17th World Conference on Earthquake Engineering, 17WCEE
Sendai, Japan - September 13th to 18th 2020

Footprint of K-means clustering algorithm used in identification of nonlinear seismic site response

Ji Kun[(1)]; Ren Yefei[(2)]; Wen Ruizhi[(3)];Yadab P. Dhakal[(4)];

[(1)] *Assistant Researcher, Key Laboratory of Earthquake Engineering and Engineering Vibration. Institute of Engineering Mechanics, China Earthquake Administration, jikun@iem.ac.cn*
[(2)] *Associate Professor, Key Laboratory of Earthquake Engineering and Engineering Vibration.. Institute of Engineering Mechanics, China Earthquake Administration, renyefei@iem.net.cn*
[(3)] *Professor, Key Laboratory of Earthquake Engineering and Engineering Vibration. Institute of Engineering Mechanics, China Earthquake Administration, ruizhi@iem.net.cn*
[(3)] *Associate Researcher, National Research Institute for Earth Science and Disaster Resilience, Japan, renyefei@iem.net.cn*

Abstract

The spectral ratio methods have been widely used in the evaluation of nonlinear seismic site response. Nevertheless, it is still inefficient and subjective to identify the stations with nonlinear site response according to empirical threshold values of spectral ratio nonlinear degree indicators. The clustering algorithm in machine learning was firstly applied in this paper to address this problem using the September 6th, 2018 Japan Hokkaido eastern Iburi earthquake as an example. Firstly we calculated the surface/borehole spectral ratios using strong ground motion data recorded at KiK-net vertical array. The degree of nonlinear site response (DNL) and percentage of nonlinear site response (PNL) were computed based on the difference between strong motion in mainshock and weak aftershocks as their linear site response reference. Then the K-means clustering algorithm was introduced and incorporated in the nonlinear site response identification using the DNL, PNL, ground motion strength (PGA) and site condition (V_{S30}) as explanatory variables. After careful multicollinearity diagnosis and confirmation of the optimum number of clustering, we successfully classified stations into two clusters, representing stations with nonlinear and linear site response respectively. The clustering results are overall in good agreement with the classification results indicated by empirical thresholds of several nonlinear indicators. The deamplification and shift of frequency could also be visually observed from the spectral ratio curves regarding ground motions in mainshock and aftershocks, which show the typical nonlinear site response characteristics. This work provides an enlightening example of using state-of-art machine learning technique to solve the traditional earthquake engineering problems.

Keywords: Nonlinear site response; spectral ratio; K-means clustering algorithm; Hokkaido eastern Iburi earthquake;

1. Introduction

It is widely recognized that the seismic response characteristics of surface soft soil become nonlinear when it is struck by strong motion. More and more evidence in subsequent earthquakes showed that the soil nonlinear behavior could be indicated and evaluated by comparative spectral ratio curves computed from weak and strong motions[1][2]. When the ground motion input exceeds a certain threshold, the shift of the resonant frequencies toward lower values and a reduction in the associated amplification would be observed in the spectral ratio curves. To quantify the difference between linear and nonlinear site responses, some indicators had been proposed to evaluate the degree of nonlinearity. The most widely used parameters are the degree of nonlinearity of site response (DNL) and the percentage of nonlinearity (PNL), which were proposed in Ref.[3] and [4] respectively.

Although both parameters have been commonly used in practice, it is not an easy job to definitely classify the sites with and without soil nonlinearity only rely on their values. That's partly because the empirical thresholds of the DNL or PNL are relatively subjective and changed with specific earthquake events and site conditions[5][6]. In addition, the identification results based on the fixed threshold of DNL and PNL are not always the same considering the variability of spectral ratio curves. Besides that, the ground motion input level and the site condition both need to be comprehensively considered by users to get

plausible classification results. It is well recognized that the larger the ground motion input level is and the softer the soil condition is, the more likely site response will behave nonlinearity. However, the corresponding classification thresholds for both the ground motion intensity and soil condition are also unfixed especially for the latter one. In practice, the judgment of nonlinear site response depends on the existing knowledge and past experience of researchers. The manual classification process will cost a lot of time especially when the number of records is large.

In view of these defects, the clustering algorithm of state-of-art machine learning technique is firstly incorporated in the seismic nonlinear site response identification in our study. Clustering analysis is one of the main tools of exploratory data mining. It is actually not one specific algorithm, but an unsupervised machine learning process that classifies unlabeled similar objects into the same group or cluster[7]. It is usually served as a useful tool for solving the multi-objective optimization (MP) problems[8], which is concerned with more than one objective function to be optimized simultaneously. The stations with nonlinear site response have similar recorded large ground motion, similar site condition covered by soft soil and similar feature in the spectral ratio curve that could be measured by DNL or PNL. These inherent similarities provide solid foundations for the use of clustering analysis to classify stations with and without nonlinear site response, which can be regarded as a typical MP problem that needs to consider all these factors as variables in objective functions to get a comprehensive result.

In this study, the recent September 6th, 2018 Hokkaido eastern Iburi earthquake (M_w6.6) was taken as an example to analyze the nonlinear seismic site response. The surface/borehole spectral ratio were calculated using strong ground motion data recorded at KiK-net vertical array. Then the values of DNL and PNL were computed using weak motions recorded in aftershocks as linear site response reference. Then the K-means clustering algorithm was introduced and applied in the nonlinear site response identification considering the DNL, PNL, PGA and the V_{S30} as clustering variables.

2. Strong ground motion data and process

The September 6th, 2018 Hokkaido eastern Iburi Earthquake (*M*w6.6) occurred as the result of shallow reverse rupture (according to the preliminary focal mechanism solution provided by the USGS, https://earthquake.usgs.gov). The hypocenter depth of the earthquake is 35.0 km and located in the island of Hokkaido. A total number of 208 KiK-net stations were triggered in the mainshock. To investigate the possible nonlinear seismic site response based on these strong motion data, we studied the stations with epicenter distance less than 200 km as shown in Fig.1.

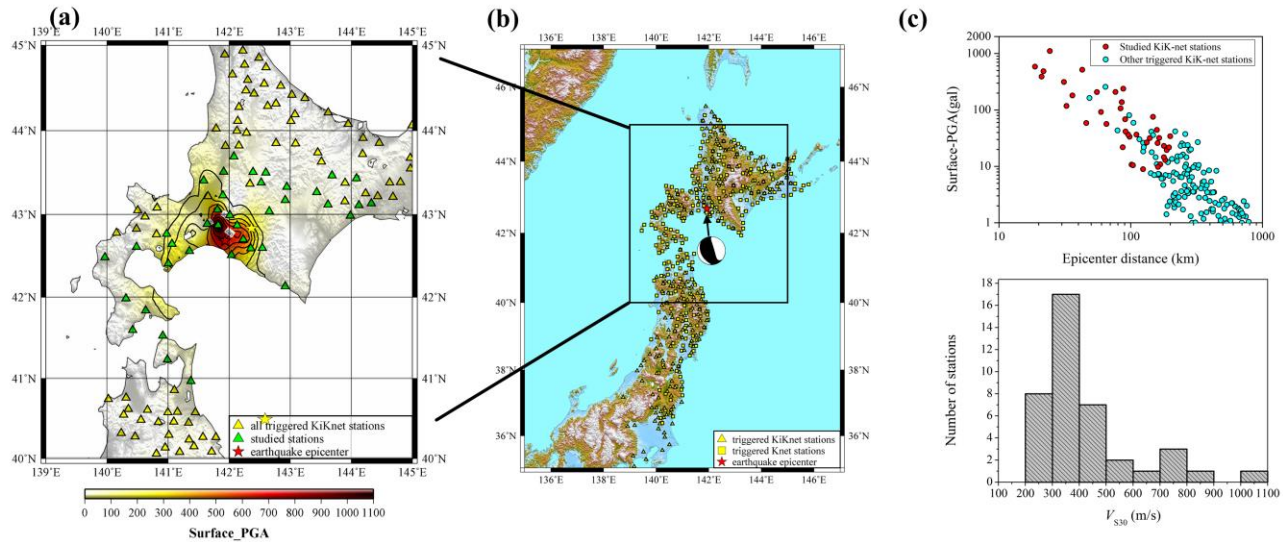


Figure 1. Location of the studied KiK-net stations and mainshock epicenter of the Hokkaido eastern Iburi earthquake; The surface PGA contour map of the studied region and the preliminary focal mechanism solution provided by USGS were also shown. (c) Surface PGA distribution with the epicenter distance and the Histograms of V_{S30} values for the selected stations

The weak motions were selected from 29 aftershocks for computation of reference linear site response according to the following criterions: (1) The geometric mean of surface-PGA of two horizontal components should be less than 30 cm/s^2 to remove stations potentially affected by soil nonlinearity. (2) At least five records matching criterion (1) were recorded in each station to ensure a relatively low scattering level of the spectral ratio curves. Finally, 39 KiK-net stations were selected to evaluate their nonlinear site response during the mainshock. The corresponding location of studied stations and the PGA contour map was shown in Fig.1. The station code, PGA (recorded at surface and borehole), and V_{S30} (a proxy for the site condition) of each station were listed in Table 1. The surface-PGAs correlating with epicenter distance of the selected records are shown in Fig.1(c), clearly illustrating that the selected records have a uniform distribution within a range of 200 km and half of them are higher than 50 gal. The histograms of the values of V_{S30} for the selected stations, two-thirds of them exhibit a value of V_{S30} smaller than 500 m/s respectively. Potential nonlinear site response may be observed in this region regarding the relatively high ground motions recorded and the relatively soft surface soil existed.

Table 1. Information of the selected 39 KiK-net strong-motion stations and the values of indicator (DNL and PNL) used for identifying the nonlinear site response

No.	Station Code	Epicenter Dis. (km)	PGA(gal)		V_{S30}(m/s)	DNL	PNL	Nonlinearity Identification result
			Borehole	Surface				
1	AOMH01	158	10.8	44.2	302	2.3	2.9	L
2	AOMH03	183	6.5	12.9	654	1.1	2.0	L
3	AOMH06	200	5.1	34.3	434	1.3	3.1	L
4	HDKH01	19	108.4	584.1	368	5.9	37.5	NL
5	HDKH03	31	36.1	312.8	341	3.6	21.8	NL
6	HDKH04	21	138.5	389.1	235	5.4	25.4	NL
7	HDKH05	46	22.5	58.6	766	2.0	5.0	L
8	HDKH07	98	9.7	33.4	459	1.3	2.4	L
9	HYMH01	168	4.9	10.8	395	1.2	4.0	L
10	HYMH02	160	10.1	25.8	498	1.1	2.9	L
11	IBUH01	24	218.7	1105.5	307	5.7	48.4	NL
12	IBUH02	22	102.2	485.8	542	4.9	32.7	NL
13	IBUH05	55	95.0	207.4	379	4.2	11.3	NL
14	IBUH06	88	46.0	238.4	304	5.4	14.5	NL
15	IBUH07	77	61.8	208.9	259	3.5	7.8	L
16	IKRH01	86	34.5	137.3	405	1.4	5.4	L
17	IKRH03	36	125.0	180.6	326	4.7	15.9	NL
18	KKWH08	66	9.0	56.5	311	2.8	11.2	L
19	KKWH12	102	6.3	10.7	771	1.9	4.7	L
20	KKWH13	96	10.5	36.7	356	2.4	4.8	L
21	KKWH14	87	9.9	21.8	538	3.5	5.9	L
22	KSRH01	188	7.0	19.4	215	4.4	7.0	L
23	KSRH02	179	7.5	23.1	219	3.2	5.8	L
24	KSRH07	196	5.5	21.7	204	1.9	5.3	L
25	KSRH09	165	8.4	31.9	230	2.0	6.2	L
26	OSMH01	179	4.4	14.5	239	2.0	4.6	L
27	OSMH02	148	25.6	75.3	325	2.6	2.8	L

28	SBSH08	84	14.8	106.9	325	3.5	10.7	L
29	SBSH09	124	5.1	8.9	719	1.0	2.1	L
30	SRCH06	111	16.9	36.6	321	2.0	6.2	L
31	SRCH07	60	20.2	92.0	316	2.0	4.0	L
32	SRCH08	91	17.5	68.4	347	4.0	10.3	L
33	SRCH09	43	109.4	515.2	241	5.4	20.2	NL
34	SRCH10	33	55.3	117.6	1027	2.4	7.3	L
35	TKCH01	161	4.9	9.8	445	1.2	3.8	L
36	TKCH03	133	9.1	26.5	372	3.2	4.7	L
37	TKCH04	91	15.4	41.4	446	2.0	7.2	L
38	TKCH05	140	6.7	31.8	337	1.5	4.4	L
39	TKCH10	104	6.8	10.4	804	1.7	5.1	L

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

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

3. Quantification of the nonlinear site response

3.1 Nonlinear site response indicator parameters

The DNL parameter (the degree of nonlinearity of site response) was proposed by Noguchi and Sasatani(2008)[6] as in equation (1).

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

Where R_{strong} is the spectral ratio value for strong ground motion in mainshock; R_{weak} is the average spectral ratio values computed using weak aftershock records; f_i is the i_{th} frequency. This parameter is calculated in the frequency range [0.5–20] Hz in this paper. N_1 is the first index of the frequency that is above 0.5 Hz, and N_2 is the last index of the frequency that is below 20.0 Hz.

In order to take into account the variability of the linear reference site response curve, the indicator PNL(the percentage of nonlinearity) was proposed by Régnier et al. (2013)[4] as follows:

$$A=\sum_{i=N_1}^{N_2}\begin{cases}(R_{strong}(i)-R^{+}_{weak}(i))\log 10(\dfrac{f_{i+1}}{f_i}) & \text{if } R_{strong}(i)\geq R^{+}_{weak}(i)\\ (R^{-}_{weak}(i)-R_{strong}(i))\log 10(\dfrac{f_{i+1}}{f_i}) & \text{if } R_{strong}(i)\leq R^{-}_{weak}(i)\\ 0\end{cases} \tag{2}$$

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

Where the $R^{-}_{weak}(i)$ and $R^{+}_{weak}(i)$ represent the values of the average linear spectral ratio curve minus and plus one standard deviation at i_{th} frequency f_i respectively; N_1 is the first index of the frequency that is above 0.5 Hz, and N_2 is the last index of the frequency that is below 20.0 Hz. It is normalized by the linear site-response spectral ratio curve in order to give an absolute estimation of the nonlinear soil behavior independent of the linear site-response amplitude.

3.2 DNL and PNL computation results

The DNL and PNL were calculated based on the records processed above. Although there is apparently positive correlation relationship between DNL, PNL and surface PGA as illustrated in Fig.2, the data are clearly scattered due to the variability of spectral ratio curves. In addition, the site condition also might has slight impact on the DNL and PNL value. The site which has high PGA and low DNL shows higher value of V_{S30}.

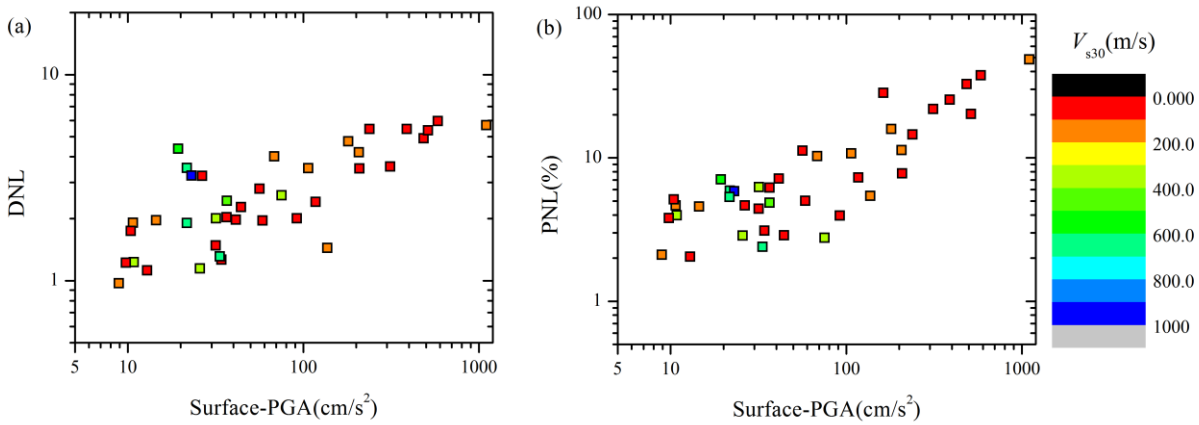


Figure 2. (a) DNL, (b) PNL versus the observed horizontal PGA at ground surface. The V_{s30} value were illustrated using different color.

The contour maps of DNL and PNL were plotted in Fig.3 using hermit interpolation, which indicates the generally spatial region of the occurrence of nonlinear site response. It turns out to be very hard to ascertain the exact region or stations with nonlinear site response if just by means of the empirical threshold of DNL or PNL. Besides the inherent scatter of DNL and PNL values, the ground motion strength and site condition will also influence the degree of the nonlinear site response. If we manually compare and check the spectral ratio curves of each station, it will cost a lot of time and lead to loss of objectivity in the results. Thus, we will use the K-means clustering algorithm to give a comprehensive explanation of the observed data in the next section.

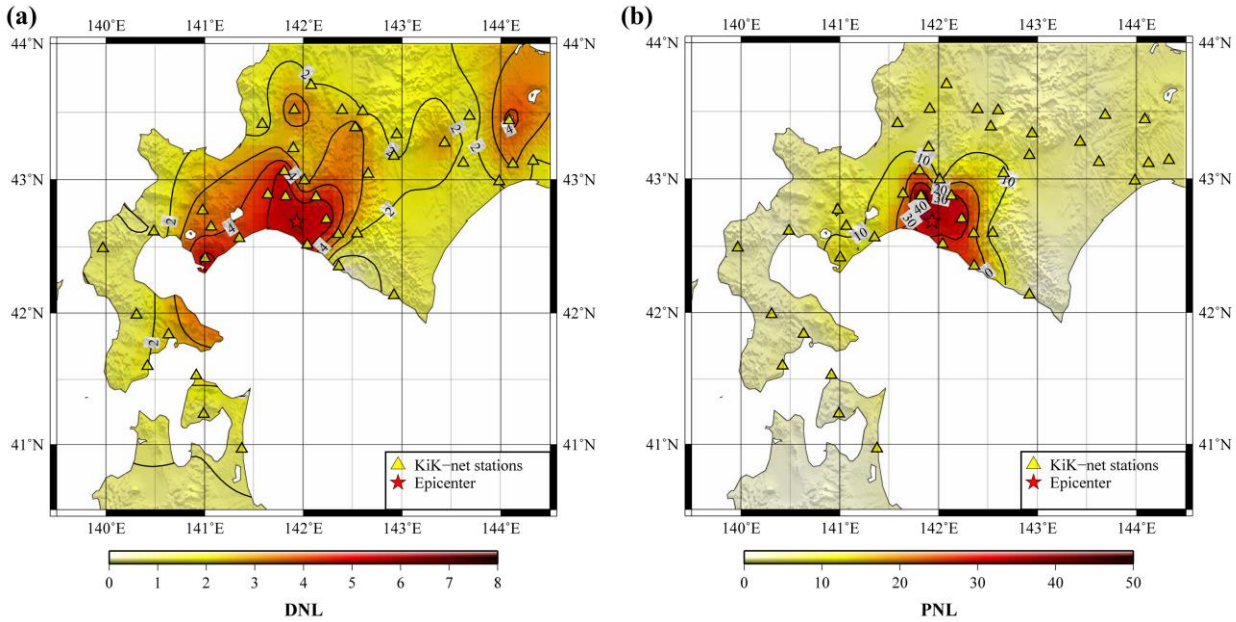


Figure 3. The DNL and PNL contour map calculated using ground motions recorded at KiK-net

4. Clustering the 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 similarity. The generated clusters consist of a set of data that are similar to each other in the same group but dissimilar from data in other groups. The clustering analysis categorizes unlabeled data based on the observations themselves only and is thus regarded as an unsupervised classification procedure, which is the most essential characteristic compared with traditional empirical-model based classification procedure.

K-means clustering is one of the most widely used methods for cluster analysis in data mining [9]. K-means clustering aims 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, …, x_n), where the n_{th} observation x_n is a d-dimensional real vector with d explanatory variables, K-means clustering partitions n observations into k sets $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the sum of variance J as defined in Eq.(4). The $\sum_{k=1}^{K}(|x_n - \mu_k|^p)^{\frac{1}{p}}$ is the Minkowaki distance to describe the "difference" between the observation x_n and the k_{th} clustering centroid point μ_k. When p equals 1 and 2 respectively, it represents the commonly used Euclidean distance and the Manhattan distance (or City-block distance) [10]

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

The main steps are:

Step1: k initial "means" are randomly generated within the data domain.

Step2: The corresponding k clusters are created by associating observation with the nearest mean.

Step3: The centroid of each of the k clusters μ_k becomes the new mean.

Step4: *Step* 2 and 3 are repeated until convergence criteria is reached, that is, it reaches the specified maximum number of iterations or the centroid of each cluster does not change.

Considering our problem, each station is treated as one observation with four explanatory variables characterizing nonlinear site response, including the ground motion intensities (Surface-PGA or Borehole-PGA), soil condition (V_{S30}), DNL and PNL. The task of K-means clustering algorithm is to partition the stations into at least two clusters representing stations with nonlinear and linear site response respectively. As one of the mature clustering analysis tools, the Clustering Toolbox of MATLAB [11] was utilized in this paper to achieve this goal.

4.2 Multicollinearity diagnosis and standardizing the clustering variables

Before we utilize the K-means clustering, multicollinearity diagnosis is needed to be carried out in the first step to guarantee that there are no completely linear correlations between explanatory variables[12]. Otherwise, one variable could be linearly expressed by other variables, and the redundant variables may cause unnecessary fluctuation in the clustering results due to a few outlier data. The variance inflation factor, VIF, is commonly used to estimate the degree of the multiconllinearity between different variables[13][14]. First running a least square linear regression that the i_{th} variable v_i is represented as a function of all the other explanatory variables. The f_i is the regression prediction value.

$$v_i = f_i + \varepsilon = a_1 v_1 + a_2 v_2 + \ldots a_{i-1} v_{i-1} + a_{i+1} v_{i+1} \ldots + b + \varepsilon \tag{5}$$

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

The VIF_i of v_i is computed as follows:

$$VIF_i = \frac{1}{1-R_i^2} \tag{6}$$

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

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

Significant multicollinearity exists when the VIF value exceeds the threshold of 10.0 (Ref.[15]). The VIF values and Pearson correlation coefficients were calculated for four explanatory variables regarding KiK-net observations as illustrated in Table.2. The VIF values for Surface-PGA(H) and PNL variables are all larger than 10.0 as shown in Table.4. The Pearson correlation coefficients between surface-PGA and PNL reached 0.95, which indicates a strong linear correlation between them. Therefore, the Surface-PGA was replaced by Borehole-PGA to represent the ground motion level. New VIF values for Borehole-PGA(H), DNL, PNL and V_{S30} were calculated respectively and listed in Table.3 as case B, indicating no predominant multicollinearity correlation with largest VIF value as 5.2.

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**

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}, 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:

$$v_i' = \frac{v_i - v_i^{min}}{v_i^{max} - v_i^{min}} \tag{8}$$

4.3 Optimum clustering number

Before we perform the K-means clustering analysis, it is necessary to determine the optimum number of clusters. As the only prior information given in the whole unsupervised clustering analysis process, the number of clustering will significantly influence the clustering result. There is no standard procedure for computation of the optimum clustering number. The most popular criteria is the Calinski and Harabasz (1974) [16]F-stopping-rule index which is based on the within-cluster sum of difference squares. It is a measure of (dis-)similarity between clusters, that is, measures the degree of homogeneity between groups. The larger the value of Calinski-Harabasz index is, the more significant the differences among groups are, and the more acceptable the clustering number is. Different clustering numbers and the corresponding Calinski-Harabasz index regarding KiK-net stations are illustrated in Fig.4(a). The comparison results indicated that two clusters has the largest value of Calinski-Harabasz index, indicating that the optimum number of clustering is two. The results are consistent with the problem which this study deals with, that is, one cluster groups the stations with nonlinear site response and another one groups the stations with linear site response. The K-means clustering analysis was then performed for the observation data from KiK-net stations which would be separated into two clusters.

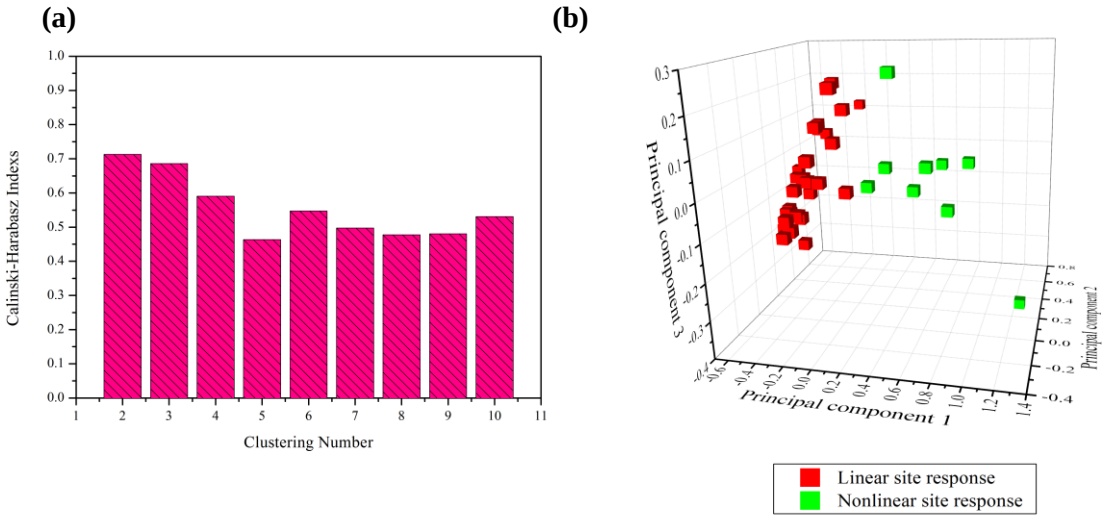


Figure 4. (a) The clustering number and corresponding Calinski-Harabasz criterion index. (b)The original observation data are represented in a 3-dimensional space using PCA method. Different color indicates the nonlinearity identification results using K-means algorithm.

4.4 Clustering results validation

It is difficult to visualize the classification results due to each observation data having four explanatory variables that requires a visualization in a four-dimensions space. Therefore, we used the principal component analysis (PCA) to reduce the dimensions of the observation data. PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the projection of the data comes to lie on new coordinates (called the principal component)[17]. The greatest variance by projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. The PCA results indicated that three components account for more than 95% of the variance, which could represent the original observation data we studied. The original four-dimensions observation data for KiK-net stations were projected into three-dimensions space as shown in Fig.4(b). The classification results using K-means algorithm are all visually separated into two clusters, indicating a relatively satisfactory classification result.

The clustering results were then compared with those based on the empirical thresholds of the widely recognized nonlinear indicator, such as DNL, PNL, PGA as shown in Fig.5. Noguchi and Sasatani (2008) [6]suggested a DNL value of 4.0 for H/V ratio method as the boundary of nonlinear site response

identification, which has been used in many studies. (e.g. [5][18]). Although there is a difference between the surface/borehole and H/V spectral ratios, the DNL_{SB} and DNL_{HV} thresholds are nearly equivalent[6]. The empirical threshold values of PNL and Borehole-PGA suggested by Régnie et al. (2013)[4] were 10 % and 50 gal respectively based on a large number of KiK-net data in Japan. Ranging from 100 gal to 200 gal, however, the surface-PGA threshold value is still debated. Our classification results were in good agreement with the threshold value of 200 gal suggested by Ref.[5]. The correlation between site condition (i.e., V_{S30}) and the degree of nonlinear site response (i.e., DNL or PNL) is obviously weak, and there is no convincing threshold for the V_{S30}. It can be observed that the sites with nonlinear site response classified by clustering algorithm (i.e., cluster A) are mostly located within the empirical nonlinearity region only except for two or three data near the boundary. It proves that the clustering results are quite convincing and robust from the perspective of nonlinearity indicators distribution.

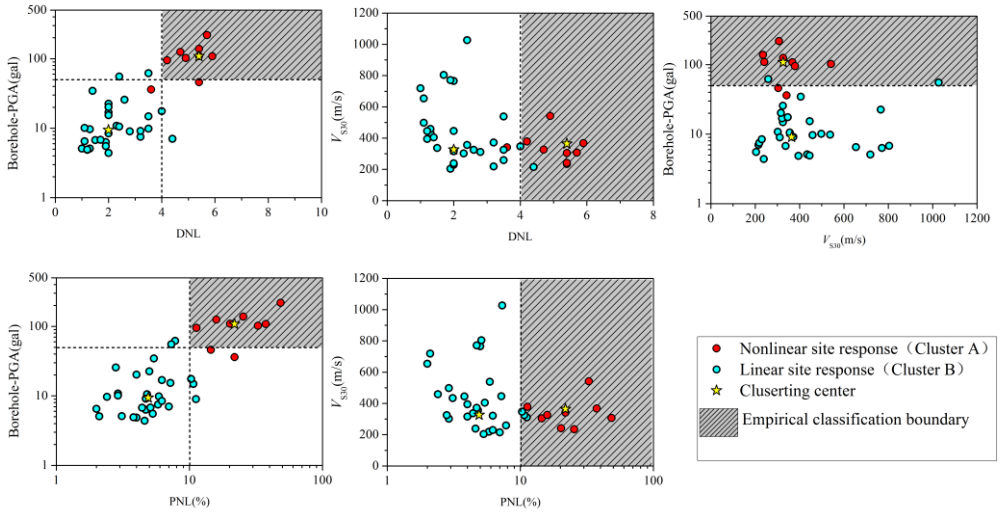


Figure 5. Values of DNL, PNL for 39 KiK-net stations versus the recorded PGA and V_{S30}. The dots with different color indicate the classification results computed using K-means clustering algorithm in this paper. The dashed lines indicate the empirical classification threshold of DNL, PNL, Borehole-PGA and Surface-PGA. The shaded area means the regions covering the indicator values of beyond their empirical thresholds which defines the nonlinear site response

The surface-borehole spectral ratio curves for stations with nonlinear site response according to clustering results were shown in Fig.6. It can observe apparently a systematic decrease of the peak frequencies associated with a decrease of their amplitude. The computed spectral ratios using the recordings of the main event were amplified at frequencies below the predominant frequency and were deamplified above it, which illustrates typical nonlinear soil behavior. These indicate that the identification results of nonlinear site response using the clustering method are reasonable and accurate. The whole clustering identification process is completely automatic and efficient without any manual intervention, while the satisfactory and objective results were obtained as shown in the Fig.5 and Fig.6.

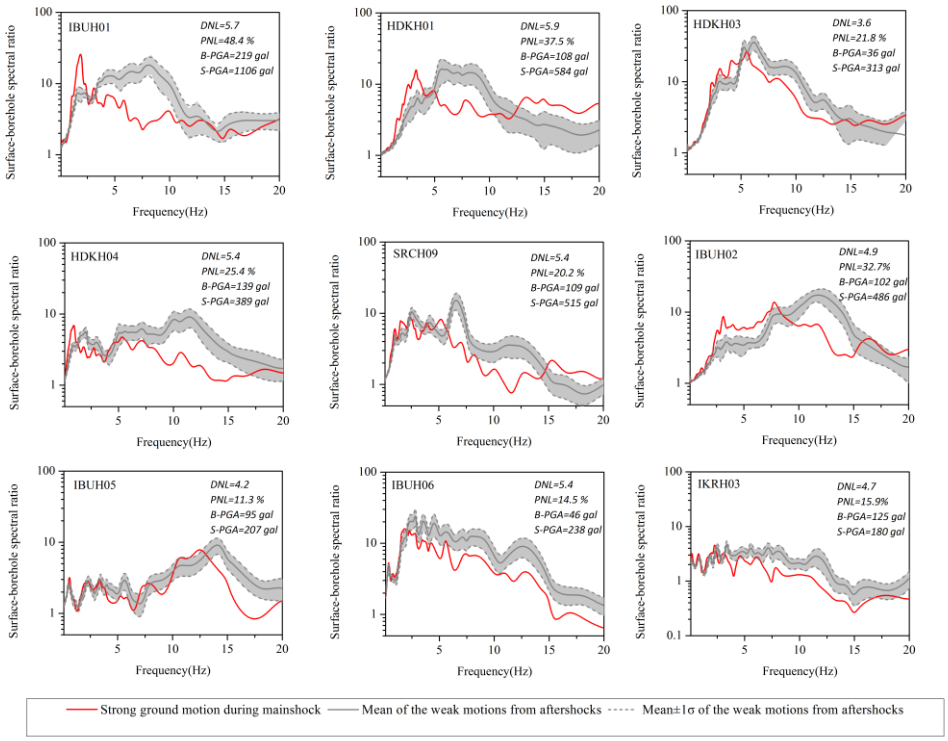


Figure 6. The surface-borehole spectral ratio curves for 9 KiK-net stations identified as those with nonlinear site response using the K-means clustering algorithm. The shaded area indicates the range of the mean plus and minus one standard deviation.

6. Conclusions

This paper firstly incorporated the clustering algorithm in machine learning to deal with the nonlinear seismic site response classification problem using the Hokkaido eastern Iburi earthquake as a study case.

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

To validate the clustering identification results, we adopted different methods to comprehensively evaluate them from different aspects. Firstly, Principal component analysis (PCA) was carried out to intuitively illustrate the clustering results by reducing the dimensions of the data. The values of DNL, PNL, PGA, and V_{S30} are visually separated into two clusters. The clustering classification results are overall consistent with the results indicated by empirical nonlinearity empirical threshold proposed by other scholars. For the identified stations with nonlinear site response, an obvious deamplification and shift of frequency could be observed between the main shock and reference linear spectral ratio curve which is the typical site nonlinearity characteristics.

This study presented an interesting example solving the seismology problems using state-of-art matching learning technique. Using classical seismology techniques with machine learning algorithm in a hybrid approach, it is possible that we could extract novel insights directly from the data and solving more and more problems in the field of earthquake engineering.

7.Acknowledgement

The strong-motion waveform records used in this study were obtained from the National Research Institute for Earth Science and Disaster Resilience (NIED) in Japan. The raw KiK-net data were downloaded from the websites at: http://www.kyoshin.bosai.go.jp/, last accessed in Feb.2019. This work was partially supported by the Science Foundation of the Institute of Engineering Mechanics, CEA [grant number. 2019B09]. Chinese National Natural Science Fund [grant number. 51908518 & 51778589]; Natural Science Foundation of Heilongjiang Province [grant number. E2017065];

8.References

[1] Wen, K. L. (1994). Non-linear soil response in ground motions. *Earthquake engineering & structural dynamics*, **23**(6), 599-608.

[2] Wen, K. L., Chang, T. M., Lin, C. M., & Chiang, H. J. (2006a). 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 (pp. 225-232).

[3] Wen, K. L., Huang, J. Y., Chen, C. T., & Cheng, Y. W. (2011). Nonlinear site response of the 2010 Darfield, New Zealand earthquake sequence. *In 4th iaspei/iaee international symposium: Effects of surface geology on seismic motion*.

[4] Régnier, J., Cadet, H., Bonilla, L. F., Bertrand, E., & Semblat, J. F. (2013). Assessing nonlinear behavior of soils in seismic site response: Statistical analysis on KiK‐net strong‐motion data. *Bulletin of the Seismological Society of America*, **103**(3), 1750-1770.

[5] Ren, Y., Wen, R., Yao, X., & Ji, K. (2017). Five parameters for the evaluation of the soil nonlinearity during the Ms8. 0 Wenchuan Earthquake using the HVSR method. *Earth, Planets and Space*, **69**(1), 116.

[6] Noguchi, S., & Sasatani, T. (2008). Quantification of degree of nonlinear site response. *In 14th world conference on earthquake engineering*, Beijing, paper ID (pp. 03-03).

[7] Bailey, K. (1994). Numerical taxonomy and cluster analysis. *In Typologies and taxonomies: an introduction to classification techniques*, pp, 34, -6524.

[8] Miettinen, K. (2012). Nonlinear multiobjective optimization (Vol. 12). *Springer Science & Business Media*.

[9] Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms (pp. 1027-1035). *Society for Industrial and Applied Mathematics*.

[10] Kriegel, H. P., Schubert, E., & Zimek, A. (2017). The (black) art of runtime evaluation: Are we comparing algorithms or implementations?. Knowledge and Information Systems, 52(2), 341-378.

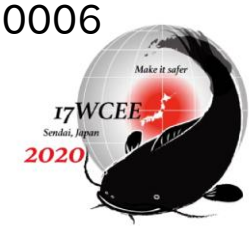
[11] Mathworks Inc (2014). Clustering Toolbox User's Guide.

[12] Spicer, J. (2005). Making sense of multivariate data analysis: An intuitive approach. Sage.

[13] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis (Vol. 821). *John Wiley & Sons*.

[14] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). *New York: springer*.

[15] Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). Applied Linear Regression Models (4th ed.). *McGraw-Hill Irwin*.



[16] Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1), 1-27.

[17] Jolliffe, I.T. (2001). Principal Component Analysis, *Series: Springer Series in Statistics, 2nd ed.*, Springer, NY, XXIX, 487 p. 28 illus. ISBN 978-0-387-95442-4.

[18] Dhakal, Y. P., Aoi, S., Kunugi, T., Suzuki, W., & Kimura, T. (2017). Assessment of nonlinear site response at ocean bottom seismograph sites based on S-wave horizontal-to-vertical spectral ratios: a study at the Sagami Bay area K-NET sites in Japan. *Earth, Planets and Space*, 69(1), 29.

