https://doi.org/10.20965/jdr.2018.p0254

Paper:

Cluster Analysis of Long-Period Ground-Motion Simulation Data with Application to Nankai Trough Megathrust Earthquake Scenarios

Takahiro Maeda*,†, Hiroyuki Fujiwara*, Toshihiko Hayakawa, Satsuki Shimono**, and Sho Akagi****

*National Research Institute for Earth Science and Disaster Resilience
3-1 Tennodai, Tsukuba, Ibaraki 305-0006, Japan
†Corresponding author, E-mail: tmaeda@bosai.go.jp
**Mitsubishi Space Software Co., ltd., Ibaraki, Japan
[Received November 17, 2017; accepted February 8, 2018]

We developed a clustering method combining principal component analysis and the k-means algorithm, which classifies earthquake scenarios based on the similarity of the spatial distribution of earthquake ground-motion simulation data generated for many earthquake scenarios, and applied it to long-period ground-motion simulation data for Nankai Trough megathrust earthquake scenarios. Values for peak ground velocity and relative velocity response at approximately 80,000 locations in 369 earthquake scenarios were represented by 15 principal components each, and earthquake scenarios were categorized into 30 clusters. In addition, based on clustering results, we determined that extracting relationships between principal components and scenario parameters is possible. Furthermore, by utilizing these relationships, it may be possible to easily estimate the approximate ground-motion distribution from the principal components of arbitrary sets of scenario parameters.

Keywords: long-period ground motion, simulation, principal component analysis, clustering

1. Introduction

In seismic hazard assessments, that lead to the quantitative estimate of seismic risk, it is necessary to perform detailed evaluations of seismic ground motion for every points likely to be affected. In addition, evaluating appropriate range of possible ground motions at each point is also crucial. As is clear from the example of the 2011 Tohoku earthquake, predicting an exact source model (earthquake scenario) of a megathrust earthquake in detail before the earthquake is impossible. Even with limited earthquake scenarios to be evaluated, knowing possible range of ground-motion would requires many ground-motion simulations to account for the uncertainties in source models.

In order to effective application of the results of these ground-motion simulations to the disaster resilience, it is necessary to easily extract useful information from the vast amount of data derived from simulations and to visualize it. The authors have built a system that aggregates the many seismic waveforms produced by the ground-motion simulations using parallel distributed processing and visualized the statistical quantities for the maximum amplitude calculated at each point [1]. The authors also have built a parallel distributed processing system that uses clustering to extract relationships between the characteristics of ground motion in each area and the source model (scenario parameters) from the ground-motion simulation results for many earthquake scenarios [2]. The authors have applied these systems to long-period ground-motion simulation data for Nankai Trough megathrust earthquake scenarios and examined their effectiveness.

In this study, we attempt to improve the efficiency of extracting earthquake scenarios having similar ground-motion distribution and establish relationships between characteristics of shaking at each area with scenario parameters. To achieve this goal, we combine k-means clustering, which we used in previous examinations, with principal component analysis (PCA) for more efficient computations. We then apply the method to long-period ground-motion simulation data for Nankai Trough megathrust earthquake scenarios.

2. Method and Results

In **Fig. 1**, we show the analysis pipeline used in our study that combines principal component analysis and clustering. The procedure consists of the following three steps.

1 We standardize the ground-motion indices of the ground-motion simulation at Q grid points to produce a mean of 0 and variance of 1. This operation is performed for each of the N scenarios.

2 Using the standardized ground-motion indices from the N scenarios, we extract P principal components

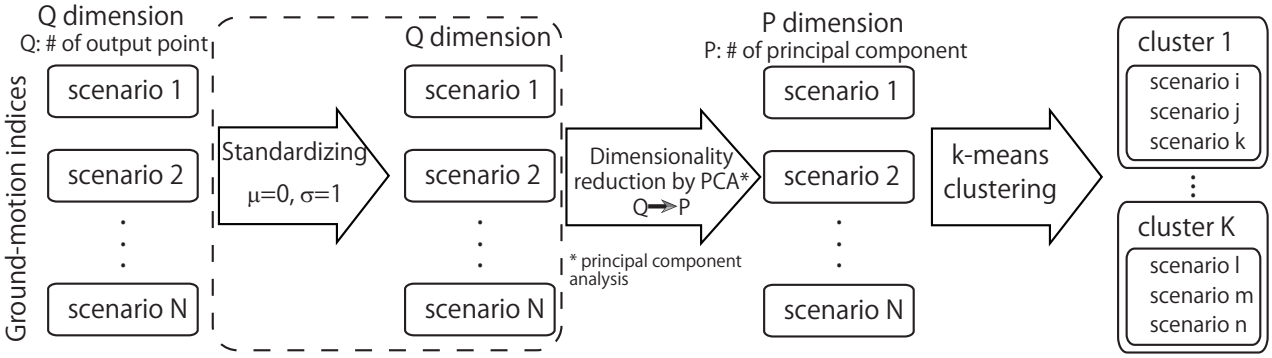


Fig. 1. Flowchart of analysis pipeline.

to reduce the number of dimensions (Q→P). With this, the ground-motion distribution from each scenario, which was represented by standardized indices at Q grid points, can be represented as coefficients for the P principal components.

3 We apply the k-means algorithm to the P-dimensional data to find clusters of scenarios.

In the following section, we describe the details of the method with application to long-period ground-motion simulation data for Nankai Trough earthquakes [3] as an example. We use as the input the peak-ground velocity and relative velocity response (damping of 5%; periods of 3, 5, 7, 10, and 20 s) at 77,609 locations based on ground-motion simulations of 369 earthquake scenarios. These ground-motion indices are produced from the simulated waveforms. First, we apply standardization to the input data so that for each scenario the mean is 0 and variance is 1. With this standardization, differences between scenarios resulting from differences in earthquake magnitude vanish, and the data are converted to a form in which the spatial distributions of ground motion can be compared directly. Although standardization is required when data with different units are used, we use data having identical units. For this reason, we also examine the results of using unstandardized data. In the following, however, unless otherwise noted, the results with standardization will be shown.

We next performed PCA (e.g., [4]) on the input data from 369 scenarios. PCA is a method that finds the P vectors having the largest variances with respect to high-dimensional input data and, by taking projections onto those directions, obtains a P-dimensional subspace that retains the characteristics of the input data. The subspace can be expressed as $t_{il} = x_{ij} \cdot w_{jl}$, where x_{ij}, w_{jl}, and t_{il}, represent the elements of the $n \times Q$ input matrix, the $Q \times P$ coefficient matrix, and the $n \times P$ principal component score matrix, in which $i = 1, \ldots, n$, $j = 1, \ldots, Q$, $l = 1, \ldots, P$, n is the number of scenarios, and Q is the dimension of the input data (number of output points). Dimension of the input can be reduced by choosing a value of P that is smaller than Q. In this study, we imposed conditions in which, for each of the six ground-motion indices, the proportion of variance for each principal component is greater than 1% and the cumulative proportion of variance is greater than 85%. We decided to use up to the 15th principal component (**Fig. 2**). The proportion of variance refers to the ratio between the variance accounted for by a principal component and the total variance.

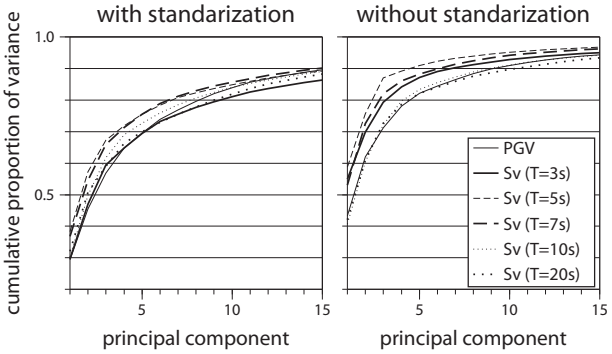


Fig. 2. Cumulative proportion of variance up to the 15th principal component. Left: using standardized data. Right: using unstandardized data. Results for the six ground-motion indices are shown.

Figure 3 shows the results of PCA. Several of the principal components from the 1st to the 15th are shown, which were calculated for relative velocity response values (periods of 3, 5, 7, and 10 s) as examples. Focusing on the 5-s period results, the 1st principal component had large amplitudes in the left half (West Japan) of the computed region and small amplitudes in the right half (East Japan), and the 2nd principal component had large amplitudes near the center of the computed region (the Kinki region) and small amplitudes on the left (Kyushu) and right (Chubu and Kanto) sides. The 3rd principal component had large amplitudes on the right side (Kanto region) and small amplitudes in the center and on the left side (West Japan). Even though the data for each period was analyzed independently, the 1st and 2nd principal components had common characteristics at each period. With this analysis, the ground-motion distribution for each scenario that had been expressed by amplitude values at 77,609 locations was expressed approximately as a linear combination of 15 principal components.

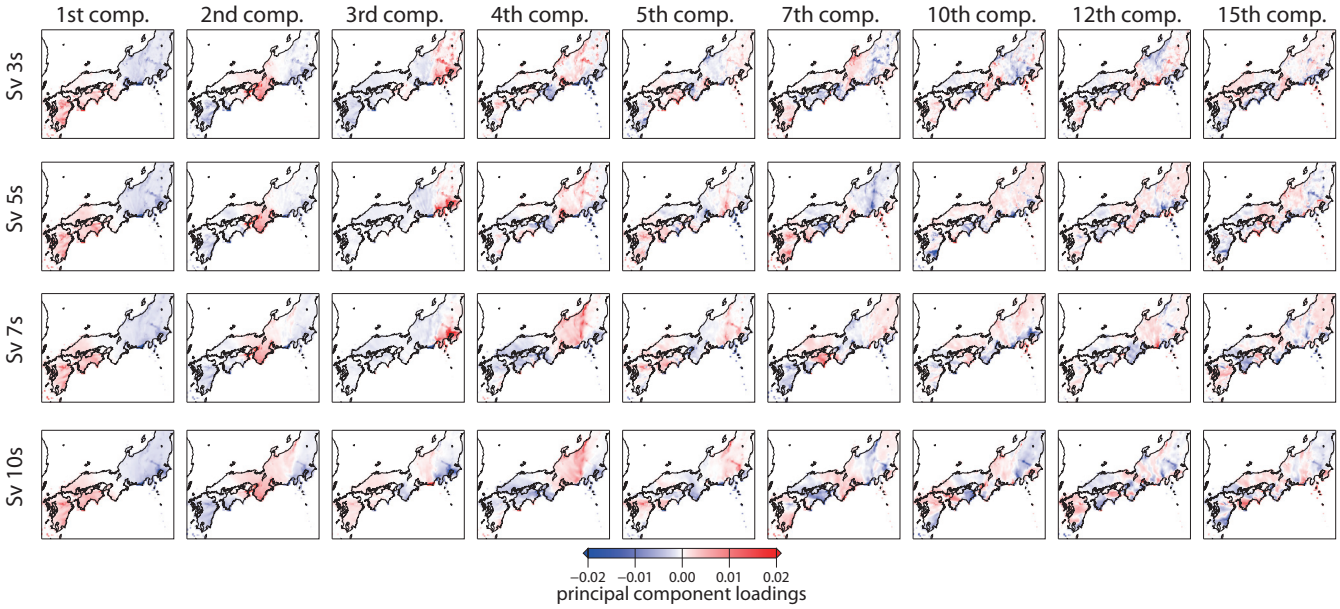


Fig. 3. The results of principal component analysis (principal components numbered one through five, seven, ten, twelve, and fifteen are shown as an example). From the top, results using the relative velocity response values (damping = 5%) at periods of 3, 5, 7, and 10 seconds. In each panel, the amplitude is normalized so that the sum of the squares of the amplitude values of all points becomes 1.

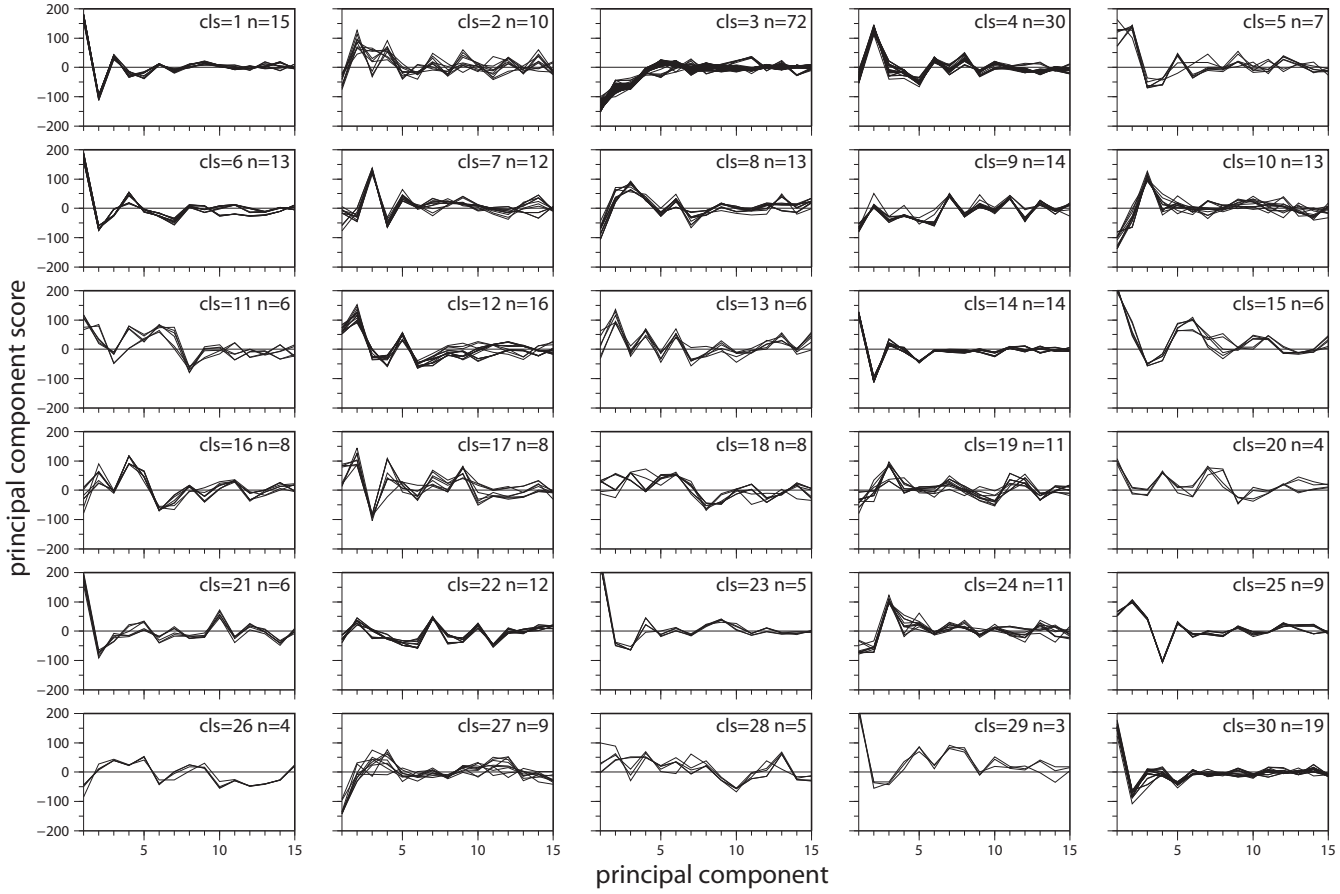


Fig. 4. Clustering results with 5-s period relative velocity response values (damping = 5%) as the input. In each panel, cls is the cluster number and n is the number of scenarios in the cluster.

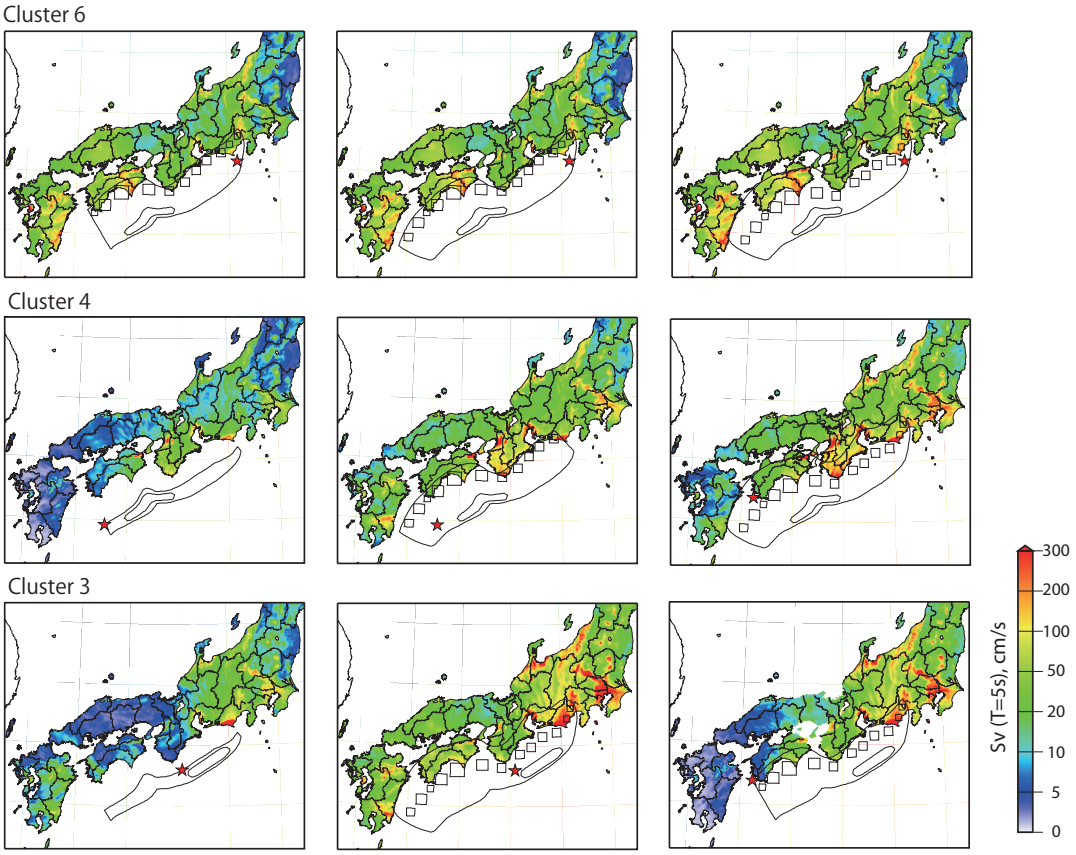


Fig. 5. Comparisons of ground-motion distribution (5-s period relative velocity response values, damping = 5%) among scenarios in the same cluster. Top: within the 6th cluster. Middle: within the 4th cluster. Bottom: within the 3rd cluster. Source models are shown inside the figure, where asperity is represented by rectangles and the rupture starting point by red stars within the source area surrounded by black lines.

We next performed clustering of earthquake scenarios using the k-means algorithm (e.g., [5]), which is a non-hierarchical clustering algorithm. The k-means algorithm categorizes the data into k clusters so that the total sum of the squared Euclidean distances between data points to their cluster centers is minimized. Before applying k-means, fixing the number of clusters k was necessary. Based on the conditions that multiple similar clusters can exist in the results and that only similar scenarios should be included within a cluster, we decided after testing multiple values that $k = 30$ was a good value. We used the Python machine learning library scikit-learn for our clustering analysis.

In **Fig. 4**, we show the results of partitioning the scenarios into 30 clusters using the 5-s period relative velocity response values. Each scenario was represented by 15 coefficients with respect to the principal components (principal component scores). This means that the ground-motion distribution for each scenario could be represented by multiplying the principal components shown in **Fig. 3** by the corresponding coefficients shown in **Fig. 4**, then summing over them. **Fig. 4** shows all scenarios within a given cluster on top of each other, and we can see that clusters were composed of scenarios with similar principal component scores and that there was little dispersion within a cluster. We note that there is no particular meaning to the ordering of the clusters. From these results, we can see for example that the 6th cluster (cls = 6) had a large contribution from the 1st principal component, and the 4th cluster (cls = 4) contained a large contribution from the 2nd principal component. The 3rd cluster included the largest number of earthquake scenarios.

Figure 5 shows the distributions for relative velocity response values (damping of 5%, period of 5 s) in earthquake scenarios contained in (as shown in the figure from top to bottom) the 6th, 4th, and 3rd clusters. Because the unstandardized amplitudes are used here, the differences in absolute values stand out. However, if focus is directed to the characteristics of the spatial distribution, the 6th cluster had a tendency for amplitudes in West Japan to be larger, and this matches the characteristics of the 1st principal component (**Fig. 3**). Similarly, the 4th cluster had a distribution in which the amplitude in the Kinki region (near the center of the computed region) was large relative to other regions, which matches the characteristics of the 2nd principal component. For the 3rd cluster, amplitudes to the east of the Chubu region (right half for the computed region) tended to be large, and this is consistent with the principal component scores from the 1st through 3rd principal components being negative. Although the 3rd cluster contained the largest number of earthquake scenarios, the occurrence probability of each earthquake sce-

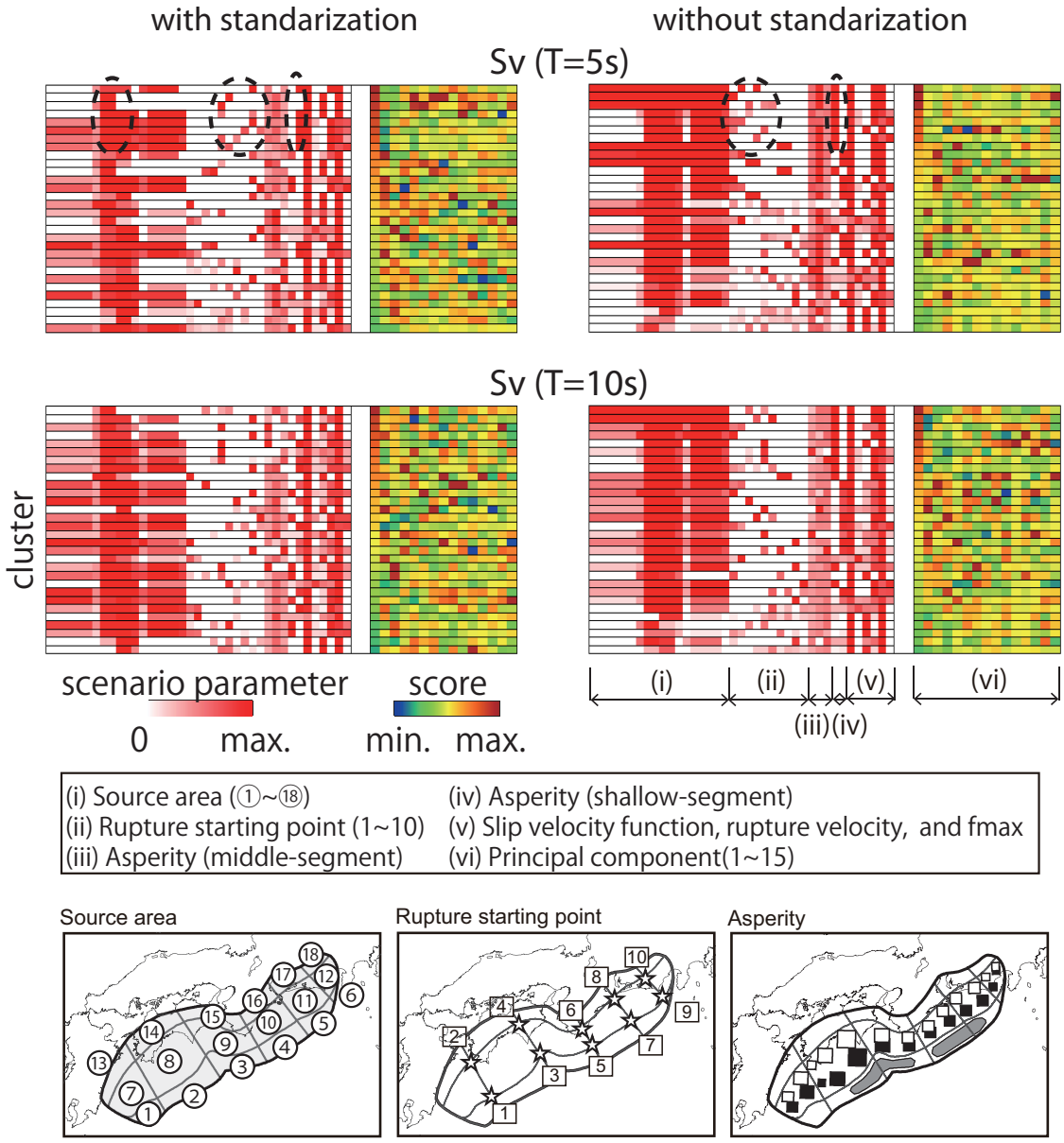


Fig. 6. Relationships between scenario parameters and principal component scores for each cluster. Top left: results for standardized 5- and 10-s period data. Top right: results for unstandardized 5- and 10-s period data. The columns to the left in white-red represent scenario parameters, those to the right in blue-yellow-red are principal component scores, and each row corresponds to a cluster. Details of the scenario parameters are shown at the bottom. Note that we only show the two patterns of middle-segment asperity (deep and shallow cases) in the asperity panel.

nario was different [3], and it is not necessarily the case that the ground-motion characteristics of the 3rd cluster represented the most likely ground-motion distribution to occur in a Nankai Trough earthquake. We succeeded in performing a cluster analysis of earthquake scenarios using the method outlined in **Fig. 1** based on the similarity of ground-motion spatial distribution.

3. Discussion

We examined the relationships between the scenario parameters (source parameters) within each cluster and the principal component scores. The scenario parameters consisted of a source area (18 small regions), rupture starting points (10 points), middle-segment asperity (three patterns), shallow-segment asperity (two patterns), slip velocity function at shallow-segment asperity (two patterns), shallow-segment rupture velocity (two patterns), and f_{max} (two patterns) for a total of 39 components. Parameters for each scenario were represented in Boolean form, with components that apply to the scenario as 1, otherwise as 0. In **Fig. 6**, we plotted the averages of principal component scores and scenario parameters for all scenarios in each cluster. The color scales were determined for each scenario parameter and principal component score so that the magnitudes could not be compared directly between different scenario parameters or principal component scores.

Fig. 6 shows the relationships between scenario parameters and principal component scores based on clustering results using relative velocity response values with periods of 5 and 10 s. The clusters were rearranged in descending order of 1st principal component score. With focus on the 5-s period results, clusters that were ranked high in the 1st principal component score had the following characteristics: 1) the rupture starting point tended to be located in the center or to the east of the source area, 2) the source areas may have been strictly in the western regions, and 3) the shallow-segment asperity was located on the western side (parts circled by dashed lines in the figure). With this combination of scenario parameters, the rupture in the shallow-segment asperity propagated from east to west, and with the forward directivity effect, this was expected to result in larger ground-motion amplitudes to the west of the source area. This coincided with the characteristics of the 1st principal component shown in **Fig. 3**, in which the amplitude was larger in West Japan, and we confirmed that a relationship existed between scenario parameters and principal component scores.

For reference, we also show the results of using unstandardized data (**Fig. 6**). With no standardization, the influence of the magnitude of earthquakes on the amplitudes was retained, and the size of the source area strongly contributed to the clustering results. With standardization, scenarios with different source areas were contained in one cluster and, therefore, different shades of color are shown in the source area parameters. In the unstandardized case, the colors are either very dark or very light, indicating that the clusters are composed of scenarios with the same source areas. In both the 5- and 10-second period cases, clusters having large contributions from the 1st principal component tended to have rupture starting points from the center to the west of the source area, and shallow-segment asperity tended to be on the eastern side. This combination of parameters was expected to make the ground-motion amplitude larger to the east of the source area, and although we omit the details because of space limitations, we confirmed that this is consistent with the characteristics of the 1st principal component with respect to the 5-s period unstandardized data.

We showed that earthquake scenarios can be categorized by clustering them into similar groups and that the relationships between principal components and scenario parameters can be extracted from the clustering results. It is possible that, by quantitative evaluation of the relationships between principal components and scenario parameters, ground-motion distributions corresponding to an arbitrary scenario parameter can be estimated. It is also possible to suggest additional earthquake scenarios by examining scenario parameters corresponding to the ground-motion distribution that cannot be represented by the current clusters.

4. Conclusion

In this study, we developed a categorization method that combines principal component analysis and the k-means algorithm, for ground-motion indices derived from large-scale high-resolution simulation data for many earthquake scenarios. We applied the method to long-period ground-motion simulation data of Nankai Trough megathrust earthquakes. As a result, earthquake scenarios that were similar in terms of ground-motion distributions were grouped together by means of clustering, and we showed that the relationships between principal components and scenario parameters can be extracted. In this investigation, we examined qualitative relationships between scenario parameters and principal component scores, and if these relationships could be evaluated quantitatively such as by inverse analysis, it was generally possible to estimate easily the approximate ground-motion distribution for an arbitrary combination of scenario parameters using the principal components. In addition, we determined that if it is possible to extract earthquake scenarios that are not categorized within the current clusters, we can effectively enrich the earthquake scenario set that forms the basis for seismic hazard assessment.

Because we used only the ground-motion amplitude data, it was still necessary to examine the appropriate ground-motion indices in order to extract useful information about seismic hazards. In addition, it is necessary to examine a different set of earthquake scenarios with different source areas. These hazard assessments are the premises for risk assessments. Another important task for future work is to perform a similar analysis for risk information, such as the exposed population evaluated from hazard information.

This study is an example of technology development for sharing and utilizing high-capacity and high-level disaster information from large-scale high-resolution numerical simulations. To advance the development of such technologies, the National Research Institute for Earth Science and Disaster Resilience maintains an earthquake hazard information-sharing system known as the Japan Seismic Hazard Information Station (J-SHIS) [6] and is developing mutual operation-type information-sharing platform, known as the e-Community Platform [7] and cloud system for joint public-private crisis management [8]. Expanding the functionalities of these systems is also a major task.

Acknowledgements

This study was supported by CREST/JST. Some figures in this paper were produced using GMT [9]. This manuscript was improved by constructive comments by two anonymous reviewers.

References:

[1] T. Maeda and H. Fujiwara, "Seismic Hazard Visualization from Big Simulation Data: Construction of a Parallel Distributed Processing System for Ground Motion Simulation Data," J. Disaster Res., Vol.11, No.2, pp. 265-271, 2016.

[2] T. Maeda and H. Fujiwara, "Seismic Hazard Visualization from Big

Simulation Data: Cluster Analysis of Long-Period Ground-Motion Simulation Data," J. Disaster Res., Vol.12, No.2, pp. 233-240, 2017.
[3] T. Maeda, A. Iwaki, N. Morikawa, S. Aoi, and H. Fujiwara, "Seismic-Hazard Analysis of Long-Period Ground Motion of Megathrust Earthquakes in the Nankai Trough Based on 3D Finite-Difference Simulation," Seismological Research Letters, Vol.87, No.5, doi:101785/0220160093, 2016.
[4] H. Hotelling, "Analysis of a complex of statistical variables into principal components," J. of Educational Psychology, Vol.24, pp. 417-441, 1933.
[5] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations," Proc. of 5th Berkeley Symposium on Mathematical Statistics and Probability, 1, University of California Press, pp. 281-297, 1967.
[6] http://www.j-shis.bosai.go.jp/map/?lang=en [accessed Nov. 17, 2017]
[7] http://ecom-plat.jp/ [accessed Nov. 17, 2017]
[8] http://ecom-plat.jp/k-cloud/index.php [accessed Nov. 17, 2017]
[9] P. Wessel and W. H. F. Smith, "New version of the Generic Mapping Tools released," Eos Transactions, American Geophysical Union, Vol.76, p. 329, 1995.

Name:
Takahiro Maeda

Affiliation:
Chief Researcher, National Research Institute for Earth Science and Disaster Resilience (NIED)

Address:
3-1 Tennodai, Tsukuba, Ibaraki 305-0006, Japan

Brief Career:
2004- Postdoctoral Fellow, Hokkaido University
2009- Postdoctoral Fellow, University of California, Santa Barbara
2010- Research Fellow, NIED
2014- Senior Researcher, NIED
2016- Chief Researcher, NIED

Selected Publications:
• Maeda, T. and H. Fujiwara, "Seismic hazard visualization from big simulation data: construction of a parallel distributed processing system for ground motion simulation data," J. Disaster Res., Vol.11, No.2, pp. 265-271, 2016.

Academic Societies & Scientific Organizations:
• Seismological Society of Japan (SSJ)
• Japan Association of Earthquake Engineering (JAEE)
• Architectural Institute of Japan (AIJ)
• American Geophysical Union (AGU)

Name:
Hiroyuki Fujiwara

Affiliation:
Manager, Integrated Research on Disaster Risk Reduction Division National Research Institute for Earth Science and Disaster Resilience (NIED)

Address:
3-1 Tennodai, Tsukuba, Ibaraki 305-0006, Japan

Brief Career:
1989- Researcher, NIED
2001- Head of Strong Motion Observation Network Laboratory, NIED
2006- Project Director, Disaster Prevention System Research Center, NIED
2011- Director, Department of Integrated Research on Disaster Prevention, NIED
2016- Manager, Integrated Research on Disaster Risk Reduction Division, NIED

Selected Publications:
• "Seismic Hazard Assessment for Japan: Reconsideration After the 2011 Tohoku Earthquake," J. Disaster Res., Vol.8, No.5, pp. 848-860, 2013.

Academic Societies & Scientific Organizations:
• Seismological Society of Japan (SSJ)
• Japan Association for Earthquake Engineering (JAEE)

Name:
Toshihiko Hayakawa

Affiliation:
Group Manager, Mitsubishi Space Software Co., Ltd.

Address:
1-6-1 Takezono, Tsukuba, Ibaraki 305-0032, Japan

Brief Career:
1998- Engineer, Mitsubishi Space Software
2014- Group Manager, Mitsubishi Space Software

Selected Publications:
• T. Hayakawa, T. Furumura, and Y. Yamanaka, "Simulation of strong ground motions caused by the 2004 off the Kii Peninsula earthquakes," Earth Planets Space, Vol.57, pp. 191-196, 2005.
• T. Furumura, T. Hayakawa, M. Nakamura, K. Koketsu, and T. Baba, "Development of long-period ground motions from the Nankai Trough, Japan, earthquakes: Observations and computer simulation of the 1944 Tonankai (Mw8.1) and the 2004 SE Off-Kii Peninsula (Mw7) Earthquakes," Pure Appl. Geophys., Vol.165, pp. 585-607, 2008.

Academic Societies & Scientific Organizations:
• Japan Geoscience Union (JpGU)
• Seismological Society of Japan (SSJ)

Name:
Satsuki Shimono

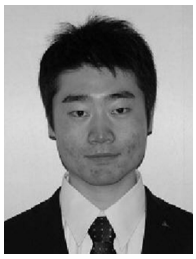
Affiliation:
Engineer, Mitsubishi Space Software Co., Ltd.

Address:
1-6-1 Takezono, Tsukuba, Ibaraki 305-0032, Japan

Brief Career:
2005 Joined Mitsubishi Space Software Co., Ltd.
2009-2012 Temporarily Transferred to ANET Co., Ltd.
2014- Professional Engineer (Applied Science)

Selected Publications:
• S. Shimono and T. Chiba, "Numerical solutions of inflating higher dimensional global defects," Phys. Rev. D, 71:084002, 2005.

Academic Societies & Scientific Organizations:
• Seismological Society of Japan (SSJ)
• Japan Geoscience Union (JpGU)
• Institution of Professional Engineers, Japan (IPEJ)



Name:
Sho Akagi

Affiliation:
Engineer, Mitsubishi Space Software Co., Ltd.

Address:
1-6-1 Takezono, Tsukuba, Ibaraki 305-0032, Japan

Brief Career:
2015- Joined Mitsubishi Space Software Co., Ltd.

Academic Societies & Scientific Organizations:
• Seismological Society of Japan (SSJ)
• Japan Geoscience Union (JpGU)