A GNSS-velocity clustering method applicable from local to global scales

Atsushi Takahashi¹, Keisuke Yano², and Masayuki Kano³

¹RIKEN, Center for Advance Intelligence Project
²The Institute of Statistical Mathematics
³Tohoku University

May 28, 2024

Abstract

We propose a hierarchical clustering methodology for Global Navigation Satellite System (GNSS) data that is applicable from local to global scales. We first adapted the conventional 2D velocity clustering metric for global-scale applications by implementing parallel translation in differential geometry. Then, we combined it with an Euler pole-based metric to incorporate the physical nature of plate motions, achieving advantages in identifying tectonic structures. This hybrid metric approach is examined through two case studies at different spatial scales to determine whether it can accurately identify tectonic plate and crustal block boundaries; one study utilizes global-scale data from the ITRF2008 plate motion model, and the other focuses on a local-scale study in Taiwan. Results obtained by using the hybrid metric consistently align better with geological data than those using either the 2D or Euler vector-based metrics alone. The proposed method is computationally efficient, enabling us to conduct two types of stability assessments: examining the robustness of clusters with synthetic noise contamination and performing leave-one-out analysis. These stability tests are demonstrated to be feasible within practical time frames.

Key points

1. We propose a GNSS clustering method that integrates existing metrics and scales from local to global.
2. Case studies at both global and local scales demonstrate that results of our method align more closely with geological information.
3. Computational efficiency enables various types of stability analyses for obtained clustering results.
A GNSS-velocity clustering method applicable from local to global scales

Atsushi Takahashi\textsuperscript{1}, Keisuke Yano\textsuperscript{2}, and Masayuki Kano \textsuperscript{3}

\textsuperscript{1}RIKEN, Center for Advance Intelligence Project
\textsuperscript{2}The Institute of Statistical Mathematics
\textsuperscript{3}Tohoku University

Key Points:

\begin{itemize}
  \item We propose a GNSS clustering method that integrates existing metrics and scales from local to global.
  \item Case studies at both global and local scales demonstrate that results of our method align more closely with geological information.
  \item Computational efficiency enables various types of stability analyses for obtained clustering results.
\end{itemize}

Corresponding author: Keisuke Yano, yano@ism.ac.jp
Abstract
We propose a hierarchical clustering methodology for Global Navigation Satellite System (GNSS) data that is applicable from local to global scales. We first adapted the conventional 2D velocity clustering metric for global-scale applications by implementing parallel translation in differential geometry. Then, we combined it with an Euler pole-based metric to incorporate the physical nature of plate motions, achieving advantages in identifying tectonic structures. This hybrid metric approach is examined through two case studies at different spatial scales to determine whether it can accurately identify tectonic plate and crustal block boundaries; one study utilizes global-scale data from the ITRF2008 plate motion model, and the other focuses on a local-scale study in Taiwan. Results obtained by using the hybrid metric consistently align better with geological data than those using either the 2D or Euler vector-based metrics alone. The proposed method is computationally efficient, enabling us to conduct two types of stability assessments: examining the robustness of clusters with synthetic noise contamination and performing leave-one-out analysis. These stability tests are demonstrated to be feasible within practical time frames.

Plain Language Summary
Clustering GNSS velocity data is a powerful tool for identifying key tectonic structures independently of geological information. Previously, two major metrics have been used for clustering: velocity similarity and Euler vector-based similarity. We have developed a hybrid metric that combines these two, enhancing its applicability for both global and local scale analyses. Our method was validated through two case studies using global plate motion data and local data from Taiwan. In both scenarios, our hybrid metric provided results that aligned more closely with geological data than those obtained using the single metrics. Additionally, this method is computationally efficient, enabling stability assessments for obtained clustering results. We then conducted stability analyses to assess the robustness of clusters in the presence of noise contamination and with elimination of partial data.
1 Introduction

In recent years, the rapid advancement of GNSS (Global Navigation Satellite System) observation networks has highlighted the significance of clustering GNSS velocity fields as a powerful tool for comprehending regional tectonics in various areas. Previous studies on clustering GNSS data can be broadly separated based on two fundamental contrasts: (i) whether they clustered data in the velocity space or the Euler pole space, and (ii) whether they used hierarchical clustering or partition-optimization algorithms.

Regarding criterion (i), there are two approaches that differ distinctly in their similarity metrics for clustering GNSS velocity fields. The first approach involves applying clustering to GNSS data distributed in east-west and north-south space, essentially conducting a 2D clustering of GNSS velocity data. This approach, initially proposed by Simpson et al. (2012), has widely supported due to its visual intuitiveness and ease of implementation, as evidenced by subsequent studies (e.g., Özdemir & Karshoğlu (2019); Takahashi et al. (2019); Granat et al. (2021)). While it is effective for relatively narrow study areas, this method has limitations. For instance, it may struggle to identify rigid rotations as clusters if the rotation pole is close to the geological center of a cluster, a limitation highlighted by Takahashi & Hashimoto (2022) in their analysis of GNSS data in New Zealand.

The second approach used Euler pole-based similarity metric, first introduced by J. C. Savage & Simpson (2013) for a case study in the Mojave region, California, and later refined by J. C. Savage & Wells (2015). This metric deals with the limitations of 2D clustering and expanded applicability of clustering approach to broader study areas. However, it may become unstable if the study area becomes regional. Therefore, ensuring consistency between the two metrics is desirable.

In criterion (ii), clustering methods are generally categorized into two principal approaches: hierarchical clustering, including agglomerative one and divisive one, and partition-optimization approaches, including the K-means algorithm and the Gaussian mixture modeling.

Hierarchical clustering aims to organize data into nested clusters (e.g., Hastie et al. (2009)). Hierarchical agglomerative clustering begins by assigning each data to its own cluster, resulting in N initial clusters, where N is number of data. Then, it merges the closest clusters in order based on similarity. This bottom-up procedure continues until all initial clusters are encompassed in a single cluster. The resulting hierarchical structure is represented by a dendrogram, a hierarchical binary tree-like representation. Clusters at different levels of the dendrogram represent different hierarchy levels of grouping. It is important to note that the derived cluster boundaries of clusters for the lower hierarchy do not cross those of clusters for the higher hierarchy. This cluster-boundary consistency is considered to be an important idea for applying it to tectonics. Moreover, each station is given equal weight all through the clustering process, this equitable treatment of respective GNSS data is particularly advantageous in scenarios where smaller clusters hold significant geological importance but might be overlooked by methods that prioritize improving misfit of large clusters. Finally, this method does not entail iterative calculations, making it relatively fast in terms of computational time.

Partition-optimization algorithms aim to partition data into cohesive groups using iterative clustering algorithm that operates in a top-down manner (Fukunaga, 2013). It begins by randomly selecting k centroids from the data. Each data is then assigned to the nearest centroid, forming initial cluster allocation. The centroids are recalculated as the centroid of data in each cluster. This process repeats until centroids stabilize, resulting in global cluster allocation of minimal misfit. While it is straightforward, it has limitations. The K-means method and its variations generally rely on initial cluster values, so that they may yield different results based on the initial conditions. Additionally, when
the number of observations is limited, the bias in station numbers can affect clustering
results. To be more specific, partition-optimization methods prioritize improving the global
misfit of clusters, which means that small clusters of significant geological importance
may not be identified. Takahashi & Hashimoto (2022) encountered this limitation in identi-
fying clusters of a small number of GNSS stations. Lastly, checking the stability of clus-
tering results via partition-optimization approaches would be more challenging as they
require unavoidable number of iterations to reduce misfits.

In this study, we propose an agglomerative hierarchical clustering method that en-
compases the cluster-boundary consistency as well as applicability from local to global
scales. To begin with, we extend the concept of 2D clustering to a more general frame-
work by incorporating parallel translation using the concept of the tangential plane, yield-
ing an parallel-translation-based similarity metric. We then reintroduce an Euler pole-
based pairwise similarity metric. These two metrics are then combined using a hyper-
parameter. A hierarchical agglomerative clustering based on this hybrid similarity met-
rpic yields the visual representation of the similarity of a velocity field by a dendrogram,
the consistent cluster boundary split, and consistency from global to local. We demon-
strate the effectiveness of this hybrid metric approach through case studies conducted
at both global and local scales, illustrating the robustness of our clustering method by
examining local-to-global scale consistency.

2 Method

Our method utilizes two similarity metrics: one is related to the translational mo-
tion (or orbital motion), the other is the rotational motion (or spin motion) (J. Savage,
2024); The first metric is defined by “parallel translation” of tangent vectors, that is vec-
tors tangent to the spherical Earth at a given point. The second metric is measured by
fitting the Euler poles.

2.1 Parallel-translation-based similarity metric

To define the first metric, we need to introduce parallel translation in differential
geometry (Spivak, 1999). For two points $P$ and $Q$ at the spherical Earth, let $r_P$ and $r_Q$
be their GNSS position coordinates, and let \( v_P \) and \( v_Q \) be tangent vectors, that is, the shortest route (a segment of a great circle) from \( P \) to \( Q \). The parallel translation \( \Pi_{(P,Q)} \) maps a tangent vector \( v_P \) at \( P \) to a velocity vector \( v_Q \) at \( Q \) so that the norm and the angle to \( \gamma(P,Q) \) is maintained; \( ||v_P|| = ||v_Q|| \) and \( \langle v_P, v_{\gamma(P,Q)} \rangle = \langle v_Q, v_{\gamma(P,Q)} \rangle \) with the inner product \( \langle \cdot, \cdot \rangle \), where \( v_{\gamma(P,Q)} \) is a velocity vector parallel to the geodesic. The parallel translation naturally induces the parallel-translation-based metric (PT metric) between velocity vectors \( v_P \) and \( v_Q \) at two points \( P \) and \( Q \), given by

\[
d_{PT}(\{v_P, r_P\}, \{v_Q, r_Q\}) = ||\Pi_{(P,Q)}(v_P) - v_Q|| = ||\Pi_{(Q,P)}(v_Q) - v_P||. \tag{1}
\]

This metric can evaluate the similarity between velocity vectors at arbitrary two points on the spherical Earth.

Figure 1 (a) displays how a velocity vector moves to the plane \( T_Q \) orthogonal to \( r_Q \) via the parallel translation. Compared to the parallel translation, velocity vectors at \( P \) and \( Q \) with the same east-west and north-south components maintain the angle with the meridian at each site. Figure 1 (b) exhibits the translation of a velocity vector \( v_P \) on maintaining the angle with the meridian to the plane \( T_Q \) orthogonal to \( r_Q \). If two points are very close, two translations match, which implies that clustering velocity vectors using parallel translation is an extension of the 2D clustering. Yet, these are different if two points do not satisfy the closeness condition as described in Figure 1 (b). See also Supporting Information 1 for further properties of the parallel translation.

### 2.2 Euler-vector based similarity metric

The PT metric takes the translational motion of a block into consideration as depicted by Figure 1 (a). For a point locating an Euler pole, the rotational motion of a block should be taken into consideration. To do this, we consider the following Euler-vector-based metric (EV metric) given by

\[
d_{EV}(\{v_P, r_P\}, \{v_Q, r_Q\}) = \sqrt{||v_P - r_P \times \hat{\omega}_{PQ}||^2 + ||v_Q - r_Q \times \hat{\omega}_{PQ}||^2}, \tag{2}
\]

where \( \hat{\omega}_{PQ} \) is defined by

\[
||v_P - r_P \times \hat{\omega}_{PQ}||^2 + ||v_Q - r_Q \times \hat{\omega}_{PQ}||^2 = \min_{\omega \in \mathbb{R}^3} \left\{ ||v_P - r_P \times \omega||^2 + ||v_Q - r_Q \times \omega||^2 \right\}, \tag{3}
\]

where the cross product of two vectors \( a, b \in \mathbb{R}^3 \) is denoted by \( a \times b \).

### 2.3 Hybrid similarity metric

Finally, to cope with both translational and rotational motions, we consider the following hybrid metric (EVPT metric) given by

\[
d_{EVPT}(\{v_P, r_P\}, \{v_Q, r_Q\}) = w_1 d_{PT}(\{v_P, P\}, \{v_Q, Q\}) + w_2 d_{EV}(\{v_P, P\}, \{v_Q, Q\}), \tag{4}
\]

where \( w_1 \geq 0 \) and \( w_2 \geq 0 \) are weights for two metrics and are taken to be 1 in this paper.

### 2.4 Hierarchical agglomerative clustering

Hierarchical agglomerative clustering is a bottom-up clustering that starts with each point as a singleton and then iteratively merges clusters until the number of clusters is one. This approach only requires a similarity metric for points and the choice of similarity for clusters called a linkage criterion. In this paper, to keep the consistency to the previous study (Takahashi et al., 2019), we adopt the average linkage criterion

\[
d(A, B) = \frac{1}{|A||B|} \sum_{(v_P, r_P) \in A} \sum_{(v_Q, r_Q) \in B} d(\{v_P, r_P\}, \{v_Q, r_Q\}) \tag{5}
\]
for two clusters $A$ and $B$ consisting of sets of velocity vectors and position coordinates, where $|\cdot|$ denotes the number of elements and $d(\{v_P, r_P\}, \{v_Q, r_Q\})$ denotes a similarity metric between two velocity vectors at different points (say, $d = d_{EVPT}$).

### 2.5 Visualization of similarity matrices by UMAP

We shall use the UMAP (Uniform Manifold Approximation and Projection; McInnes et al. (2018)) to visualize pairwise similarities in the 2D space, making them more intuitive. UMAP embeds data endowed with a (possibly complex) similarity into a lower-dimensional space (often a 2D space) by optimizing a cost function designed to preserve the original pairwise similarity structure. This approach helps visually comprehend how the data similarity varies when measured with three different metrics.

### 3 Case study with global data

#### 3.1 Dataset

To examine our proposed method on a global scale, we applied our method to the dataset used in establishing the ITRF2008 reference frame (Altamimi et al., 2012) (Fig. 2a). This dataset consists of 206 IGS (International GNSS Service) sites satisfying specific criteria: (1) each site must have a time-span of observations exceeding three years, (2) they must be situated at least 100 km away from Bird’s plate boundaries (Bird, 2003), outside deformation zones according to the criteria of Argus & Gordon (1996) and the strain map of Kreemer et al. (2003, 2006), and be distant from GIA (Glacial Isostatic Adjustment) regions, and (3) their normalized post-fit velocity residuals (raw residuals divided by their a priori uncertainties) must be smaller than 3, with raw residuals less than 3 mm/a (Altamimi et al., 2012). Therefore, the secular motions at each site is supposed to reflect each plate’s motion, and by successfully replicating the site classification onto plates, we can examine that our method effectively operates on a global scale.

#### 3.2 Results and Discussion

Figure 2 displays the clustering results with $K = 4$ (major clustering results) and $K = 14$ (minor clustering results with $K$ equaling the number of the plate allocation of the ITRF2008 plate motion model) of the EVPT metric. Figures SI 1 and SI 2 display the corresponding results for the PT and the EV metrics, respectively. We first explain the PT and EV metric results and then explain the hybrid metric results.

The PT metric (Figure SI 1) first introduce the Pacific cluster (light blue), followed by Nazca (deep blue) and India-Australia-Arabia (yellow). With $K = 14$, GNSS sites in Antarctica, South America, Somalia, and Nubia are grouped into a single red cluster. In central Russia, orange arrows are grouped with arrows in North America, indicating a strong similarity in motion.

In the EV metric results (Figure SI 2), the Australian plate is introduced first, followed by the Pacific and Nazca clusters. At $K = 14$, South America (light green) is correctly separated from Eurasia (red). However, GNSS sites in northern Nubia are grouped with those in India and Arabia. Additionally, geographically distant cluster distribution is characteristic of this metric. Examples include some central Russian sites and a site in the South Atlantic Ocean grouped with those in North America, the Antarctic cluster including a site in central Russia and Korea, and a site in North America clustered with a site in the Maldives in the Indian Ocean, despite their unlikely connection, forming geographically distant groups or pairs.

The hybrid metric results (Figure 2) are more geographically consistent to the plate allocation of the ITRF2008 plate motion model, reducing misclassification, which is also
shown in the matrix representing EVPT pairwise similarity (Figure SI 3). The dendrogram (Figure 3) exhibits the whole picture of the clustering result. At $K = 2$, the initial cluster division is the Pacific cluster versus the rest. At $K = 3$ and $K = 4$, the Nazca and Australia-India-Arabia (AIA) clusters emerge. There is a notable gap between $K = 4$ and $K = 5$, suggesting that $K = 4$ can be the number of major clusters. At $K = 14$, we examine the minor cluster distribution. The nested sub-clusters of AIA (Australia, India, and Arabia) are introduced, and they correctly match the ITRF2008 plate labels. However, subdivisions also appear within Australia (green and navy arrows in New Caledonia and New Zealand’s North Island). Subdivisions are also seen in Nazca (blue and gray arrows) and the Pacific (magenta), which is also confirmed in the PT metric results. Meanwhile, Eurasia, Nubia, Amurian, and Somalia are grouped into the same cluster (red arrows in Figure 2c). Contamination between South America and Antarctica is also present in the EVPT metric results.

The orders of splits in hierarchical clustering based on the PT and EVPT metrics are examined by computing the pairwise similarity of the 14 Euler vectors in the ITRF plate motion model using the Euclidean distance (Figure SI 4). The Euler vectors of the Pacific plate are the most distinct, followed by Nazca and Australia-India-Arabia. Figure SI 4 indicates block-like structures on the diagonals, meaning that the Euler vectors of Australia-India-Arabia are close to each other but not similar to the other plates. The ITRF2008 model shows that the Euler poles of Eurasia, Amurian, Nubia, and Somalia are close to each other. Thus, the orders of splits based on the PT and EVPT metrics align with the Euler vector dissimilarity of the ITRF2008 plate motion model.

The UMAPs of the similarities based on the three metrics show different data shapes in the embedded 2D spaces (Figure 4). Contamination between South America and Antarctica in all results is visualized through their cohesive distribution. Sites in India and Arabia are close to the Eurasian sites using the EV metric, while the PT metric keeps the Australia-India-Arabia data isolated. This occurs because the parallel translation preserves the norm of velocity as well as the angle to the geodesic connecting two sites. When two sites are geographically close, they tend to have similar velocity vectors, i.e., they are expected to be almost parallel and have similar norms. So, the PT similarity metric between these sites is supposed to be greater. The distinct and cohesive distribution of Australia, India, and Arabia in the UMAP of the PT similarity metric can be attributed to this effect. However, relying solely on the PT metric can overemphasize similarities of regional velocity vectors, potentially causing locally isolated clusters and overlooking rotational motions. While the EV metric does not exhibit this shortcoming as it takes rigid (rotational) motions into consideration, it can cluster geographically distant groups or pairs that are difficult to attribute to the same rigid rotation as it only assesses the pairwise fit of a rigid rotation without using any geographical constraint. The UMAP of the EVPT similarity metric shows a cohesive data shape but is not fully constrained to local velocity similarities. It demonstrates that the hybrid metric balances the characteristics of both PT and EV metrics: maintaining local velocity similarity while also considering the rigid rotation as a constraint, leading to more reasonable clustering than any single metric alone.
Figure 2: (a) The original GNSS horizontal velocities with arrows with respect to the ITRF2008 reference frame. The colors of the arrows denote the plate allocation of the ITRF2008 plate motion model. (b) clustering result using the EVPT metric when the number $K$ of clusters is 4. The cluster allocation is denoted by colors of arrows. (c) clustering results for $K = 14$ with the same manner.
Figure 3: Dendrogram of global data obtained by the EVPT similarity metric. The colored rectangles at the bottom represent original label of belonging plates by Altamimi et al. (2012) shown in the right upper legends.

Figure 4: UMAP plots for (a) PT similarity metric, (b) EVPT similarity metric, and (c) EV similarity metric. The symbols and colors represent original plate labels of the ITRF2008 plate motion model.
4 Case study in Taiwan

4.1 Dataset and Geological settings

Another experiment for evaluating the effectiveness of the method on a local scale, we also applied our method to GNSS data in Taiwan to examine the clustering results of the previous study (Takahashi et al., 2019). We used the same secular motion data from 281 continuously monitored GNSS stations spanning from 2007 to 2013, as published by Tsai et al. (2015). These stations were operated by the Central Weather Bureau, Institute of Earth Science of Academia Sinica, and Central Geological Survey. Prior to conducting time series analysis, outliers were removed using a specific standard deviation (SD) threshold (set at 3.5 times the normal SD) and a moving average approach with a 30-day window, shifting every 10 days. Subsequently, the secular velocity of each GNSS site was determined from the positional time series data by applying the model proposed by Nikolaidis (2002), which distinguishes between seasonal and secular components.

The primary tectonic setting of Taiwan is characterized by the collision of the Luzon Arc on the Philippine Sea plate with the Eurasian continental margin at a rate of up to 82 mm/year (DeMets et al., 2010), which causes frequent seismic activity. The geological domains in Taiwan are divided into five main regions (Ho, 1986): the Coastal Range, which originated from the Luzon Arc; the Longitudinal Valley Fault, marking the suture zone between the Philippine Sea Plate and the Eurasian Plate; the Central Mountain Range, uplifted by converging plates; the Western Foothills, characterized by fold-and-thrust belts; and the Coastal Plain (Figure 5a). Frequent and intense seismic activities are observed along the Longitudinal Valley Fault. The 2024 Mw. 7.2 devastating Hualien earthquake occurred along the northern part of the Lingding Fault (Figure 5b), which locates the northern end of the Longitudinal Valley Fault. To the east of Taiwan, the Philippine Sea Plate is subducting beneath the Eurasian Plate along the Ryukyu Trench (Takada et al., 2007). Behind the Ryukyu Trench, back-arc spreading is ongoing and its western end of this spreading zone is landing on the Ilan Plain (Hu et al., 2002).

4.2 Results and Discussion

Figure 6 displays the clustering results with $K = 4$ (major clustering results) and $K = 11$ (minor clustering results with $K$ equaling the number of the cluster allocation determined by Takahashi et al. (2019)) of the PT and EVPT metrics. Figure SI 3 displays the corresponding results for the EV metrics. We first explain the PT and EV metric results and then explain the hybrid metric results.

The clustering result based on the PT metric is compatible with previous research, as expected, and it replicates the results of Takahashi et al. (2019). For instance, at $K = 4$ (major clusters), we identify clusters corresponding to the Coastal Range and the Philippine Sea Plate (blue), the southeastward local motion (brown), and the southwestward motion in southwestern Taiwan (yellow). The boundary from the PT metric (Figure 6a) and that of Takahashi et al. (2019) terminates at the middle of the Longitudinal Valley, imposing the Chimei fault (Figure 5b), although geological studies (Shyu et al., 2006) reported that it has been inactive for at least 3 kyrs. At $K = 11$, the clusters from the previous study are replicated.

In contrast, the primary cluster boundary obtained with EV metric appears along the Longitudinal Valley Fault, though one of the GNSS stations in the coastal Range is allocated to red cluster. The clusters identified using the EV metric become geographically inconsistent as the number of clusters increases. At $K = 4$, the yellow cluster is surrounded by the red cluster, and finer cluster splits further worsen this issue. For
Figure 5: (a) Geological domain and the surface traces of active faults of Taiwan according to Ho (1986) and Lin et al. (2012). Colored regions denote following domain; Western Foothills (green); Central Range (blue); Coastal Range (red). The Longitudinal Valley is a region between red and blue. The Coastal plain is on west side of the Western Foothills. The dashed rectangle indicate region in Figure 5b. The numbered circles indicate following region names: (1) Ilan Plain, (2) Hualien, (3) Lanhsu Islands, and (4) Pingtung Plain. (b) Surface traces of active faults in the Longitudinal Valley.
Figure 6: Maps of GNSS velocity data in Taiwan and clustering results using PT/EVPT similarity metrics. (a) the clustering result with $K = 11$ by PT metric that is consistent to that of the 2D-clustering (Takahashi et al., 2019); (b) the clustering result with $K = 4$ by PT metric; (c) the clustering result with $K = 11$ by EVPT metric; (d) the clustering result with $K = 4$ by EVPT metric.

For example, at $K = 11$, the Lanhsu Island cluster contains two GNSS sites from the Coastal Range, and the pink and red clusters are scattered across the Coastal Plain. As in Figure 7 (c), the UMAP of the EV metric does not exhibit a cohesive data shape, suggesting that minor clusters are not well-determined by this metric.

When using the EVPT metric, the primary cluster boundary, introduced along the Longitudinal Valley Fault rather than the Chimei fault, resembles that of the EV metric result. The cluster covering the Coastal Range has two subclusters: Lanhsu Island and the northern part of the Coastal Range. The former subcluster is recognized in the EV metric clusters. However, the latter subcluster, indicated by four black arrows in Figure 6, has not been identified in the results of either metric and is uniquely attributed to the hybrid metric. This cluster aligns well with the Lingding-Rueysui fault series, where the 2024 Hualien earthquake occurred. For larger $K$ values, clusters are geographically consistent and similar to those from the PT metric results. However, a single cluster identified in the PT metric is not replicated by the EVPT metric, and the single station is
Figure 7: UMAP plots for GNSS velocity data in Taiwan. (a) UMAP for the PT metric; (b) UMAP for the EVPT metric; (c) UMAP for the EV metric. The symbols, the colors, and the cluster names represent the clustering result with $K = 11$ of Takahashi et al. (2019).

5 Stability Analyses

This section presents stability analyses on the clustering results of our method. These analyses can be easily conducted owing to the computational benefits of our method. Besides its computational efficiency, an implementational feature of the hierarchical agglomerative clustering is that it avoids entrapment in local optima, a common issue with partition-optimization algorithms, by progressively merging clusters based on a defined similarity metric.

Here, we conduct two types of stability tests. The first test assesses the robustness of the clusters against synthetic noise contamination and is conducted on the global dataset. The noise level was determined by the standard deviation of each site provided by Al-tamimi et al. (2012). The second test evaluates how well the cluster allocation can be replicated when only partial data is available. This second test is demonstrated using
the dataset of Taiwan. In general, the computation time for stability assessment using subsampling is short, enabling thorough stability checks.

The first test depicted by Figure SI 1 uses the clustering result with \( K = 14 \) based on the EVPT metric. The overall computational time for the whole analysis is around 25 minutes using Python 3 on Google Colaboratory, where creating an \( N \times N \) similarity matrix for each synthetic setup consumes most of the computational time. Figure SI 1 displays the relation matrix \((R_{ij})_{1 \leq i,j \leq N=206}\) defined by

\[
R_{ij} = \sum_{m=1,\ldots,M} 1_{C_m[i]=C_m[j]} / M
\]

where \( M \) is the number of the synthetic-noise patterns (here \( M = 100 \)), and \( C_m = \{C_m[i] : i = 1,\ldots,N\} \) is the cluster allocation vector of the \( m \in \{1,\ldots,M\} \)-th trial for \( N = 206 \) sites.

For a pair \((i,j)\) in the same cluster of the original result, \( R_{ij} = 1 \) means the original result is perfectly replicated even in the noise contamination, while \( R_{ij} = 0 \) means the original clustering result is never replicated. For a pair \((i,j)\) of the different cluster in the original result, \( R_{ij} = 0 \) and \( R_{ij} = 1 \) mean that the original cluster allocation is always replicated and never replicated, respectively. From Figure SI 1, we may conclude that the original clustering result is robust to the presence of noise contamination. One of the reasons for this robustness would be that the signal-to-noise ratio of the global data is quite high, thanks to the rigorous selection and processing of the data.

![Stability check for Taiwan dataset. (a) the map of influential points in the Adjusted Rand Index (ARI). Stations with lower ARI values have higher influence on the clustering. (b) the map of influential points in Shannon’s entropy. The cluster structure related to a station with a higher entropy value seems to be more fragile. (c) the map of leave-one-out probabilities. Clusters with high probabilities are more robust against leave-one-out resampling.](image)

The second test is applied to the clustering result of EVPT metric with \( K = 11 \) in Taiwan. First, we check the scalability of our method in the computational time of the stability test by increasing the number of used stations. For each number of used station, we randomly picked stations up, computed the pairwise EVPT similarity matrix, and applied the hierarchical agglomerative clustering, yielding a clustering result with \( K = 11 \). It takes around 20 seconds to prepare the similarity matrix. The computation time without it is plotted in Figure SI 2. For any number of used stations, the computation time is less than 0.002 second, almost instantaneous, implying the fast computation for stability assessment with subsampling from the similarity matrix. In other words,
once a similarity matrix is computed, the hierarchical clustering of the sub-sampled data from the matrix is instantaneous.

Second, we take a closer look at the leave-one-out (jackknife) analysis, in which one station is left out and the rest, $N - 1$ stations, are used for clustering, where $N = 281$. Specifically, this analysis computes 281 patterns, each corresponding to the absence of a single station. Through this analysis, we can assess the influence of individual datum on the clustering results using two different measures.

One measure is the Adjusted Rand Index (ARI). The ARI evaluates how similar the two sets of clustering results are and is calculated using the following formula:

$$\text{ARI} = \frac{\text{RI} - E[\text{RI}]}{1 - E[\text{RI}]},$$

where RI is the (unadjusted) Rand index given by

$$\text{RI} = \frac{\text{True Positive} + \text{True Negative}}{\text{Number of total pairs} \times \frac{N(N-1)}{2}},$$

with True Positive meaning the number of pairs assigned to the same cluster and True Negative meaning the number of pairs assigned to different clusters, and $E[\text{RI}]$ indicates the expectation of the Rand index with respect to the hypergeometric distribution. The ARI values range from -1 to 1, where 1 indicates perfect agreement between two clusterings, 0 indicates no better agreement than random, and negative values suggest worse than random clustering results.

For each pattern, we calculate the Adjusted Rand Index (ARI) between the resulting cluster and the original result (Figure 6 c).

Figure 8(a) shows the ARI values for all pattern, with each pattern corresponding to an absent station. Interestingly, influential stations in terms of ARI do not necessarily appear in the cluster boundaries, rather they often appear in boundaries of geological domains; the boundary of Western Foothills and the Coastal plain, the boundary of Central Range and Western Foothills, and the Logitudinal Valley.

The other measure is Shannon’s entropy for pairwise links

$$H(R_{ij}) = -R_{ij} \log(R_{ij}) - (1 - R_{ij}) \log(1 - R_{ij}) \quad (1 \leq i \neq j \leq N),$$

where $R_{ij}$ is the relation matrix. To quantify the influence of the $i$-th station, we take an average of $H(R_{ij})$ with respect to $j$:

$$H(i) = \sum_{j \neq i} H(R_{ij})/(N - 1) \quad (i = 1, \ldots, N).$$

The value of $H(i)$ shows ambiguity of the cluster structures related to the $i$-th station. The higher values of $H(i)$ is that the cluster structure related to the $i$-th station is more unstable, while $H(i) = 0$ means that the cluster structure is almost deterministic. The identified influential points in entropy measure are somehow similar to those in ARI. However, the single station introduced as a single cluster with the PT metric (a black arrow in Figure 6) has higher entropy value, implying that current allocation of this station to the red cluster might not be so confident.

The leave-one-out analysis also offers a confidence value for each cluster. Figure 8 (c) shows the map of leave-one-out probabilities defined as below. For each station $i$, we compute

$$p_{C[i]} = \sum_{m=1}^{N} \frac{1 \{ j : C[j]=C[i] \} = \{ k : C^{-m}[k]=C[i] \}}{N},$$

with $C^{-m}$ denoting the pattern obtained by excluding the $m$-th station from the original clustering.
where $C = (C[i])_{i=1,...,N}$ is the original cluster allocation and for $m = 1, \ldots, N$, $C^{-m} = (C^{-m}[i])_{i=1,...,m-1,m+1,...,N}$ is the cluster allocation based on data without the $m$-th station. The leave-one-out probability gives a confidence value for a cluster that each station belongs to. This corresponds to a leave-one-out version of the bootstrap probability (Felsenstein, 1985; Suzuki & Shimodaira, 2006) in the phylogenetic tree analysis. The value 1 means that the cluster is perfectly replicated in every trial despite leaving one data out. The cluster covering the Longitudinal Valley has the highest confidence value, while the clusters including Western Foothills and the Coastal plan has lower confidence values as the contrast of velocities in these area is not clear as seen in Figure 6. Relatively higher confidence value of the newly introduced cluster by the EVPT metric along the east side of the Lingding-Rueysui fault series is considered to be reliable.

Finally, we emphasize that the computational efficiency of our method also contributes to conventional iterative clustering approaches. Generally, the $K$-means method consumes significant computational resources to search for a good starting point for data clustering. By using clusters derived from our method as initial allocations, the initialization time for the $K$-means based iterative method can be reduced, addressing one of its major weaknesses and reducing the computation cost of conventional methods.

6 Conclusions

We have proposed a hierarchical clustering method of GNSS data that is applicable from global to local scale. The hybrid metric outperformed previously used metrics in both global and local case studies, estimating tectonic plate and crustal block boundaries consistent with geological data. This new method is computationally efficient, allowing for error analysis through bootstrapping or other resampling techniques within a practical timeframe. Using clusters derived from this method as initial allocations can reduce the initialization time of the $K$-means based method, dealing with one of its weaknesses and improving the overall speed of conventional methods.

Acknowledgments

The authors thank Yoshikazu Terada for helpful comments on an earlier version of this study. This study was supported by the MEXT Project for Seismology toward Research Innovation with Data of Earthquake (STAR-E) (Grant number JPJ010217), JSPS KAKENHI (Grant numbers JP21K03694, JP21H05205, JP21H05206, JP23H00466, JP24H01019, and 23K11024). We thank Wessel et al. (2019) for developing Generic Mapping Tools (GMT) version 6 and PyGMT, which were used to generate geographical maps presented in this study.

References


Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the boot-


Supporting Information 1 Further properties of parallel translation

This supplement provides further properties of parallel translation.

First, the parallel translation of $\mathbf{v}_P$ along $\gamma(P, Q)$ is calculated as follows. Calculate the following orthogonal vectors:

$$e_1 = r_P, \quad e_3 = \frac{r_P \times r_Q}{\|r_P \times r_Q\|}, \quad e_2 = e_3 \times e_1.$$  \hspace{1cm} (A1)

and then calculate the following angle:

$$\theta_{PQ} = \arctan2(\langle e_2, r_Q \rangle, \langle e_1, r_Q \rangle),$$  \hspace{1cm} (A2)

where $\arctan2$ is the two-argument arctangent. Then, the parallel translation is given by

$$\Pi_{\gamma(P, Q)}(\mathbf{v}) = \langle \mathbf{v}, e_2 \rangle \cos(\theta_{PQ}) e_2 - \langle \mathbf{v}, e_2 \rangle \sin(\theta_{PQ}) e_1 + \langle \mathbf{v}, e_3 \rangle e_3.$$  \hspace{1cm} (A3)

Second, the parallel translation becomes the translation in the 2D space when two points are very close as discussed in the following. Assume that $P$ and $Q$ are very close. Then, we get $\theta_{PQ} \approx 0$, implying

$$\Pi_{\gamma(P, Q)}(\mathbf{v}) \approx \langle \mathbf{v}, e_2 \rangle e_2 + \langle \mathbf{v}, e_3 \rangle e_3.$$  \hspace{1cm} (A4)

For the simplicity, we shall consider the two cases with $P$ and $Q$ having the same latitude or the same longitude. Consider the former case, where $P$ and $Q$ has the same latitude, the axis $\pm e_3$ corresponds to the East-West axis and then the axis $\pm e_2$ corresponds to the North-South axis. So, $\Pi_{\gamma(P, Q)}$ becomes the translation in the 2D space. Consider the latter case, where $P$ and $Q$ has the same longitude, the axis $\pm e_3$ corresponds to the North-South axis and then the axis $\pm e_2$ corresponds to the East-West axis. So, $\Pi_{\gamma(P, Q)}$ becomes the translation in the 2D space. Consider the case where $P$ and $Q$ do not have the same latitude or the same longitude.
Supporting Information 2  Supporting figures of case study with global data

This supplement provides supporting figures for Section 2.
Figure SI 1: PT metric results of the global data with the numbers of clusters $K = 2, 4,$ and 14, arranged vertically in one panel. Colors in each figure correspond to the cluster allocations for each number of clusters.
Figure SI 2: EV metric results of the global data with the numbers of clusters $K = 2, 4$, and 14, arranged vertically in one panel. Colors in each figure correspond to the cluster allocations for each number of clusters.
Figure SI 3: Similarity matrix representing the EVPT similarity metrics for global data. The similarity metrics $d_{EVPT}(P,Q)$ for all pair $(P,Q)$ are plotted as the matrix. The color in the cell corresponds to the value of the metric with the scaling $\exp(d_{EVPT}(P,Q)) - 1$, where the scaling is used for visual clarification. The colors in the row indices correspond to the colors for the plate allocation of the ITRF2008 plate motion model.
Table SI 1: The 14 Euler vectors of ITRF2008 Plate motion model (Altamimi et al. (2012)).

<table>
<thead>
<tr>
<th>Plate</th>
<th>$\Omega_x$</th>
<th>$\Omega_y$</th>
<th>$\Omega_z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCFC</td>
<td>-0.411</td>
<td>1.036</td>
<td>-2.166</td>
</tr>
<tr>
<td>NAZC</td>
<td>-0.33</td>
<td>-1.551</td>
<td>1.625</td>
</tr>
<tr>
<td>AUST</td>
<td>1.504</td>
<td>1.172</td>
<td>1.228</td>
</tr>
<tr>
<td>INDI</td>
<td>1.232</td>
<td>0.303</td>
<td>1.54</td>
</tr>
<tr>
<td>ARAB</td>
<td>1.202</td>
<td>-0.054</td>
<td>1.485</td>
</tr>
<tr>
<td>N.AM</td>
<td>0.035</td>
<td>-0.662</td>
<td>-0.1</td>
</tr>
<tr>
<td>S.AM</td>
<td>-0.243</td>
<td>-0.311</td>
<td>-0.154</td>
</tr>
<tr>
<td>ANTA</td>
<td>-0.252</td>
<td>-0.302</td>
<td>0.643</td>
</tr>
<tr>
<td>SUND</td>
<td>0.047</td>
<td>-1</td>
<td>0.975</td>
</tr>
<tr>
<td>EURA</td>
<td>-0.083</td>
<td>-0.534</td>
<td>0.75</td>
</tr>
<tr>
<td>AMUR</td>
<td>-0.19</td>
<td>-0.442</td>
<td>0.915</td>
</tr>
<tr>
<td>NUBI</td>
<td>0.095</td>
<td>-0.598</td>
<td>0.723</td>
</tr>
<tr>
<td>SOMA</td>
<td>-0.08</td>
<td>-0.745</td>
<td>0.897</td>
</tr>
<tr>
<td>CARB</td>
<td>0.049</td>
<td>-1.088</td>
<td>0.664</td>
</tr>
</tbody>
</table>

Figure SI 4: A heatmap illustrates the pairwise Euclidian dissimilarity of the 14 Euler vectors of ITRF2008 plate motion model.
Figure SI 5: UMAP plots for (a) PT similarity metric, (b) EVPT similarity metric, and (c) EV similarity metric. The symbols and colors represent original plate labels of the ITRF2008 plate motion model. Colors of symbols in each figure correspond to the clustering results obtained using the corresponding similarity metric.
Figure SI 6: Zoomed UMAP plots of Figure SISI 5 for (a) PT similarity metric, (b) EVPT similarity metric, and (c) EV similarity metric. The symbols represent original plate labels of the ITRF2008 plate motion model. Colors of symbols in each figure correspond to the clustering results obtained using the corresponding similarity metric.
Supporting Information 3  Case study in Taiwan

Figure SI 1: Dendrograms of EV, EVPT, and PT for Taiwan data.
Figure SI 2: Clustering result for $K=11$ using EVPT metric. For clarity, we use light blue instead of black to indicate the cluster in Hualien.
Figure SI 3: 3x2 grid example arranging the results of EV, EVPT, and PT.
Figure SI 1: The relation matrix at $K=14$ of the EVPT metric for global data. The colors in the row indices correspond to the colors for the cluster allocations without noise injection.
Figure SI 2: Computational time of resampling with Taiwan data. The mean curve is plotted as the blue solid line. Regions within the 25%-75% quantiles are hatched in blue.

Figure SI 3: The leave-one-out relation matrix at $K = 14$ of the EVPT metric for global data, where the matrix is calculated as the mean of relation matrices produced in the leave-one-out procedure. The colors in the row indices correspond to the colors for the cluster allocations.
Figure SI 4: The leave-one-out relation matrix at $K = 11$ of the EVPT metric for Taiwan data, where the matrix is calculated as the mean of relation matrices produced in the leave-one-out procedure. The colors in the row indices correspond to the colors for the cluster allocations.
Figure SI 5: Stability check for the EVPT results at $K = 14$ for global data. (a) the map of influential points in ARI. (b) the map of leave-one-out probabilities.