Hierarchical Cluster Analysis of Dense GNSS Data and Interpretation of Cluster Characteristics

（高密度 GNSS データの階層型クラスター解析とクラスターの特徴の解釈）

Atsushi Takahashi

1. Introduction

Spatial distribution of crustal deformation is a fundamental problem for understanding tectonics in plate convergence zones. One of the most effective observation for measurement of crustal deformation is GNSS. It is widely recognized that crustal deformation is localized in a narrow zone owing to dense GNSS observations. Identifying and assessing such an area is critical for both in scientific perspective and hazard assessment. For these purposes, block models have been used to obtain fault characteristics such as slip rate or its deficit. However, block models have an important shortcoming. How to design tectonic blocks and surrounding faults depends on researchers and it tends to be subjective. It is critical as the observation data are inverted to determine rotations of tectonic blocks and slip deficits on the faults.

Recently, Simpson et al. (2012) proposed to apply a cluster analysis to GNSS velocity data for identifying tectonic blocks. The similar research using a different clustering algorithm (Savage and Simpson, 2013) also succeeded in identifying tectonic block structures. Despite these innovative studies, a few case studies have done so far. Furthermore, the general behavior of the method has not yet been examined.

To investigate the general behavior of the method, I adopt the hierarchical agglomerative clustering algorithm and develop a statistical scheme that enables us to visualize and quantify cluster ambiguity associated with observation noise contamination. Then I apply the hierarchical clustering with developed statistical scheme to GNSS velocity fields in Taiwan, Kyushu, and New Zealand. Finally, I discuss advantages and limitations of the method throughout the case studies.

2. Methods

Following the method of Simpson et al. (2012), I used the hierarchical agglomerative clustering algorithm to classify data in this study. Observed GNSS horizontal velocity data are projected in the 2-dimensional velocity space (Fig. 1a). Then, the nearest pair of data is searched and merged into a temporal representative point introduced at their centroid (Fig. 1b). In order to describe cluster linkages in a hierarchical representation, a tree space is introduced and the pair of data are bridged in the tree space (Fig. 1c). Note that the bridge girder height is given by the distance between the merged data pair. This operation is repeated until a single cluster remains in the velocity space (Fig. 1d). Through the linkage procedures, the data are organized into a hierarchical tree, called dendrogram. Clusters are obtained by cutting a tree at a certain cluster hierarchy. If I cut a tree at a higher cluster hierarchy, smaller number of clusters are obtained (Fig. 1e). On the other hand, larger number of clusters are derived if a cut is at lower cluster hierarchy (Fig. 1f). The higher the cluster hierarchy is, the more important velocity gaps are.

The clustering yields cluster boundaries for each given number of clusters, but whether the boundaries are trustworthy or not hasn't been examined from a statistical viewpoint. I introduce a statistical scheme to assess cluster stability. To be concrete, hundreds of synthetic datasets are prepared by composition of the secular velocities and noise. Then, for each given number of clusters, the synthetic datasets are clustered using the hierarchical clustering. Trust worthy clusters are replicated through the clustering trials, but the same is not true for doubtful clusters. A matrix representation whose elements are pairwise cluster reproducibility enables us to confirm cluster stability intuitively. In addition to the matrix representation, I introduce information entropy (Shannon, 1948) to quantify instability of cluster members. The introduction of information entropy also enables us to evaluate cluster instability in the map and velocity spaces at a glance.

3. A Case Study in Taiwan

The first case study is the application to GNSS velocity data in Taiwan. The Luzon volcanic arc on the Philippine Sea Plate is colliding against the Eurasian continental margin. As the persistent collision formed the Taiwan Island, basic tectonic framework is collision tectonics. I identified the 4 major clusters in the dendrogram space and their boundary appeared along significant geological boundaries. For example, the primary cluster boundary appeared along the Longitudinal Valley Fault Zone, which binds the Eurasian Plate and Philippine Sea Plate. Some subclusters appeared locally associated with active faults in east and southwest Taiwan (Fig. 2a). The correspondence of clusters of a high hierarchy to important tectonic boundaries and clusters of a lower hierarchy to local tectonic boundaries suggests that the hierarchical description is informative for interpreting tectonic implications of the cluster separations. Most of cluster boundaries are stable, but a boundary between the Coastal Plain and the Western Foothills was illuminated with high entropy (Fig. 2b). This feature is interpreted as that low angled thrust faults in the eastern edge of the Coastal Plain make velocity gap smoother.

4. A Case Study in Kyushu

The second case study is the investigation of the GNSS velocity field in Kyushu. I analyzed the GEONET data from March 2006 to August 2009 for avoiding temporal events such as slow slips events (Yoshioka et al., 2015) and large postseismic deformation of the Tohoku earthquake. I identified 3 major clusters in Kyushu based on dendrogram. One cluster occupies the region around the Bungo Channel. This may reflect locking between overlying and subducting plates beneath the Bungo Channel. The secondary cluster appeared in southernmost Kyushu. The separation is associated with a block motion of the Northern Ryukyu Arc. The leftover covers most of Kyushu, which had 4 subclusters at lower hierarchy. The stability test indicates that the boundary between them are diffused, not block like boundary. Though subclusters were indicated, I need more advanced method of noise analysis to clearly resolve the cluster boundaries. Comparison of previous study indicated that the cluster boundary may have experienced temporal changes.

5. A Case Study in New Zealand

The third case study of velocity field in New Zealand gives a different view from that for Taiwan. I analyzed GNSS data published by Beavan et al. (2016). The observation had conducted during the period from 1995 to 2013. The primary boundary appeared along the Alpine Fault, but it included southern part of the North Island. This may reflect interplate coupling beneath the North Island. The secondary separation appeared along the North Island Dextral Fault Belt (NIDFB), but not appeared along the Taupo Volcanic Zone. The third separation splits the Chatham Islands. The cluster separations at lower hierarchy indicate clusters parallel to Alpine Fault. The stability assessment indicates that cluster boundary is not clearly segmented. None of the surface traces of the Marlborough fault zone clearly coincide with any cluster boundary.

6. Discussion

An empirical relationship is suggested from the case study in Taiwan that clusters of the higher hierarchies correspond to significant tectonic sources, while those of the lower hierarchies are associated with relatively shallower tectonic sources such as active faults. The two case studies support the former part of the empirical relationships but not necessarily for the latter part. The latter part is reflecting regional rheological properties. If brittle lithosphere is stronger, block like deformation prevails, and the lower cluster hierarchy reflect minor active faults. If ductile upper lithosphere is stronger, velocity field at ground surface reflect ductile flow at depth, and surface tectonic structures will be subsidiary ones. In this case, correspondence between cluster boundaries and tectonic structures become poor.

The advantages and limitations of the clustering are also revealed through the case studies. Comparison with block models of previous studies in the three areas indicates that boundaries are obtained objectively because the clustering is free from any knowledge of tectonics. In addition, the stability assessment method illustrated significance of the tectonic boundaries. A hierarchical representation of the cluster linkage provides an important support for interpretation of regional tectonics. On the other hand, a block model prevails the clustering approach in yielding fault properties such as slip deficit and its spatial distribution. The respective pros and cons of the method should be complementary so that they should be combined in the future.

The curvature of the Earth affects results if analytical area is broad or close to a rotation pole. I perform a comparison between the Euler pole clustering method proposed by Savage and Wells (2015) and the hierarchical clustering using the same dataset in New Zealand. The study indicates that the split of the Chatham Island in the conventional hierarchical clustering is due to the curvature of the Earth.

Clustering results can be changed according to observation period. The slow slip events on a plate interface can disturb the GNSS velocity field, and can change cluster boundary temporally. This prevent us from identifying tectonic boundaries, but this feature is applicable for identifying slow slip events.

It is always argued which approach is better for clustering, a top-down approach or a bottom-up approach. A k-medoid clustering of GNSS data in Taiwan demonstrated that cluster boundaries in larger number of clusters do not coincide with those of smaller number of clusters. In other words, the cluster boundaries change according to the given number of clusters. On the other hand, as demonstrated in the case study of Taiwan, the hierarchical clustering provides a useful hierarchical representation among clusters, which contributes to the cluster interpretability. Based on the consistent cluster partitions, the hierarchical representation of clusters has more advantages for tectonic implications than those of the top-down approach will provide.

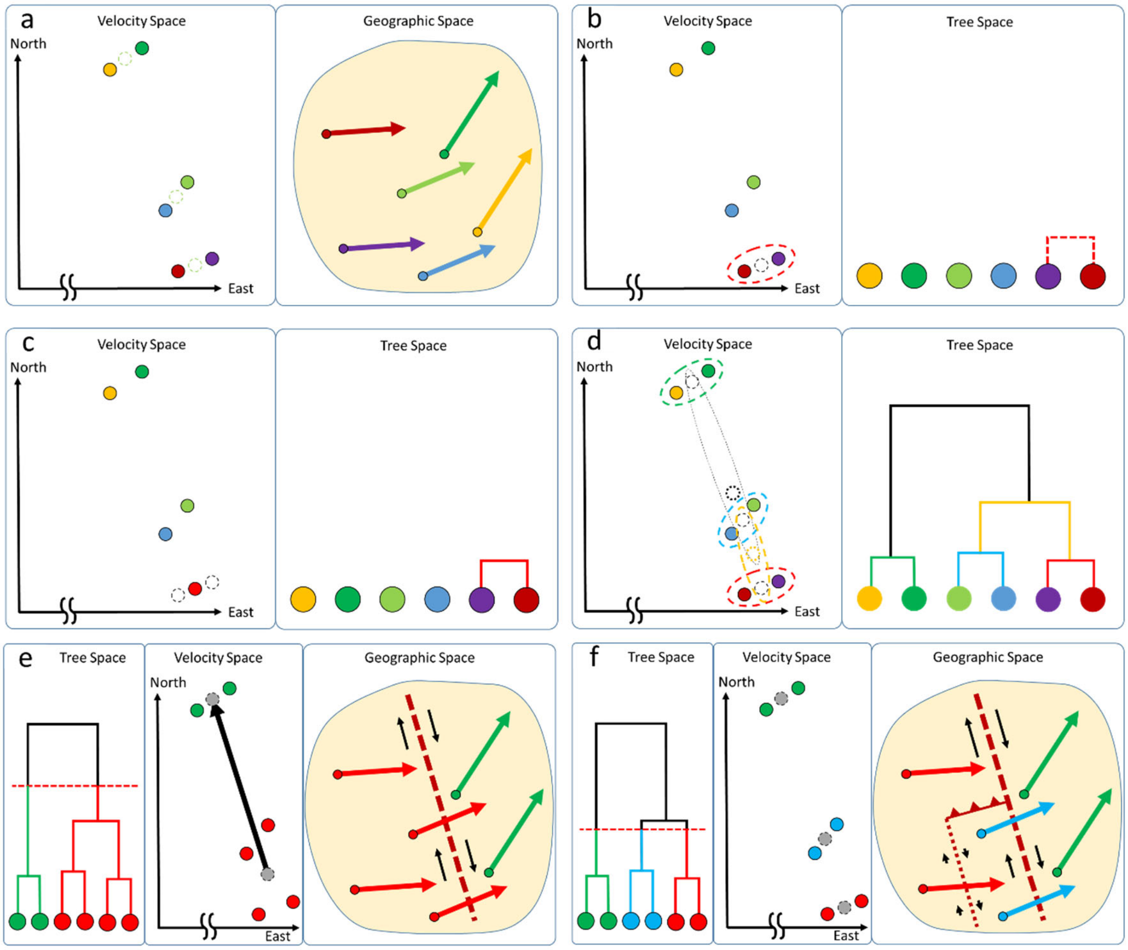

Figure 1 Panels explains the hierarchical agglomerative clustering algorithm for application to GNSS data. (a) Observed GNSS data are projected into the velocity space. Then the data are their own clusters. In this case, there are 6 initial clusters. (b) A pair of clusters whose distance is the closest is merged into their centroid and bridged in the tree space. To be concrete, the red and purple are the closest pair, and distance between them is the height of the horizontal bridge girder in the tree space. (c) The nearest pair of clusters is merged into their centroid. (d) Steps b and c are repeated until there is one single cluster remains. Finally, the data are connected each other in the tree space. The clusters are derived by cutting the organized tree at a certain height. (e) Cutting tree at a higher level, we can obtain small number of clusters. The red line crosses the tree branches twice, then two clusters are obtained. It corresponds to a significant velocity discontinuity. (f) Cutting the tree at a lower level introduces finer clusters. In this case, three clusters are derived since the red line crosses the tree branches three times. As the cutting height becomes lower than example in Figure 1e, an additional cluster boundary appears associated with a relatively moderate velocity discontinuity.

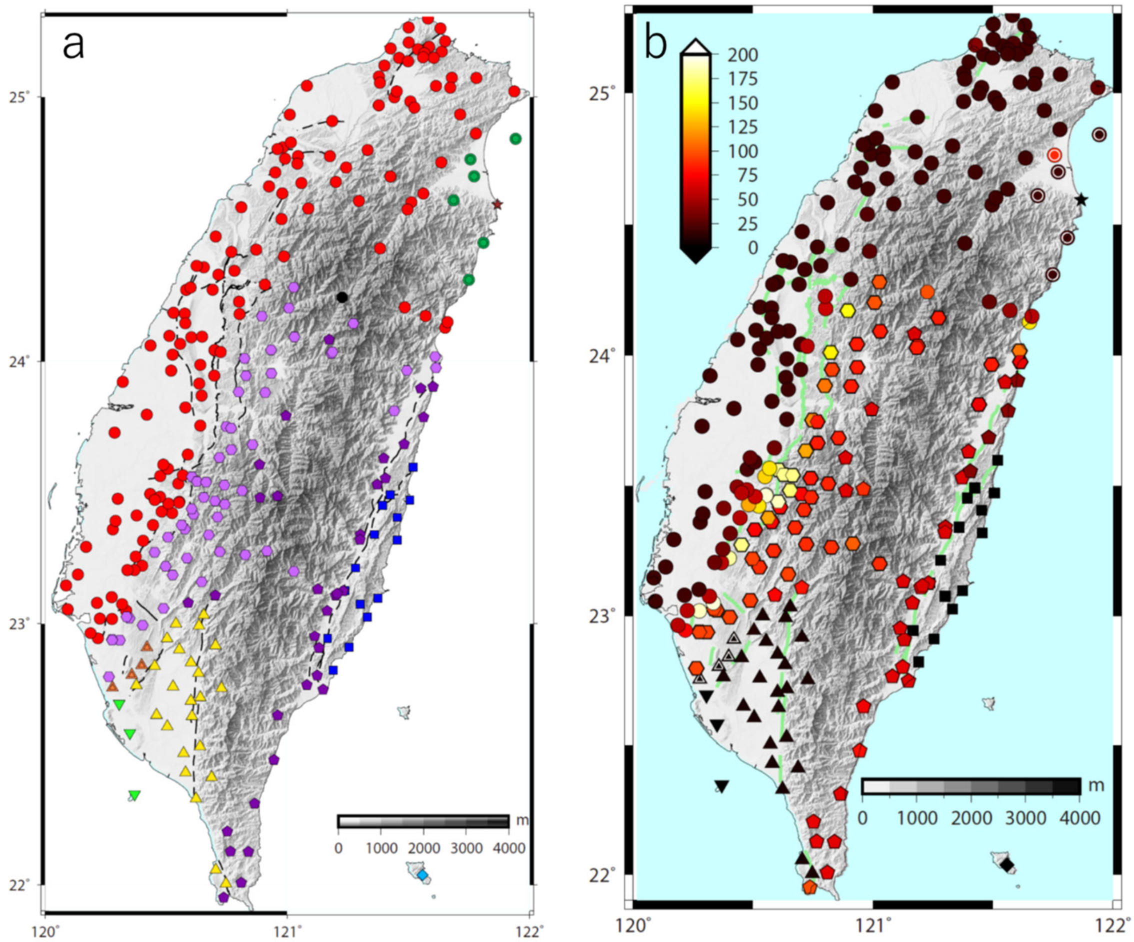

Figure 2 (a) The hierarchical clustering result for the case of 11 clusters in Taiwan. Black dashed lines are surface traces of known active faults. The elevation is indicated by a scale in the right bottom. (b) The entropy plot of cluster analysis in Taiwan for the same number of clusters. The color scale on the upper left indicates the entropy value at each GNSS station. Symbols are common among (a) and (b).