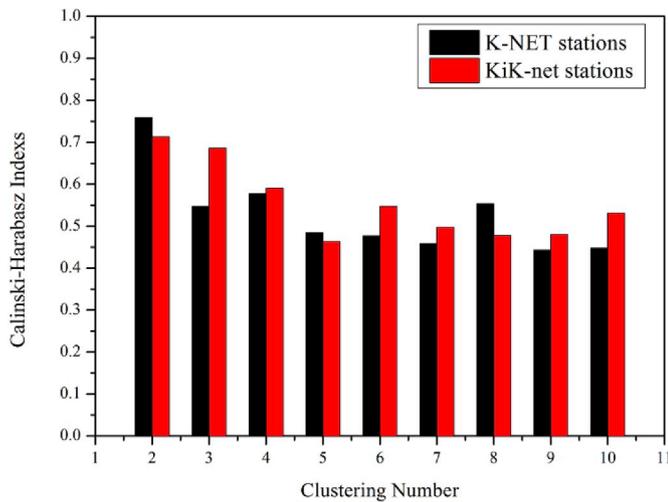


Fig. 5. Clustering number and corresponding Calinski-Harabasz index relating to K-NET and KiK-net stations.

4. Clustering stations with/without nonlinear site response

4.1. Clustering analysis and K-means clustering algorithm

Clustering analysis refers to the process of organizing items into groups based on their degree of similarity. The generated clusters comprise sets of data that are similar to each other in the same group but dissimilar from data in other groups. Clustering analysis categorizes unlabeled data based only on the observations themselves. Thus, it is regarded as an unsupervised classification procedure, which is the most essential characteristic compared with traditional empirical-model-based classification procedures.

K-means clustering is one of the methods used most widely for clustering analysis in data mining [23]. The objective of K-means clustering is to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean value, serving as a prototype of the cluster. Given a set of observations (x_1, x_2, \dots, x_n) ,

where the n -th observation x_n is a d -dimensional real vector with d explanatory variables, K-means clustering partitions the n observations into k sets $S = \{S_1, S_2, \dots, S_k\}$ to minimize the sum of the variance (J) as defined in Eq. (4). The term $\sum_{k=1}^K (|x_n - \mu_k|^p)^{\frac{1}{p}}$ is the Minkowski distance that is used to describe the “difference” between the observation x_n and the k -th clustering centroid point μ_k . When p is equal to 2, it represents the Manhattan distance (or City-block distance) which was used in this study [24].

$$J = \sum_{n=1}^N \sum_{k=1}^K (|x_n - \mu_k|^p)^{\frac{1}{p}} \tag{4}$$

The main steps of the procedure are as follows:

- Step 1: k initial “means” are generated randomly within the data domain.
- Step 2: The corresponding k clusters are created by associating observations with the nearest mean.
- Step 3: The centroid of each of the k clusters μ_k becomes the new mean.
- Step 4: Steps 2 and 3 are repeated until the convergence criterion is reached, i.e., it reaches the specified maximum number of iterations or the centroid of each cluster does not change.

In our problem, each station is treated as one observation with four explanatory variables characterizing nonlinear site response, i.e., ground motion intensity (Surface-PGA or Borehole-PGA), soil condition (V_{S30} or V_{S20}), DNL and PNL. The task of the K-means clustering algorithm is to separate the stations into at least two clusters representing those with nonlinear site response and those with linear site response. As one of the mature clustering analysis tools, the Clustering Toolbox of software MATLAB [25] was used in this study to achieve this objective.

4.2. Multicollinearity diagnosis and standardization of clustering variables

Before employing K-means clustering, multicollinearity diagnosis was performed to guarantee there were no linear correlations between the explanatory variables [26]; otherwise, one variable could be

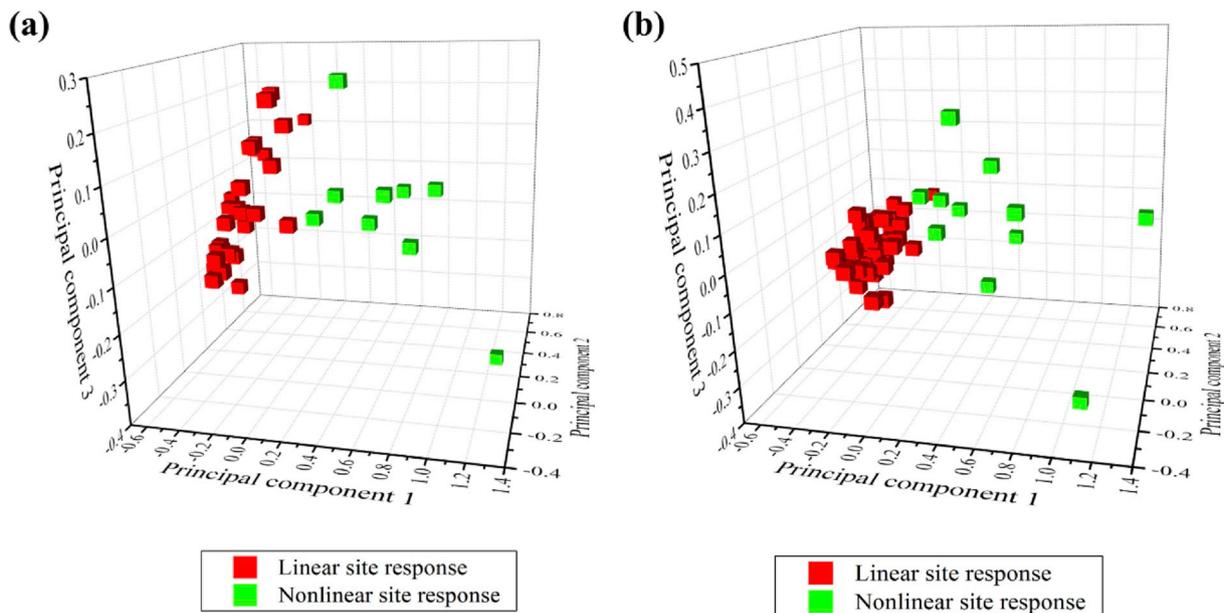


Fig. 6. Observation data for (a) KiK-net and (b) K-NET stations represented in 3-dimensional space using the PCA method. Different colors indicate the nonlinearity identification results using the K-means algorithm. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

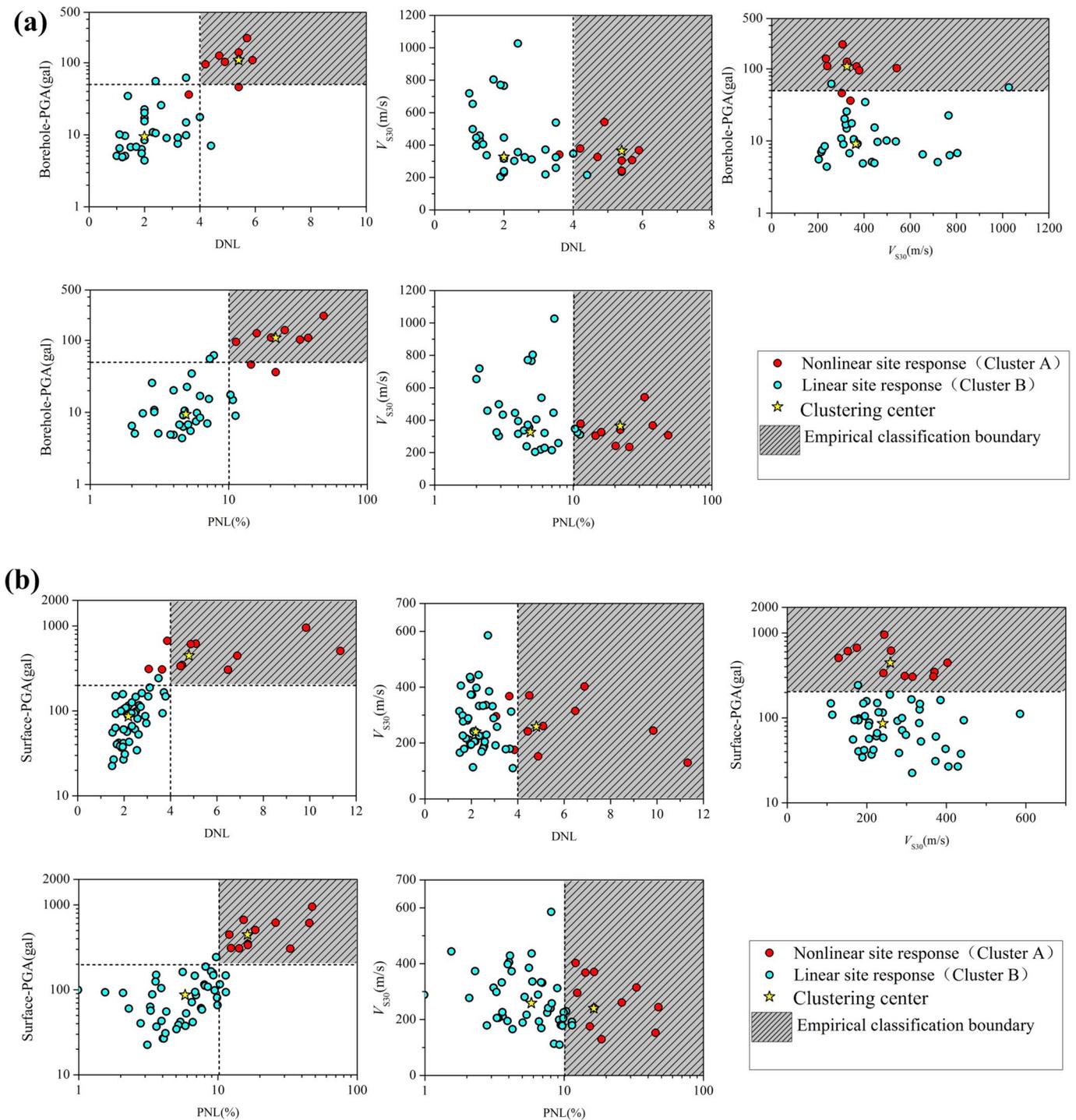


Fig. 7. Values of DNL, PNL for (a) 39 KiK-net stations and (b) 56 K-NET stations versus the recorded PGA and V_{S30}/V_{S20} . Dots of different color indicate the classification results computed using K-means clustering algorithm described in this paper. Dashed lines indicate the empirical classification thresholds of the DNL, PNL, Borehole-PGA and Surface-PGA. Shaded area indicates the region where indicator values are larger than the empirical thresholds to define nonlinear site response. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

expressed linearly by other variables, and the redundant variables might cause unnecessary fluctuation in the clustering results because of a few outliers. The variance inflation factor (VIF) is commonly used to estimate the degree of multicollinearity between different variables [27, 28]. First, a least square linear regression is undertaken in which the i_{th} variable v_i is represented as a function of all the other explanatory variables. The parameter f_i is the regression prediction value:

$$v_i = f_i + \varepsilon = a_1 v_1 + a_2 v_2 + \dots + a_{i-1} v_{i-1} + a_{i+1} v_{i+1} + \dots + b + \varepsilon \quad (5)$$

where b is a constant parameter and ε is the error term.

The VIF_i of v_i is computed as follows:

$$VIF_i = \frac{1}{1 - R_i^2} \quad (6)$$

where R_i is the coefficient of the multiple correlation for Eq. (5), defined as Eq. (7). The term \bar{v} is the mean v_i value of the n observed data $\bar{v} = \frac{1}{n} \sum_{j=1}^n v_{i,j}$:

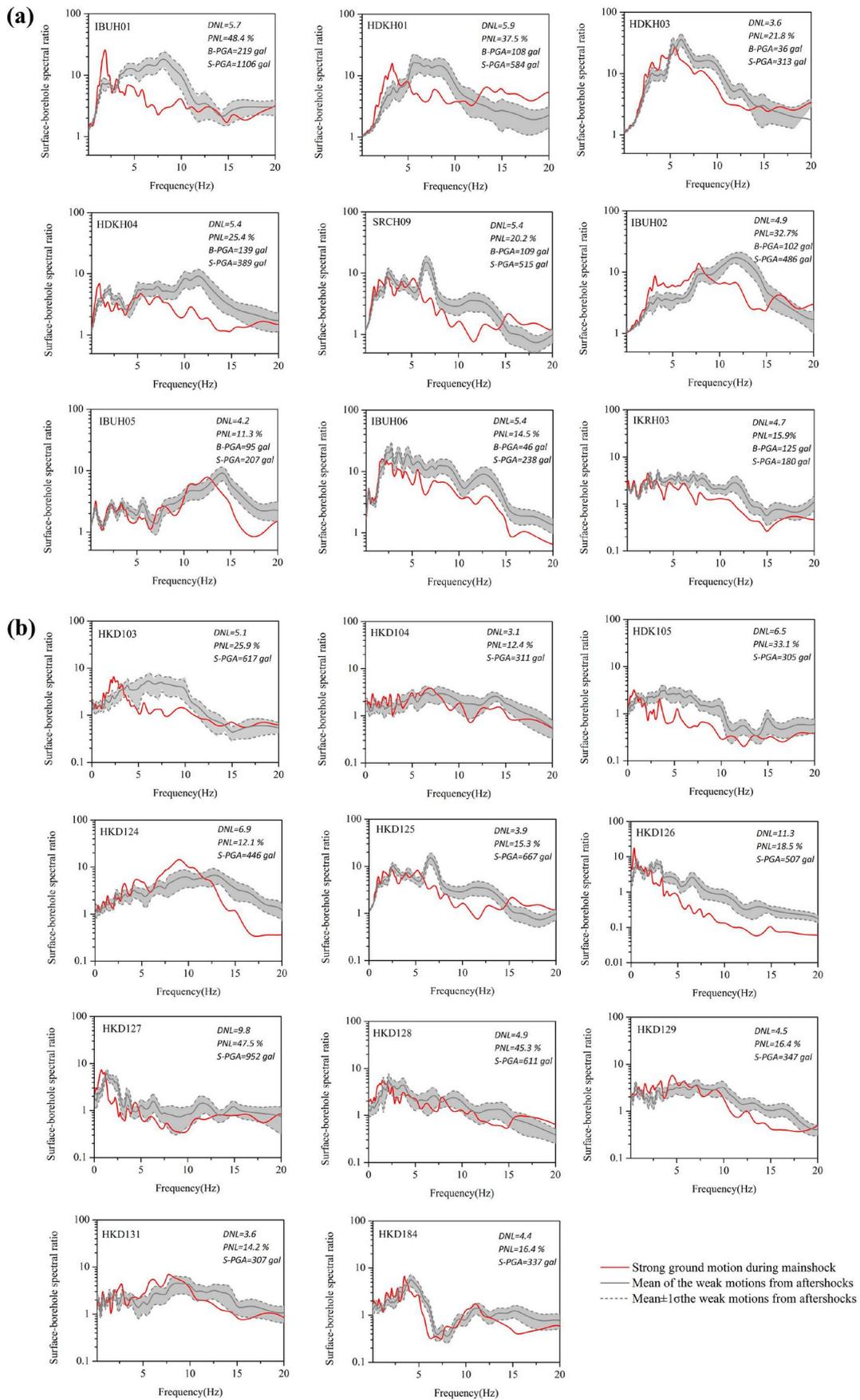


Fig. 8. Surface-borehole spectral ratio curves for (a) 9 KiK-net stations and (b) 11 K-NET stations that were identified as having nonlinear site response based on the K-means clustering algorithm. Shaded area indicates range of the mean plus/minus one standard deviation.

Table 3

Ten different weighed linear combination of the normalized variables.

		M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
KiK-net stations	w1	0.20	0.20	0.20	0.20	0.30	0.30	0.30	0.40	0.40	0.50
	w2	0.20	0.30	0.40	0.50	0.20	0.30	0.40	0.20	0.30	0.20
	w3	0.50	0.40	0.30	0.20	0.40	0.30	0.20	0.30	0.20	0.20
	w4	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
K-NET stations	w1	0.20	0.20	0.20	0.20	0.30	0.30	0.30	0.40	0.40	0.50
	w2	0.20	0.30	0.40	0.50	0.20	0.30	0.40	0.20	0.30	0.20
	w3	0.50	0.40	0.30	0.20	0.40	0.30	0.20	0.30	0.20	0.20
	w4	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10

“0.20”: bold number means that the first several ranking stations are identical with the clustering recognition results in this paper.

$$R_i = 1 - \frac{\sum_{j=1}^n (v_{i,j} - f_{i,j})^2}{\sum_{j=1}^n (v_{i,j} - \bar{v}_i)^2} \quad (7)$$

Significant multicollinearity exists when the VIF value exceeds the threshold value of 10.0 [29]. The VIF values and Pearson’s correlation coefficients were calculated for the four explanatory variables relating to the K-NET and KiK-net observations, as presented in Tables 1 and 2, respectively.

For the K-NET observations (Table .1), the VIF values of the four explanatory variables: Surface-PGA, V_{S20} , DNL and PNL are all <4.0 , indicating negligible multicollinearity and that all four variables could be used in clustering analysis. For the KiK-net stations, the VIF values for Surface-PGA and PNL are all >10.0 . The Pearson’s correlation coefficient between surface-PGA and PNL is 0.95, which indicates strong linear correlation. Therefore, the Surface-PGA was replaced by the Borehole-PGA to represent the level of ground motion and new VIF values were calculated for Borehole-PGA, DNL, PNL and V_{S30} and are listed in Table 2 as case B. The VIF values indicated no predominant multicollinearity (the largest VIF value is 5.2).

In addition to the diagnosis for multicollinearity, scaling of the variables is also an important procedure that should be performed prior to K-means clustering. If variables are measured on different scales, the effect of variables with small scale might be submerged in the variables with larger scale, which might produce misleading results. Significant difference exists in the scale of measurement of the PGA, V_{S30} , V_{S20} , DNL and PNL values. Therefore, we used the extreme method to obtain nondimensional explanatory variables, as shown in Eq. (8), where the term v_i' is the nondimensional result of variable v_i , and v_i^{\max} and v_i^{\min} are the maximum and minimum values, respectively, among the n observations:

$$v_i' = \frac{v_i - v_i^{\min}}{v_i^{\max} - v_i^{\min}} \quad (8)$$

4.3. Optimum clustering number

Before performing the K-means clustering analysis, it is necessary to determine the optimum number of clusters. As the only prior information used in the entire unsupervised clustering analysis process, the number of clusters will influence the clustering result substantially. There is no standard procedure for computation of the optimum number of clusters. The most popular criterion is the F-stopping-rule index of Caliński and Harabasz [30], which is based on the within-cluster sum of squares of differences. It is a measure of the homogeneity level between clusters. The larger the value of the Caliński–Harabasz index, the more significant the differences among the groups and the more acceptable the clustering number. Different clustering numbers and the corresponding Caliński–Harabasz index relating to the K-NET and KiK-net stations are illustrated in Fig. 5. It can be seen that two clusters have the largest value of the Caliński–Harabasz index, indicating that two is the optimum clustering number. The results are consistent with

the problem central to this study, i.e., one cluster groups stations with nonlinear site response and the other cluster groups stations with linear site response. The K-means clustering analysis was then performed to separate the observational data obtained from the KiK-net and K-NET stations into two clusters.

4.4. Clustering result validation

It is difficult to visualize the classification results because each observation element has four explanatory variables that require visualization in four-dimensional space. Therefore, we used principal component analysis (PCA) to reduce the dimensions of the observational data. PCA is mathematically defined as an orthogonal linear transformation that transforms data into a new coordinate system such that the projection of the data comes to lie on the new coordinates (called the principal component) [31]. The greatest variance by projection of the data on the first coordinate is called the first principal component; the second greatest variance on the second coordinate is called the second principal component, and so on. The PCA results indicate that three components account for $>95\%$ of the variance, which could represent the original observational data studied here. The original four-dimensional observation data for the KiK-net and K-NET stations were projected into three-dimensional space, as shown in Fig. 6. The classification results using the K-means algorithm are all presented visually as two separate clusters, indicating a reasonably satisfactory classification result.

The clustering results were then compared with those based on empirical thresholds of widely used nonlinear indicators, e.g., the DNL, PNL and PGA (Fig. 7). Noguchi and Sasatani [14] suggested a DNL value of 4.0 for the H/V ratio method as the boundary of nonlinear site response identification, and this has been used in many subsequent studies [13]. Although there is a difference between the surface/borehole and H/V spectral ratios, the DNL_{SB} and DNL_{HV} thresholds are nearly equivalent [14]. The empirical threshold values of the PNL and Borehole-PGA suggested by Régnier et al. [12] were 10% and 50 gal, respectively, based on a large number of KiK-net data obtained in Japan. However, debate continues regarding where within the range of 100–200 gal the surface-PGA threshold value should be set. Our classification results were found in good agreement with the threshold value of 200 gal suggested by Ren et al. [13], as shown in Fig. 7(b). The correlation between site condition (i.e., V_{S30} or V_{S20}) and the degree of nonlinear site response (i.e., DNL or PNL) is obviously weak, and there is no convincing threshold for either V_{S30} or V_{S20} . It can be observed that the sites with nonlinear site response classified by clustering algorithm (i.e., cluster A in Fig. 7) are located mostly within the empirical nonlinearity region, except for two or three data points near the boundary. It proves that clustering results are reasonably convincing and robust from the perspective of the distribution of nonlinearity indicators.

The surface-borehole or H/V spectral ratio curves for stations with nonlinear site response, according to our clustering results, are shown in Fig. 8. The computed spectral ratios using the recordings of the main-shock, which were amplified at frequencies below the predominant

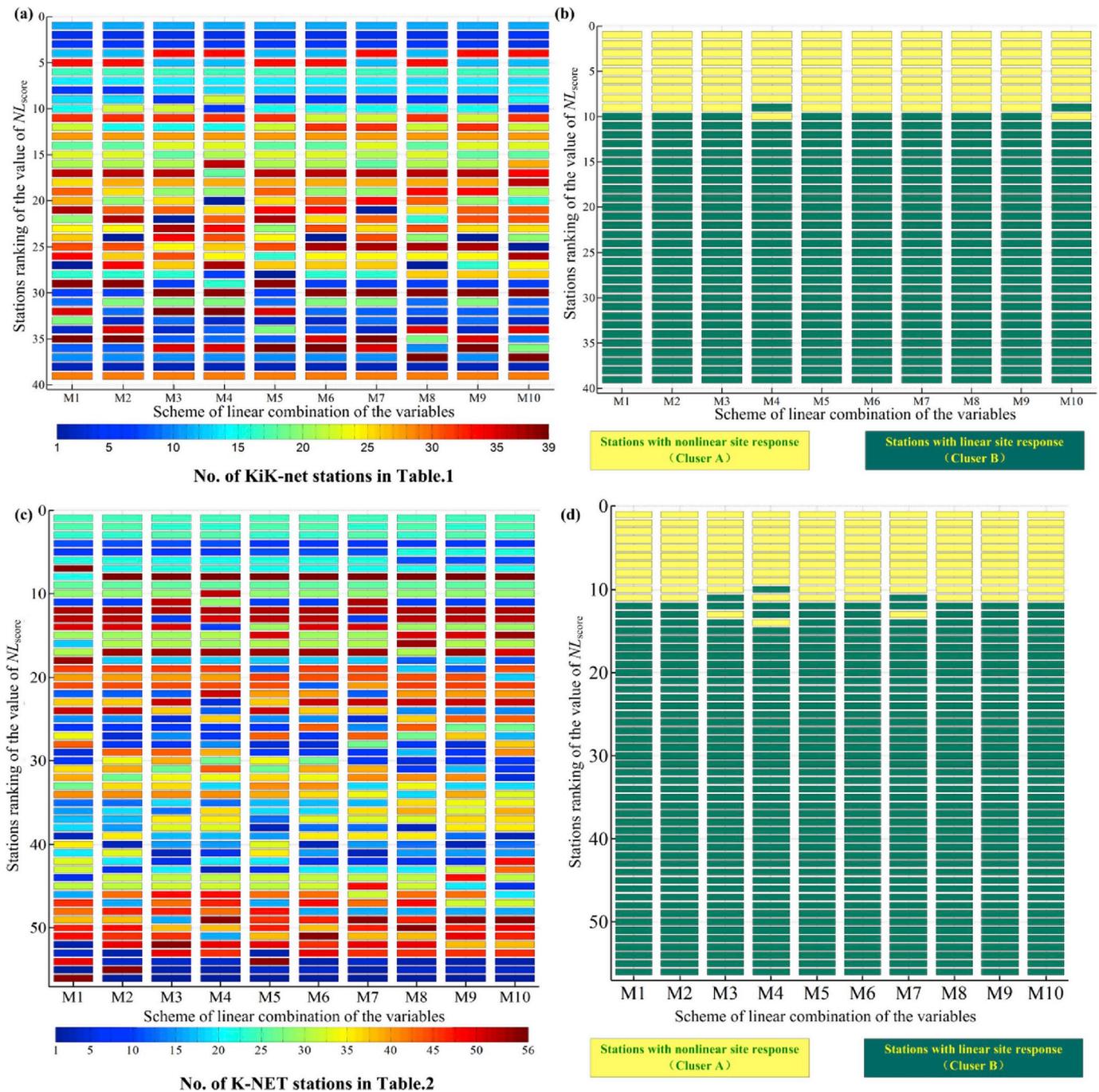


Fig. 9. Ranking of NL_{score} indicating the possibility of occurrence of nonlinear site response for (a) KiK-net stations and (c) K-NET stations for 10 schemes of different linear combinations of the normalized variables. Each cell represents one station with different colors representing the stations listed in Tables A1 and A2. The ranking results are compared with the clustering results obtained in this study for: (b) KiK-net stations and (d) K-NET stations. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

frequency and were reduced above it, illustrate typical nonlinear soil behavior. For KiK-net stations, the DNL values were between 3.6 (at HDKH03) and 5.9 (at HDKH01). All the DNL values were larger than the empirical threshold of 4.0 except for the station HDKH03, of which the PNL value is 21.8% and reduced predominant frequency could be clearly observed. The PNL values are between 11.3% and 48.4%. The predominant frequencies of the sites were reduced by approximately 9% (at SRCH09) to 78% (at IBUH01). It can be seen that a systematic decrease of the peak frequencies is associated with a decrease of their amplitude. For K-NET stations, the DNL values were between 3.6 (at HKD131) and 11.3 (at HKD126). For HKD131, PNL is 14.2% and the predominant

frequency reduced from 10.0 Hz to 7.69 Hz. The PNL values are between 12.1% and 47.5%. The predominant frequencies of the K-NET sites were reduced by approximately 17% (at HKD184) to 83% (at HKD105). It indicates that the results of the identification of nonlinear site response using the clustering method are reasonable and accurate. The entire clustering identification process, which is completely automatic and efficient without any manual intervention, is demonstrated to produce satisfactory and objective results, as shown in Figs. 7 and 8.

It is worth noting that accurate identification of predominant frequency of spectral ratio curves could not be easily accomplished by programs especially when multiple peaks existed. In addition, the

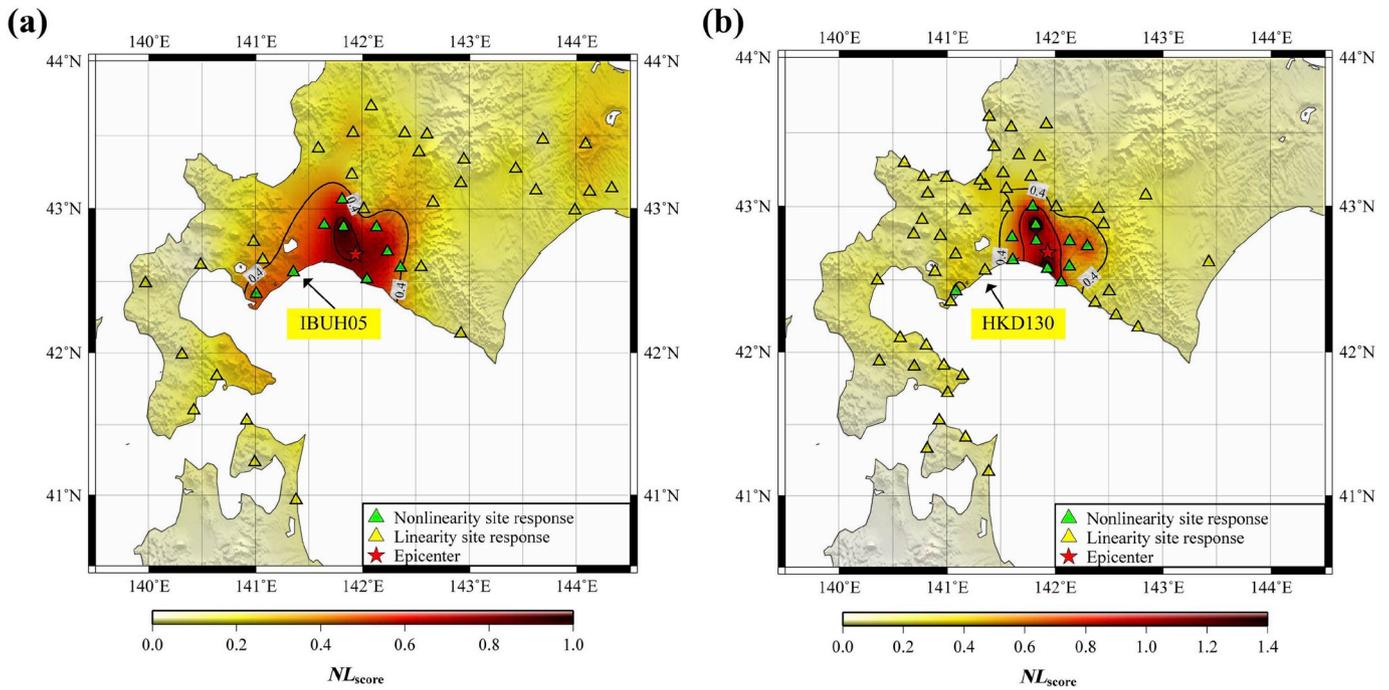


Fig. 10. The contour maps of the NL_{score} for studied (a) KiK-net stations and (b) K-NET stations. The stations with nonlinear site response are indicated with green triangles. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

predominant frequency shift is part of change in spectral ratio curve shape, which could be described by the DNL and PNL. Larger DNL and PNL values mean generally larger reduction of amplitude and larger shift of frequencies. The purpose we apply the clustering method is to minimize the impact of subjective judgment and improve classification efficiency. Therefore it is suggested that the frequency shift indicators (like FNL [12], Rfp [13]) are not first choices of clustering variables, but could be applied in checking or further analysis of the classification results.

5. Ranking of nonlinear-site-response occurrence probability

To illustrate the nonlinear site response occurrence probability intuitively, the PGA, DNL, PNL and site condition parameters were linearly combined with different weights to calculate a comprehensive ranking score. As shown in Fig. 7, the PGA, DNL and PNL values have strong positive correlation, whereas site condition has weak negative correlation with them. Therefore, it is mandatory that the corresponding weights for V_{S30} or V_{S20} be smaller than for the other nonlinearity indicator variables. Equation (9) defines the ranking score of nonlinear site response (NL_{score}), which reflects the possibility of the occurrence of nonlinear site response. The larger the value of NL_{score} the more likely the occurrence of nonlinear site response. The variables were normalized using the extreme method, as in Eq. (8). Ten possible schemes of linear combinations of weights that satisfy Eq. (9) are presented in Table 3.

$$\begin{cases} NL_{score-KiKnet} = w_1 BoreholePGA + w_2 DNL + w_3 PNL - w_4 V_{S30} \\ NL_{score-Knet} = w_1 SurfacePGA + w_2 DNL + w_3 PNL - w_4 V_{S20} \\ w_1 + w_2 + w_3 + w_4 = 1; \\ w_1, w_2, w_3, w_4 > 0; \\ w_1, w_2, w_3 > w_4 \end{cases} \quad (9)$$

The KiK-net and K-NET stations were ranked based on their NL_{score} value based on different weight combinations of the variables, as shown in Fig. 9. The station with largest NL_{score} value was ranked as first and the station with the smallest value was ranked last. For illustration purposes, we used cells with different colors to represent the stations listed in Table A1 and A2. For the KiK-net stations, it can be seen that the

top nine stations are identical with the results of Cluster A (nonlinear site response), except for the M4 and M10 schemes (Fig. 9). The ranking difference among the other stations is also reasonably close. The same tendency can be observed for the K-NET stations. The consistency between the ranking and clustering recognition results indicates that the clustering identification results are robust and reliable. Considering the small difference within the ranking results among the different weighting schemes, we used the M6 combination scheme to compute NL_{score} , i.e., w_1, w_2, w_3 and w_4 is equal to 0.3, 0.3, 0.3 and 0.1, respectively. The NL_{score} computation results of the stations are listed in Tables A1 and A2.

The corresponding contour maps of the NL_{score} for KiK-net and K-net stations were compared in Fig. 10. Based on the result of NL_{score} (Fig. 9) and cluster analysis, the stations with NL_{score} value larger than 0.4 in Hokkaido earthquake would be classified as nonlinearity ones for both KiK-net and K-NET stations. The identified site nonlinearity region shape (indicated with contour line of $NL_{score} = 0.4$) is basically consistent regarding the contour map of KiK-net and K-NET stations. The stations for both arrays with identified nonlinearity site responses is observed located in the same region except for HKD130 and IBUH05. The DNL value of IBUH05 and HKD130 is 4.2 and 2.8 respectively, which is near the empirical nonlinearity boundary. The IBUH05 show slight site nonlinearity as shown in Fig. 8 while HKD130 is not classified into nonlinearity according to the observation of the spectral ratio curves. It is indicated that although the clustering algorithm could efficiently classify the nonlinearity response of stations in the first step, the stations located near the classification boundary still should be of special concern for further analysis. The results given by clustering algorithm are reference rather than the final conclusion of the nonlinearity site response which is complex and influenced by many factors.

6. Conclusions

This study was the first to incorporate the clustering algorithm of the machine learning technique to address the problem of classification of nonlinear seismic site response. We used the data from the 2018 $M_w 6.6$ Hokkaido Iburu-Tobu earthquake as a case study.

We calculated the surface/borehole and horizontal/vertical (H/V) spectral ratios using strong ground motion data recorded at KiK-net vertical array and K-NET stations. The degree of nonlinear site response (DNL) and the percentage of nonlinear site response (PNL) were then computed respectively based on the difference between the strong motion in the mainshock and ground motion of the weak aftershocks as their linear site response reference. Using the values of DNL, PNL, PGA and V_{S30} (V_{S20}) as clustering variables, the K-means clustering algorithm was incorporated in the nonlinear site response identification process. Multicollinearity diagnosis was applied to guarantee the explanatory variables were not correlated linearly. After confirmation of the optimum clustering number, we grouped the stations into two clusters representing observations with nonlinear site response and those with linear site response.

To validate the clustering identification results, we adopted different methods to evaluate them from different aspects. First, principal component analysis (PCA) was performed to illustrate the clustering results intuitively by reducing the number of dimensions of the data. Following which the values of the DNL, PNL, PGA and V_{S30} (V_{S20}) were presented visually in two separate clusters. The clustering classification results for the KiK-net and K-NET stations were found consistent with results indicated based on empirical nonlinearity thresholds proposed in other studies. For the stations identified with nonlinear site response, an obvious reduction of amplification and shift of frequency could be observed between the mainshock and the reference linear spectral ratio curves, which is the most common characteristics of nonlinear site response. Finally, the NL_{score} of each station, indicating the occurrence probability of nonlinear site response, was calculated using all possible linear weighted combinations of the four normalized explanatory variables. The top ranking stations with high nonlinearity occurrence probability were found consistent with the clustering nonlinearity recognition results. All validation work proved that the unsupervised clustering method proposed in this study was efficient in obtaining convincing and robust results.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.soildyn.2019.105907>.

APPENDIX

Table A1

Information of the selected 39 KiK-net strong-motion stations and the values of nonlinear site response indicator (DNL and PNL).

No.	Station Code	Epicenter Dis. (km)	PGA(gal)		V_{S30} (m/s)	DNL	PNL	Nonlinearity identification result	NL_{score}
			Borehole	Surface					
1	AOMH01	158	10.8	44.2	302	2.3	2.9	L	0.18
2	AOMH03	183	6.5	12.9	654	1.1	2.0	L	0.05
3	AOMH06	200	5.1	34.3	434	1.3	3.1	L	0.10
4	HDKH01	19	108.4	584.1	368	5.9	37.5	NL	0.76
5	HDKH03	31	36.1	312.8	341	3.6	21.8	NL	0.41
6	HDKH04	21	138.5	389.1	235	5.4	25.4	NL	0.70
7	HDKH05	46	22.5	58.6	766	2.0	5.0	L	0.14
8	HDKH07	98	9.7	33.4	459	1.3	2.4	L	0.10
9	HVMH01	168	4.9	10.8	395	1.2	4.0	L	0.10
10	HVMH02	160	10.1	25.8	498	1.1	2.9	L	0.08
11	IBUH01	24	218.7	1105.5	307	5.7	48.4	NL	0.98
12	IBUH02	22	102.2	485.8	542	4.9	32.7	NL	0.63
13	IBUH05	55	95.0	207.4	379	4.2	11.3	NL	0.46
14	IBUH06	88	46.0	238.4	304	5.4	14.5	NL	0.50
15	IBUH07	77	61.8	208.9	259	3.5	7.8	L	0.36
16	IKRH01	86	34.5	137.3	405	1.4	5.4	L	0.16
17	IKRH03	36	125.0	180.6	326	4.7	15.9	NL	0.57
18	KKWH08	66	9.0	56.5	311	2.8	11.2	L	0.26
19	KKWH12	102	6.3	10.7	771	1.9	4.7	L	0.11

(continued on next page)

Table A1 (continued)

No.	Station Code	Epicenter Dis. (km)	PGA(gal)		V _{S30} (m/s)	DNL	PNL	Nonlinearity identification result	NL _{score}
			Borehole	Surface					
20	KKWH13	96	10.5	36.7	356	2.4	4.8	L	0.19
21	KKWH14	87	9.9	21.8	538	3.5	5.9	L	0.25
22	KSRH01	188	7.0	19.4	215	4.4	7.0	L	0.34
23	KSRH02	179	7.5	23.1	219	3.2	5.8	L	0.26
24	KSRH07	196	5.5	21.7	204	1.9	5.3	L	0.18
25	KSRH09	165	8.4	31.9	230	2.0	6.2	L	0.19
26	OSMH01	179	4.4	14.5	239	2.0	4.6	L	0.17
27	OSMH02	148	25.6	75.3	325	2.6	2.8	L	0.22
28	SBSH08	84	14.8	106.9	325	3.5	10.7	L	0.31
29	SBSH09	124	5.1	8.9	719	1.0	2.1	L	0.04
30	SRCH06	111	16.9	36.6	321	2.0	6.2	L	0.19
31	SRCH07	60	20.2	92.0	316	2.0	4.0	L	0.18
32	SRCH08	91	17.5	68.4	347	4.0	10.3	L	0.34
33	SRCH09	43	109.4	515.2	241	5.4	20.2	NL	0.63
34	SRCH10	33	55.3	117.6	1027	2.4	7.3	L	0.19
35	TKCH01	161	4.9	9.8	445	1.2	3.8	L	0.10
36	TKCH03	133	9.1	26.5	372	3.2	4.7	L	0.24
37	TKCH04	91	15.4	41.4	446	2.0	7.2	L	0.18
38	TKCH05	140	6.7	31.8	337	1.5	4.4	L	0.13
39	TKCH10	104	6.8	10.4	804	1.7	5.1	L	0.09

“NL” means that the stations were identified with nonlinear site response.

“L” means that the stations were identified with linear site response.

Table A2

Same as Table A1 but for 56 K-NET strong-motion stations.

No.	Station Code	Epicenter Dis. (km)	Surface PGA (gal)	V _{S20} (m/s)	DNL	PNL	Nonlinearity identification result	NL _{score}
1	AOM001	158	26.8	405	1.5	4.0	L	0.06
2	AOM002	181	22.5	314	1.5	3.1	L	0.07
3	AOM003	159	55.5	165	1.5	4.3	L	0.12
4	AOM007	178	26.7	429	2.0	4.1	L	0.10
5	HKD040	45	71.6	289	3.0	6.5	L	0.26
6	HKD097	117	37.0	212	1.9	3.6	L	0.14
7	HKD099	80	60.9	224	2.5	7.6	L	0.22
8	HKD102	42	86.7	332	2.9	7.0	L	0.25
9	HKD103	24	617.7	261	5.1	25.9	NL	0.80
10	HKD104	16	310.5	295	3.1	12.4	NL	0.39
11	HKD105	25	305.0	315	6.5	33.1	NL	0.82
12	HKD106	50	157.3	199	1.9	9.2	L	0.22
13	HKD107	51	52.8	336	2.6	5.9	L	0.20
14	HKD108	68	115.4	240	2.2	8.0	L	0.21
15	HKD109	86	42.0	217	1.8	5.4	L	0.13
16	HKD118	95	63.1	298	1.6	3.3	L	0.10
17	HKD120	77	149.9	226	1.6	3.6	L	0.16
18	HKD121	71	116.4	230	2.6	10.3	L	0.27
19	HKD122	58	88.3	206	2.2	3.4	L	0.19
20	HKD123	33	111.6	585	2.7	8.1	L	0.19
21	HKD124	37	446.1	403	6.9	12.1	NL	0.82
22	HKD125	13	666.9	175	3.9	15.3	NL	0.67
23	HKD126	15	506.9	129	11.3	18.5	NL	1.38
24	HKD127	24	952.5	244	9.8	47.5	NL	1.52
25	HKD128	16	611.4	152	4.9	45.3	NL	0.88
26	HKD129	33	347.0	370	4.5	16.4	NL	0.56
27	HKD130	55	161.8	385	2.8	5.6	L	0.25
28	HKD131	81	307.3	368	3.6	14.2	NL	0.44
29	HKD132	88	147.9	192	3.0	11.4	L	0.34
30	HKD134	75	165.0	312	3.7	8.9	L	0.38
31	HKD135	93	40.4	179	1.7	2.8	L	0.12
32	HKD138	131	38.6	281	1.8	5.2	L	0.12
33	HKD139	114	125.3	333	2.3	3.6	L	0.19
34	HKD140	98	58.4	241	2.1	7.7	L	0.18
35	HKD142	105	56.8	206	2.1	3.3	L	0.17
36	HKD144	103	41.7	193	2.0	6.6	L	0.16
37	HKD145	88	66.1	226	2.3	9.9	L	0.22
38	HKD146	108	30.9	373	2.0	4.2	L	0.12
39	HKD151	137	34.4	189	2.6	5.0	L	0.21
40	HKD154	136	80.9	204	2.6	9.8	L	0.25
41	HKD155	122	104.7	195	2.0	3.9	L	0.18
42	HKD156	139	92.1	277	1.6	2.1	L	0.12
43	HKD157	122	146.3	334	2.5	6.8	L	0.23
44	HKD158	119	108.7	113	2.1	8.4	L	0.23
45	HKD159	137	98.5	180	2.5	9.5	L	0.26

(continued on next page)

Table A2 (continued)

No.	Station Code	Epicenter Dis. (km)	Surface PGA (gal)	V_{S20} (m/s)	DNL	PNL	Nonlinearity identification result	NL_{score}
46	HKD169	159	37.7	437	1.9	5.8	L	0.10
47	HKD175	112	43.1	398	2.2	3.9	L	0.14
48	HKD176	98	99.8	289	1.9	1.0	L	0.13
49	HKD177	90	60.3	374	2.0	2.3	L	0.12
50	HKD178	77	93.7	179	3.7	11.4	L	0.38
51	HKD179	70	94.4	169	2.4	6.8	L	0.24
52	HKD180	72	148.1	110	3.8	9.3	L	0.42
53	HKD181	59	242.8	178	3.5	9.7	L	0.42
54	HKD182	49	187.6	258	3.1	8.1	L	0.34
55	HKD183	74	93.7	444	2.3	1.5	L	0.15
56	HKD184	34	336.5	242	4.4	16.4	NL	0.57

"NL" means that the stations were identified with nonlinear site response.

"L" means that the stations were identified with linear site response.

References

- Trifunac MD, Todorovska MI. Nonlinear soil response—1994 Northridge, California, earthquake. *J Geotech Eng* 1996;122(9):725–35.
- Beresnev IA, Field EH, Van Den Abeele K, Johnson PA. Magnitude of nonlinear sediment response in Los Angeles basin during the 1994 Northridge, California, earthquake. *Bull Seismol Soc Am* 1998;88(4):1079–84.
- Hartzell S. Variability in nonlinear sediment response during the 1994 Northridge, California, earthquake. *Bull Seismol Soc Am* 1998;88(6):1426–37.
- Pavlenko O, Loh CH. Nonlinear identification of the soil response at Dahan downhole array site during the 1999 Chi–Chi earthquake. *Soil Dyn Earthq Eng* 2005;25(3):241–50.
- Noguchi S, Sasatani T. Nonlinear soil response and its effects on strong ground motions during the 2003 Miyagi-Oki intraslab earthquake. *Zisin* 2011;63:165–87 (in Japanese with English abstract).
- Rong M, Wang Z, Woolery EW, Lyu Y, Li X, Li S. Nonlinear site response from the strong ground-motion recordings in Western China. *Soil Dyn Earthq Eng* 2016;82:99–110.
- Midorikawa S. Nonlinearity of site amplification during strong ground shaking. *Zisin* 1993;2(46):207–16.
- Wen KL. Non-linear soil response in ground motions. *Earthq Eng Struct Dyn* 1994;23(6):599–608.
- Wen KL, Chang TM, Lin CM, Chiang HJ. Identification of nonlinear site response during the 1999, Chi-Chi, Taiwan earthquake from the H/V spectral ratio. In: Third international symposium on the effects of surface geology on seismic motion, Grenoble, France, 30 Aug.-1 Sep., 2006; 2006. p. 225–32.
- Wen KL, Chang TM, Lin CM, Chiang HJ. Identification of nonlinear site response using the H/V spectral ratio method. *Terr Atmos Ocean Sci* 2006;17(3):533–46.
- Wen KL, Huang JY, Chen CT, Cheng YW. Nonlinear site response of the 2010 Darfield, New Zealand earthquake sequence. In: 4th iaspei/iaee international symposium: effects of surface geology on seismic motion; 2011.
- Régnier J, Cadet H, Bonilla LF, Bertrand E, Semblat JF. Assessing nonlinear behavior of soils in seismic site response: statistical analysis on KiK-net strong-motion data. *Bull Seismol Soc Am* 2013;103(3):1750–70.
- Ren Y, Wen R, Yao X, Ji K. Five parameters for the evaluation of the soil nonlinearity during the Ms8.0 Wenchuan Earthquake using the HVSR method. *Earth Planets Space* 2017;69(1):116.
- Noguchi S, Sasatani T. Quantification of degree of nonlinear site response. In: 14th world conference on earthquake engineering, Beijing; 2008. paper ID (pp. 03–03-0049).
- Bailey K. Numerical taxonomy and cluster analysis. In: *Typologies and taxonomies: an introduction to classification techniques*, 34; 1994. p. 6524.
- Chen Y. Automatic microseismic event picking via unsupervised machine learning. *Geophys J Int* 2018;212(1):88–102.
- Riahi N, Gerstoft P. Using graph clustering to locate sources within a dense sensor array. *Signal Process* 2017;132:110–20.
- Miettinen K. *Nonlinear multiobjective optimization*, 12. Springer Science & Business Media; 2012.
- Yamagishi H, Yamazaki F. Landslides by the 2018 Hokkaido iburi-tobu earthquake on september 6. *Landslides* 2018;15(12):2521–4.
- Dhakal YP, Kunugi T, Kimura T, Suzuki W, Aoi S. Peak ground motions and characteristics of nonlinear site response during the 2018 Mw 6.6 Hokkaido eastern Iburi earthquake. *Earth Planets Space* 2019;71(1):56.
- Wang HY, Wang SY. A new method for estimating vs (30) from a shallow shear-wave velocity profile (depth < 30 m). *Bull Seismol Soc Am* 2015;105(3):1359–70.
- Arthur D, Vassilvitskii S. k-means++: the advantages of careful seeding. In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics; 2007. p. 1027–35.
- Kriegel HP, Schubert E, Zimek A. The (black) art of runtime evaluation: are we comparing algorithms or implementations? *Knowl Inf Syst* 2017;52(2):341–78.
- Mathworks Inc. *Clustering Toolbox user's guide*. 2014.
- Spicer J. *Making sense of multivariate data analysis: an intuitive approach*. Sage; 2005.
- Montgomery DC, Peck EA, Vining GG. *Introduction to linear regression analysis*, 821. John Wiley & Sons; 2012.
- James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning*, 112. New York: Springer; 2013.
- Kutner MH, Nachtsheim CJ, Neter J. *Applied linear regression models*. fourth ed. McGraw-Hill Irwin; 2004.
- Calinski T, Harabasz J. A dendrite method for cluster analysis. *Commun Stat Theor Methods* 1974;3(1):1–27.
- Jolliffe IT. *Principal component analysis, series: springer series in statistics*. second ed., XXIX. NY: Springer; 2001, ISBN 978-0-387-95442-4. p. 487. 28 illus.