Discovering Seismic Interactions after the 2011 Tohoku Earthquake by Co-occurring Cluster Mining

Ken-ichi Fukui
The Institute of Scientific and Industrial Research, Osaka University
fukui@ai.sanken.osaka-u.ac.jp, http://www.ai.sanken.osaka-u.ac.jp

Daiki Inaba†¹
(affiliation as previous author)

Masayuki Numao (affiliation as previous author)
numao@ai.sanken.osaka-u.ac.jp

keywords: knowledge discovery, data mining, earthquake co-occurrence, hierarchical clustering, seismic interaction

Summary
In this study, we extract earthquake co-occurrence patterns for investigating mechanical interactions in the affected areas. To extract seismic patterns, both co-occurrence among seismic events in the event sequence and distances between the hypocenters to find hot spots must be considered. Most previous researches, however, have considered only one of these aspects. In contrast, we utilized co-occurring cluster mining to extract seismic patterns by considering both co-occurrence in a sequence and distance between hypocenters. Then, we acquired affected areas and relationships between the co-occurrence patterns and focal mechanisms from the 2011–2012 hypocenter catalog. Some results were consistent with seismological literature. The results include highly affected areas that may indicate asperity, and change of focal mechanisms before and after the Tohoku Earthquake.

1. Introduction
Earthquakes are one of the most devastating natural disasters, and seismologists have been studying ways to predict their occurrence in an attempt to minimize casualties and damage [Kanamori 03]. In seismology, a Bayesian statistics approach [Ogata 06] is commonly used to analyze the periodicity of earthquake occurrence and to make long-term earthquake forecasts. In Bayesian statistics, probabilistic models are applied for particular events or areas. However, short-term prediction has never been successful, since plate movement and the resulting earthquakes involve chaotic and complex processes [Geller 97]. Consequently, recent seismological research has shifted from earthquake prediction back to investigating the physical mechanisms of earthquakes [Faulkner 10].

In this study, we extract earthquake co-occurrence patterns among areas afflicted by the 2011 Tohoku Earthquake. A co-occurrence pattern consists of seismic events in different geographical areas, but frequently occurring with near time (Figure 1). In Figure 1, earthquake A₁ and A₂, and B₁ and B₂ are geographically close; moreover, A₁ and B₁, and A₂ and B₂ occur with near time. The purpose of this study is not earthquake prediction, but to advance the understanding of the mechanical interactions of seismic induced activity. Most research has investigated limited geographical areas or isolated seismic events, which were selected on the basis of particular seismologists’ knowledge; thus, little understanding of the interaction, or induced activity among seismic events or areas has been gained. Therefore, interesting and important areas of investigation may not have been identified.

Numerous studies utilize data mining for the analysis of seismic activities, including density-based clustering [Lei 10] and fuzzy clustering [Ansari 09], which are used to find earthquake hot spots. Lee et al. used quantitative

†¹ He is currently working at Ricoh Company, Ltd
association rule mining and found relationships between features such as depth and magnitude, and location and frequency [Lee 09]. Martínez-Álvarez et al. utilized a quantitative association rule and regression to investigate earthquake prediction in a specific area [Martínez-Álvarez 11].

For our objective, we need to consider the sequence of co-occurrence among seismic events in addition to the distance between hypocenters, within each area. Our purpose cannot be achieved by a straightforward approach, i.e., first clustering hypocenters to find hot spots and then extracting frequent patterns with clusters as items. The co-occurrence among multi-dimensional events in the sequence is not considered at the clustering step; the clusters may include events that are not actually related to the co-occurrence, and vice versa. As a result, co-occurrence patterns of earthquakes will not be extracted appropriately.

Apart from Earthquake application, various works have been done in spatio-temporal data mining. Especially, space-time scan statistics [Kulldorff 01] detects outbreak regions (clusters) in certain period based on statistical tests, for example detecting region(s) and period(s) of infection disease. However, space-time scan statistics does not extract co-occur of different regions, the purpose is to detect single region and its period.

To the best of our knowledge, Ohsawa’s study [Ohsawa 02] is the only one that has extracted co-occurrence frequent patterns of earthquakes among active faults using KeyGraph, which was originally designed as a keyword-extraction method. KeyGraph extracts a graph structure with high-frequently co-occurring events which is called island, as well as relatively low-frequency events but highly co-occur with events in island which is called bridge. They applied KeyGraph to earthquake events and successfully discovered some co-occurrence between active faults in Japan between 1985 and 1992. However, originally the KeyGraph is for categorical variables such as words, therefore in [Ohsawa 02] every earthquake event is re-assigned to the nearest active fault rather than using position coordinate of hypocenters, then co-occurrence was analyzed based on active faults. Therefore, KeyGraph cannot be applied to off-the-coast earthquakes such as Tohoku Earthquake on which this work focuses.

To compensate for the problems associated with the two-step method, Inaba et al. [Inaba 12a, Inaba 12b] proposed a unique method called co-occurring cluster mining (CCM), which can determine a range of clusters adapted by the co-occurrence of events in a sequence. Inaba et al. reported that CCM has a capability to extract accurate co-occurring clusters than the straightforward two-step method in terms of F-measure. Then, they applied CCM to extract damage patterns from acoustic emission events on a solid oxide fuel cell. Like a fuel cell, the Earth’s inner crust is mainly composed of solid state matter, so the application of CCM might reveal mechanical interactions. We applied CCM to the 2011-2012 hypocenter catalog data for Japan, which revealed affected areas. The extracted patterns present seismic interactions for subduction zone earthquakes. By validation from seismological literature, the results included highly affected areas that may indicate asperities, and a change of focal mechanisms both before and after the Tohoku Earthquake.

2. Co-occurring Cluster Mining

The CCM has properties of both event clustering, and association rule or frequent pattern mining in a sequence of the events. Clustering [Xu 08] tries to find groups based on similarity of the events in the data space, while association rule mining [Agrawal 93] tries to extract frequently appearing sets of items, referred to as events in this paper, in the series of items. Meanwhile, the CCM extracts and lists pairs of clusters that have high intra-cluster density and simultaneously high inter-cluster co-occurrence in the sequence.

2.1 Problem Statement

In this section, we define the characteristics of data that our current work focuses on, and then define requirements of the co-occurrence pattern.

Definition 1 (event sequence). Suppose data set \( D \) with \( N \) numerical event data points \( x_i = (x_{i,1}, \ldots, x_{i,v}) \), \( (i = 1, \ldots, N) \) in \( v \)-dimensional space are obtained in order \( x_1 \prec \cdots \prec x_N \).

Definition 2 (segment). Suppose an event sequence is divided into segments. More specifically, let data set \( D \) be denoted by

\[
D = [x_1, \cdots, x_i][x_{i+1}, \cdots, x_j][x_{j+1}, \cdots, x_k][x_{k+1}, \cdots, x_N],
\]

where \( i < j < k < N \) and “\([\]\)” refers to a segment (similar to market basket analysis).

The segments above are measured in minutes, days, and so on; furthermore, the length of these segments need not be regular. Given the above, the extracted co-occurrence patterns must satisfy the following three requirements:

Requirement 1 (co-occurrence). For two sets composed of events \( A, B \subset D \) (\( A \cap B = \emptyset \)), the co-occurring ratio of \( A \) and \( B \) must be high. Co-occurrence can be evaluated by the Jaccard coefficient by counting
the number of segments that contain A and B, and A or B.

**Requirement 2 (frequency).** The number of times in which A and B co-occur in an event sequence must be high. Such occurrence frequency can be evaluated, for example, by the support score by counting the number of segments that contain A and B.

**Requirement 3 (similarity).** For two event sets A and B, events in A must be similar and events in B must be similar as well. The within-cluster similarity can be evaluated for example by the sum of squares within clusters (SSW), or the average distance among all data points in a cluster, where the distance is measured by Euclidean distance of latitude and longitude between hypocenters in this work.

Requirements 1 and 2 are derived from frequent pattern mining between event sets (clusters), whereas requirement 3 is derived from the clustering of events. Requirements 1 and 3 are evaluated by functions and 2 is satisfied by a threshold, as explained in the next subsection.

Given the above, we define the co-occurring cluster and co-occurrence pattern as follows:

**Definition 3 (co-occurring cluster).** If two sets A, B ⊂ D satisfy the above three requirements, set A is a co-occurring cluster of B and vice versa.

**Definition 4 (co-occurrence pattern).** With co-occurring clusters A and B, P(A, B) = {A, B | A ∩ B = ∅} is called a co-occurrence pattern.

### 2.2 Evaluation Function

In this section, we define an evaluation function to search for co-occurrence patterns defined in the above section. We search pairs of clusters A, B ⊂ D that maximize the following evaluation function:

\[
\mathcal{L}(A, B) = \sqrt{\mathcal{F}(A, B) \cdot \mathcal{G}(A, B)}.
\]  

Function \(\mathcal{F}(A, B)\) evaluates the co-occurrence ratio for requirement 1. The higher the \(\mathcal{F}(A, B)\) value is, the higher the co-occurrence ratio. Note that because requirement 1 denotes the co-occurrence among many separated segments, co-occurrence in the short and sequential period must be excluded. Therefore, even if events from A and B co-occur several times in the same segment, this is considered only once. Function \(\mathcal{G}(A, B)\) denotes similarity within a cluster for requirement 3. The higher the \(\mathcal{G}(A, B)\) value is, the more dense the clusters are. Contents of \(\mathcal{F}\) and \(\mathcal{G}\), in case of the earthquake application, are described in sec. 3.2.

Evaluation function (1) is defined as the product of \(\mathcal{F}\) and \(\mathcal{G}\) to simultaneously satisfy the requirements of co-occurrence and similarity. By normalizing \(\mathcal{F}, \mathcal{G} \in [0,1]\), both requirements of co-occurrence and similarity can be equally evaluated. In addition, requirement 2 for occurrence frequency can be satisfied using minimum support \(\text{Supp}_{\text{min}}\) as a threshold.

### 2.3 Algorithm

The pseudocode and conceptual diagram of the CCM algorithm are presented in Algorithm 1 and Figure 2, respectively. First, the algorithm generates possible sub-clusters from the dendrogram obtained by aggregative hierarchical clustering (AHC) in the data space. All combinations of sub-clusters can be candidate patterns (Step 1, Fig. 2(a)). Second, the algorithm evaluates each candidate pattern via function \(\mathcal{L}\), which evaluates the co-occurrence degree of sub-clusters in the event sequence and similarity within each sub-cluster in the data space. If the evaluation value exceeds the minimum thresholds \(\mathcal{L}_{\text{min}}\) and \(\text{Supp}_{\text{min}}\), then these patterns are added to an output pattern list \(\mathcal{P}\) (Step 2, Fig. 2(b)). Here the definition of support score (l. 6 of Algorithm 1)— i.e., \(\text{Supp}(H_i, H_j) = \text{count}(H_i \cap H_j) / S\), where \(\text{count}(H_i)\) denotes the number of segments that contain event(s) with cluster label \(H_i\), and \(S\) is the total number of segments. Third, the algorithm checks the in-

---

**Algorithm 1 Co-occurring cluster mining algorithm**

**Input:** multi-dimensional event sequence with segments:

\[
D = [x_1, \ldots, x_i; x_{i+1}, \ldots, x_j; \ldots; x_k]_N^k
\]

dendrogram by hierarchical clustering from \(D' = \{x_k\}_k=1^N\) \n
\(\mathcal{H}\)

minimum evaluation function value: \(\mathcal{L}_{\text{min}}\)

minimum support value: \(\text{Supp}_{\text{min}}\)

**Output:** Co-occurrence patterns:

\(P = \{P_k | A \cap B = \emptyset, A, B \subseteq D\}\)

1. \#Step 1: Generate sub-clusters
2. Generate possible sub-clusters from \(\mathcal{H}\):
   \(H_1, H_2, \ldots, H_N \subseteq \mathcal{H}\)
3. \#Step 2: Evaluate candidate patterns
4. Initialize \(k \leftarrow 0\);
5. for all combinations of \(H\) do
6. if \(H_i \cap H_j = \emptyset\) and \(\mathcal{L}(H_i, H_j) > \mathcal{L}_{\text{min}}\) and 
   \(\text{Supp}(H_i, H_j) > \text{Supp}_{\text{min}}\) then
   7. \(P_k(A, B) \leftarrow \{H_i, H_j\}\);
   8. \(k \leftarrow k + 1\);
6. end if
9. end for
10. \#Step 3: Eliminate co-occurrence patterns with inclusion relation
11. for all combinations of \(P\) do
12. if \(P_i \cap P_m \neq \emptyset\) then
13. Remove \(P_i\) from \(P\) such that
   \(i = \arg \min \{\mathcal{L}(P_i), \mathcal{L}(P_m)\}\);
15. end if
16. end for
2.4 Toy Example

Suppose ten events in sequence with three segments are obtained: $D = \{x_{1}, x_{2}, x_{3}\} \cup \{x_{4}, x_{5}, x_{6}\} \cup \{x_{7}, x_{8}, x_{9}, x_{10}\}$. Then suppose a dendrogram of clustering merge process by AHC is obtained as shown in Figure 3. The clustering is produced based only on $D' = \{x_{k}\}_{k=1}^{10}$, disregarding order of the events, for example $x$ has latitude and longitude of an earthquake hypocenter in this work.

In step 1, CCM generates possible sub-trees (clusters) from the dendrogram. Then in step 2, evaluate all combinations of sub-trees as candidates of co-occurrence pattern. Here, focus attention on one of the candidates, a cluster pair $A$ and $B$ in Figure 3. Cluster elements are $x_{1}, x_{7}, x_{10} \in A$ and $x_{2}, x_{8} \in B$. With these cluster labels, the event sequence can be $[x_{1}^{(A)}, x_{2}^{(B)}, x_{3}] [x_{4}, x_{5}, x_{6}, x_{7}, x_{8}, x_{9}, x_{10}]$. The events in cluster $A$ and $B$ frequently co-occur in the sequence. If evaluation score $L(A, B)$ exceeds the threshold $L_{min}$, and also co-occurrence frequency exceeds the minimum support for example one over three segments, then $P_{1}(A, B)$ is a co-occurrence pattern.

However, if $P_{2}(A, B')$ also satisfies the conditions, similar patterns can be extracted. In order to eliminate similar patterns, hierarchical property of AHC is utilized in step 3. Compare evaluation scores $L(A, B)$ and $L(A, B')$, then remove the lower one. However, $P_{1}(A, B)$ and $P_{3}(B, C)$ should be different patterns. Elimination of pattern is checked if both co-occurring clusters in two patterns are in inclusion between the patterns, and removes patterns from $\mathcal{P}$ that have a lower evaluation score (Step 3, Fig. 2(c)).

In the algorithm description, $P_{1} \cap P_{m}$ (l. 13 of Algorithm 1) means $A_{1} \cap A_{m}, A_{1} \cap B_{m}, B_{1} \cap A_{m}$, or $B_{1} \cap B_{m}$.

3. Application to the 2011 Tohoku Earthquake

3.1 Data Preprocessing

We applied CCM to the hypocenter catalog data recorded from January 1st, 2011 to December 31st, 2012, released by the JMA\(^{1}\). Each earthquake event $e_{i}$ is represented by a tuple $<x_{i}, t_{i}, M_{i}>$, where a vector $x_{i} = (lat_{i}, lon_{i})$ is latitude and longitude of the hypocenter, and $t_{i}$ and $M_{i}$ denote corresponding origin time and magnitude. We omitted events distant from Japan, i.e., only events between 23°N and 50°N, and between 129°E and 156°E, were used. The events within this region with a magnitude greater than four were used for the analysis. 5954 seismic events were recorded in that period; all these events are plotted in Figure 4.

Regarding AHC in CCM, we used latitude and longitude as the attributes for merging clusters, and Euclidean distance is used as the distance metric. We did not include depth because there were only minor differences in depth in the same areas. We empirically employed Ward’s

\(^{1}\) The data is distributed via Japan Meteorological Business Support Center: http://www.jmbsc.or.jp (in Japanese)
method by checking the dendrogram; e.g., a chaining effect and monotonicity.

Regarding segmentation of the seismic event sequence, much as in the fuel cell application [Inaba 12a, Inaba 12b], the key idea is based on Ohsawa’s study [Ohsawa 02]—i.e., the segment division was performed utilizing the magnitude of the seismic events. When a large energetic event occurs, the structure of the Earth’s inner crust changes, and the seismic process will transit to another condition; however, in the Tohoku Earthquake, quakes greater than M6.0 occurred near the main shock*2 at short time intervals. Therefore, the length of the segments decreased immediately after the main shock. We introduced time constraints for segment length to eliminate very short segments that contained only a few events in a segment.

On the basis of the above ideas, a seismic event sequence can be divided into segment $s = [x_s, \ldots, x_{s+l}]$ by the following two conditions:

$$M_{s+i} \leq M_\sigma \text{ and } M_{s+i} > M_\sigma \ (i = 0, \ldots, l-1),$$

$$t_{s+i} - t_s > t_\sigma,$$

where $M_i$ and $t_i$ denote corresponding magnitude and origin time of earthquake event $x_i$, respectively; $M_\sigma$ is set to M6.0, and $t_\sigma$ is 1 h. With these conditions, we obtained 91 segments, which is approximately the same condition as that seen in Ohsawa’s study.

### 3.2 Design of the Evaluation Function

We designed the evaluation function $\mathcal{L}$ as follows; $\mathcal{F}(A, B)$ is defined as:

$$\mathcal{F}(A, B) = \frac{1}{1 + \exp(-\alpha J(A, B) - 0.5)}, \quad (4)$$

$$J(A, B) = \frac{\text{count}(A \cap B)}{\text{count}(A \cup B)}., \quad (5)$$

where $\text{count}(X)$ is the number of segments that contain $X$. The Jaccard coefficient $J(A, B)$ is wrapped by a sigmoid function to sharpen around 0.5, where $\alpha$ is a sharpness parameter. The higher $\alpha$, the higher resolution around Jaccard coefficient 0.5 can be obtained.

While $\mathcal{G}(A, B)$ is defined by the following function:

$$\mathcal{G}(A, B) = \exp\left(-\frac{\text{SSW}(A)^2 + \text{SSW}(B)^2}{\sigma}\right). \quad (6)$$

where SSW indicates the sum of squares within clusters. Here, a Gaussian function is used to adjust the bias of SSW in a dataset, where a radius $\sigma$ is a parameter to control correction of the bias. Note that $A$ and $B$ are independently evaluated, i.e., $\mathcal{G}(A, B) = g(A)g(B)$, where $g(A) = \exp(-\text{SSW}(A)^2/\sigma)$.

Here, two parameters $\alpha$ and $\sigma$ are to adjust outputs of $\mathcal{F}$ and $\mathcal{G}$ so as to widely distribute in [0,1].

### 3.3 Parameter Study

First, we studied the effect of parameters; minimum evaluation score $\mathcal{L}_{\min}$, minimum support score $\text{Supp}_{\min}$, magnitude threshold $M$, a parameter $\alpha$ in $\mathcal{F}$, and $\sigma$ in $\mathcal{G}$. Figure 5 shows the change in one of the above parameters with base parameters as $\mathcal{L}_{\min} = 0.7$, $\text{Supp}_{\min} = 0.05$, $M = 6.0$, $\alpha = 10$, and $\sigma = 0.03$. Parameters $\mathcal{L}_{\min}$ and $\text{Supp}_{\min}$ are thresholds controlling the number of patterns to list, whereas the score of patterns, including non-listed patterns, are fixed. As $\mathcal{L}_{\min}$ or $\text{Supp}_{\min}$ decreases, the number of patterns increase exponentially (Figs. 5(a) and 5(b)). The other threshold, $M$, determines length and the number of segments (Fig. 5(c)). This parameter is somewhat sensitive, as the number of patterns are drastically increased when changing $M$ from 6.0 to 6.1. When increasing parameter $\alpha$, the number of patterns and the average $\mathcal{L}$ score increases, while the average $\mathcal{G}$ score decreases (Fig. 5(d)). This is because a higher $\alpha$ gives more weight to $\mathcal{F}$; accordingly this causes a relative decrement of $\mathcal{G}$. Finally, when increasing $\sigma$, the number of patterns increase, while the average $\mathcal{L}$ suddenly drops after 0.04 (Fig. 5(e)).

The base parameters were determined by trial and error, as the number of patterns should be lower than 30 to 40, so as to analyze the patterns manually, average $\mathcal{L}$ should be high and, in addition, $\mathcal{F}$ and $\mathcal{G}$ should be balanced. Empirically, we carried out about ten trials to determine the base parameters. Note that, although we cannot conclude these are the best parameters, the extracted patterns are stable around these parameters; i.e., similar patterns in terms of geographical plotting of the patterns can be obtained.
3.4 Extracted Patterns

The evaluation scores of the extracted patterns and the number of co-occurrences are listed in Table 1. The base parameters in the above mentioned section were used. We validated each extracted pattern by Fisher’s exact test with a cross-table that consists of the number of co-occurrences of two events in the same segment, from a pair of co-occurring clusters. This statistical association cannot be achieved by the Jaccard coefficient.

Some patterns, P₁, P₂, and P₂₇ are high in both \( \mathcal{F} \) and \( \mathcal{G} \). A few patterns, P₂₄ and P₂₈ have a high \( \mathcal{F} \) and relatively low \( \mathcal{G} \). The rest of the patterns are a relatively low \( \mathcal{F} \) and high \( \mathcal{G} \). Most of the numbers of co-occurrence are between 5 and 9, with only two patterns, P₃₀ and P₃₁, occurring approximately 30 times.

Figure 6 illustrates the representative extracted patterns. For example, P₅ in Fig. 6(a), high score of \( \mathcal{G} \) indicates that a cluster of off the coast of Iwate Pref. (upper cluster in the figure) has closer hypocenters, and same as in a cluster of off the coast of Miyagi Pref. (lower cluster). While high score of \( \mathcal{F} \) indicates that earthquakes in the cluster of off the coast of Iwate Pref. and that of off the coast of

Table 1  Scores of the extracted patterns from 2011-2012 earthquakes

<table>
<thead>
<tr>
<th>ID</th>
<th>( \mathcal{L} )</th>
<th>( \mathcal{F} )</th>
<th>( \mathcal{G} )</th>
<th># co-occurrence</th>
<th>ID</th>
<th>( \mathcal{L} )</th>
<th>( \mathcal{F} )</th>
<th>( \mathcal{G} )</th>
<th># co-occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>P₁</td>
<td>0.9421</td>
<td>0.8949</td>
<td>0.9918</td>
<td>5</td>
<td>P₁₆</td>
<td>0.7891</td>
<td>0.6354</td>
<td>0.9800</td>
<td>5</td>
</tr>
<tr>
<td>P₂</td>
<td>0.9424</td>
<td>0.8949</td>
<td>0.9923</td>
<td>5</td>
<td>P₁₇</td>
<td>0.7891</td>
<td>0.6354</td>
<td>0.9800</td>
<td>5</td>
</tr>
<tr>
<td>P₃</td>
<td>0.7923</td>
<td>0.6354</td>
<td>0.9879</td>
<td>5</td>
<td>P₁₈</td>
<td>0.7625</td>
<td>0.6354</td>
<td>0.9150</td>
<td>5</td>
</tr>
<tr>
<td>P₄</td>
<td>0.7044</td>
<td>0.5000</td>
<td>0.9924</td>
<td>5</td>
<td>P₁₉</td>
<td>0.7384</td>
<td>0.6117</td>
<td>0.8914</td>
<td>6</td>
</tr>
<tr>
<td>P₅</td>
<td>0.7037</td>
<td>0.5000</td>
<td>0.9906</td>
<td>5</td>
<td>P₂₀</td>
<td>0.7947</td>
<td>0.7310</td>
<td>0.8639</td>
<td>6</td>
</tr>
<tr>
<td>P₆</td>
<td>0.7050</td>
<td>0.5000</td>
<td>0.9941</td>
<td>5</td>
<td>P₂₁</td>
<td>0.7376</td>
<td>0.6117</td>
<td>0.8895</td>
<td>6</td>
</tr>
<tr>
<td>P₇</td>
<td>0.7282</td>
<td>0.6354</td>
<td>0.9892</td>
<td>5</td>
<td>P₂₂</td>
<td>0.7165</td>
<td>0.5949</td>
<td>0.8630</td>
<td>7</td>
</tr>
<tr>
<td>P₈</td>
<td>0.7831</td>
<td>0.6354</td>
<td>0.9651</td>
<td>5</td>
<td>P₂₃</td>
<td>0.7192</td>
<td>0.6354</td>
<td>0.8142</td>
<td>5</td>
</tr>
<tr>
<td>P₉</td>
<td>0.7850</td>
<td>0.6354</td>
<td>0.9698</td>
<td>5</td>
<td>P₂₄</td>
<td>0.8200</td>
<td>0.9241</td>
<td>0.7277</td>
<td>6</td>
</tr>
<tr>
<td>P₁₀</td>
<td>0.7905</td>
<td>0.6354</td>
<td>0.9835</td>
<td>5</td>
<td>P₂₅</td>
<td>0.8367</td>
<td>0.7773</td>
<td>0.9007</td>
<td>5</td>
</tr>
<tr>
<td>P₁₁</td>
<td>0.7769</td>
<td>0.6354</td>
<td>0.9499</td>
<td>5</td>
<td>P₂₆</td>
<td>0.7210</td>
<td>0.6354</td>
<td>0.8183</td>
<td>5</td>
</tr>
<tr>
<td>P₁₂</td>
<td>0.7749</td>
<td>0.6117</td>
<td>0.9818</td>
<td>6</td>
<td>P₂₇</td>
<td>0.8602</td>
<td>0.8949</td>
<td>0.8268</td>
<td>5</td>
</tr>
<tr>
<td>P₁₃</td>
<td>0.7875</td>
<td>0.6354</td>
<td>0.9761</td>
<td>5</td>
<td>P₂₈</td>
<td>0.7763</td>
<td>0.8949</td>
<td>0.6734</td>
<td>5</td>
</tr>
<tr>
<td>P₁₄</td>
<td>0.7652</td>
<td>0.6354</td>
<td>0.9217</td>
<td>5</td>
<td>P₂₉</td>
<td>0.7363</td>
<td>0.6513</td>
<td>0.8324</td>
<td>9</td>
</tr>
<tr>
<td>P₁₅</td>
<td>0.7742</td>
<td>0.6354</td>
<td>0.9433</td>
<td>5</td>
<td>P₃₀</td>
<td>0.7324</td>
<td>0.6651</td>
<td>0.8064</td>
<td>29</td>
</tr>
<tr>
<td>P₃₁</td>
<td>0.7255</td>
<td>0.7523</td>
<td>0.6997</td>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 6 illustrates the representative extracted patterns. For example, P₅ in Fig. 6(a), high score of \( \mathcal{G} \) indicates that a cluster of off the coast of Iwate Pref. (upper cluster in the figure) has closer hypocenters, and same as in a cluster of off the coast of Miyagi Pref. (lower cluster). While high score of \( \mathcal{F} \) indicates that earthquakes in the cluster of off the coast of Iwate Pref. and that of off the coast of
Miyagi Pref. frequently co-occurred. Some co-occurring clusters could be geographically close, or inclusion relations of AHC, or identical clusters. Most of the patterns are located off the coast of the Tohoku area or the boundary between the North American and Pacific Plates, even though earthquakes also occurred in mid-island, as shown in Figure 4. This is because the generating mechanisms are different; i.e., due to subduction of the plate off-coast, and due to active mid-island faults. Active faults can be isolated, since every active fault has a different force direction, whereas subduction of the plate affects a wide range with similar force that may induce interactions on the seashore.

Some co-occurrence patterns can be connected with such close or inclusion co-occurring clusters, which suggest a highly affected area, in the sense that the area has interactions between several other areas. Three highly affected areas were obtained (Figs. 6(a) to 6(c)), where all the areas were located off the coast of Tohoku and near the main shock of the Tohoku Earthquake. From Fig. 6(d), pat-
terns $P_2$, $P_{16}$, and $P_{17}$ are located near Iwaki City, where swarm earthquakes were reported (the co-occurring clusters are close). Some more distant patterns were obtained, e.g., $P_{23}$ in Fig. 6(e) and $P_{30}$ in Fig. 6(f), where $P_{30}$ has the largest number of co-occurrences within a relatively wide area.

We compared the extracted patterns with those prior to the Tohoku Earthquake, from Oct. 1997 to Feb. 2011, with 12062 seismic events whose magnitudes were greater than M4.0. Most of the high score patterns were located near Miyake Island, south of Tokyo, as in Fig. 7(a). A remarkable historical event here is the large eruption on Miyake Island on June 26, 2000, which induced swarm earthquakes in the west of the island, resulting in close co-occurring clusters in this region.

The other type is the pattern in the plate boundary northeast of Hokkaido (Fig. 7(b)), where similar patterns were also obtained after the Tohoku Earthquake (Fig. 6(e)). In fact, the boundary between the North American and Pacific Plates in this region has been stuck strongly, according to GPS observation in 1997–2001 (Figure 8). Hence, we inferred that the internal energy had been accumulating, and eventually a large number of interactions between earthquakes were produced in this region.

The last ones are distant patterns from $P_{25}$ to $P_{27}$ between the plate boundary northeast of Hokkaido and mid-island in the Tokyo area (Fig. 7(c)). Although these patterns are very far apart, this interaction cannot be ignored; since these co-occurring clusters are both still on the North American Plate, interactions may happen. Furthermore, these patterns have a high co-occurrence with $G > 0.7$ and occur more than 60 times.

Obviously, dense co-occurring clusters are only near Miyake Island before the Tohoku Earthquake (Fig. 7(a)), whereas various dense co-occurring clusters are obtained after it (Figure 6). This result may be due to interactions induced by the huge amount of energy released by the Tohoku Earthquake.

3.5 Relation to Geographical Measurement

Figure 9 shows the correspondence between the estimated highly affected areas with geographical measurements. The three highly affected areas (Figs. 6(a) to 6(c)) are illustrated on slip distribution maps by Miyazawa et al.[Miyazaki 11].

From the figures, area (1) has a relation to the transient slip, area (2) has relations to both the transient and the aftershock slip, and area (3) is related to the coseismic slip, respectively. It may be that the highly affected areas indicate asperities, including hidden ones, before the Tohoku Earthquake. Asperity is a seismological term describing a spot near a plate subduction boundary that is strongly adhered, but suddenly slips when it reaches its durability limit. Moreover, seismology literature proved the existence of interactions between asperities using a 3-D simulation of a subduction plate boundary[Ariyoshi 09].

3.6 Co-occurrence Relationships of Focal Mechanisms

Finally, this section discusses the co-occurrence relations of focal mechanisms. The focal mechanism describes the direction of the force (pressure or tension) and the slip type of the fault (normal or reverse fault). We used centroid moment tensor (CMT) solutions, released by JMA, which includes 1032 CMT solutions for seismic events in 2011*4, and 1103 solutions from Oct. 1997 to Feb. 2011.

We then calculated conditional probabilities $Pr(M_a|M_b)$, given a mechanism type $M_b$ and its belonging co-occurrence patterns, a probability containing $M_a$ in the other co-occurring cluster can be defined by the following equations:

$$Pr(M_a|M_b) = \frac{\sum_{x_i} \sum_{x_j \in P} l_{i,j}(a,b)}{\sum_{k} \sum_{x_i} \sum_{x_j \in P} l_{i,j}(a,k)}.$$  

$$l_{i,j}(a,b) = \begin{cases} 1 & \text{if } CMT(x_i) = M_a \text{ and } CMT(x_j) = M_b, \\ 0 & \text{otherwise,} \end{cases}$$  

where $CMT(x_i)$ denotes a type of CMT solution for an event $x_i$.

Table 2 shows the conditional probabilities before and after the Tohoku Earthquake. Before the Tohoku Earthquake, from Oct. 1997

*3 Several major recording criteria have been changed since Oct. 1997.

*4 The CMT solutions for 2012 have not been released yet at Sep. 2013.
quake (left side of slash), CMT types with a probability greater than 0.2 are mainly R-DS; however, it clearly changes to N-DS after the Tohoku Earthquake (right side of slash). Seismological literature reported that the stress field of the Earth’s crust has been changed from reverse faults to normal faults before and after the Tohoku Earthquake [Yoshida 12]. Our result is consistent with the literature; in addition, the result shows interactions between focal mechanisms.

### 4. Conclusion

We utilized a co-occurring cluster mining (CCM) method to discover seismic interactions after the 2011 Tohoku Earthquake. The CCM algorithm is a unique method that has both properties of clustering in the data space and simultaneously mining frequent co-occurring clusters. The results were partly consistent with the historical events and seismology literature, which indicate highly affected areas that may be related to asperity and change in the stress field before and after the Tohoku Earthquake. Though the CCM succeeded to extract co-occurrence patterns for subduction zone earthquakes, availability for other types of earthquake such as intra plate earthquake and volcanic earthquake should be examined.

In this work, we validated availability of CCM method for earthquake co-occurrence analysis by collating with seismology literature. However, by analyzing the results more in detail, not only based on existing knowledge, our approach is expected to discover novel but reasonable knowledge on seismic interactions, and to provide hypotheses toward revealing earthquake occurrence mechanism. That is, this work also indicates one of the directions on earthquake research for data-intensive scientific discovery[Hey 09].

### Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 24650068 and by the Management Expenses Grants for National Universities Corporations from MEXT of Japan.

![References](Agrawal 93) Agrawal, R., Imielinski, T., and Swami, A.: Mining Association Rules Between Sets of Items in Large Databases, in Proceedings of the 1993 ACM SIGMOD International Conference on
Management of Data, SIGMOD ’93, pp. 207–216, ACM (1993)