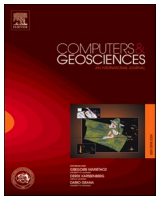
Computers & Geosciences 169 (2022) 105243

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo



On analyzing GNSS displacement field variability of Taiwan: Hierarchical Agglomerative Clustering based on Dynamic Time Warping technique

Utpal Kumar [a], Cédric P. Legendre [b], Jian-Cheng Lee [c], Li Zhao [d,e], Benjamin Fong Chao [c,*]

[a] Berkeley Seismological Laboratory, University of California, McCone Hall, 215 Haviland Path #4760, Berkeley, CA, USA
[b] Institute of Geophysics, Czech Academy of Science, Boční II/1401, 141 31 Praha 4, Spořilov, Czech Republic
[c] Institute of Earth Sciences, Academia Sinica, No. 128, Section 2, Academia Rd, Nangang District, Taipei City, Taiwan
[d] School of Earth and Space Sciences, Peking University, No.5 Yiheyuan Road Haidian District, Beijing, China
[e] Hebei Hongshan Geophysical National Observation and Research Station, Peking University, No.5 Yiheyuan Road Haidian District, Beijing, China

A B S T R A C T

We investigate the feasibility of using the Dynamic Time Warping (DTW) technique to cluster continuous GNSS displacements in Taiwan. Using the DTW distance as the measure for waveform similarity, we combine the DTW method with the Hierarchical Agglomerative Clustering (HAC) algorithm. This is in contrast to the conventional clustering approach that uses the Euclidean distance, considering the average long-term crustal motion, but inherently neglects full-waveform temporal variations. Here we apply the DTW-based HAC algorithm adopting DTW distance as the waveform similarity measure on 11 years worth of 3-D displacement data from 115 continuous GNSS network stations in Taiwan. We demonstrate the efficacy of the DTW-based HAC method in distinguishing the GNSS spatiotemporal variabilities that are consistent with the known, complex tectonic behavior of the region. An open-source Python package has been developed and made available to perform the HAC analysis.

1. Introduction

Cluster analysis has been widely applied in many fields for data exploration and for associating objects with relatively low similarity measures into groups. It is increasingly becoming a necessity with the exponential growth of real-world data. For example, in geodesy-related studies, the GNSS (Global navigation satellite system) data provide increasingly detailed pictures of local tectonics, including lithospheric deformation patterns and fault identifications (Yu et al., 1997; Rau et al., 2008; Angelier et al., 2009; Wu et al., 2009; Chang et al., 2012). When applied to the GNSS data the clustering approach is powerful in searching for spatially coherent patterns of deformations and revealing relations among the data for better interpretability (Mohapatra et al., 2012).

Recently, Euclidean distance-based clustering methods have been used to quantitatively identify the velocity discontinuities linked to regional crustal block structures (Simpson et al., 2012; Savage and Simpson, 2013b; Takahashi et al., 2019). They rely on the Euclidean distance (ED) to quantify the differences in time series. This approach is data-driven, considers no *a priori* geological information, and has several advantages over the traditional subjective block analysis approach (Thatcher, 2009) that is usually guided by the choice of tectonic blocks, their number and size, the distribution, and geometry of known faults. However, this approach considers the average secular trend of 2-D horizontal velocity and does not take into account the short-period temporal variability in the horizontal waveforms. In addition, vertical variations are ignored altogether. These lead to reduced sensitivity to localized temporal crustal variations as well as to the crustal deformation associated with normal and reverse faults (Takahashi et al., 2019). These limitations hamper the resolution of the geodynamical features in tectonically complex regions that feature rapid tectonic variations and where the 2-D horizontal velocity vectors may be inadequate to accurately represent the active tectonics.

Anomaly detection techniques capture patterns that significantly deviate from the expected behavior. Several clustering-related anomaly detection techniques have been explored in the past (e.g., Izakian and Pedrycz, 2013; Iwata and Umeno, 2017; Xia et al., 2020), however, most of them have employed the ED as the dissimilarity measure.

Here we experiment with a new approach for incorporating the full time series information that combines the clustering analysis with the Dynamic Time Warping (DTW) method. DTW is a well-known method that generically uses dynamic programming to evaluate the similarity/dissimilarity of time series with respect to their shape information (Angeles-Yreta et al., 2004; Strle et al., 2009; Hale, 2013; Venstad, 2013;

* Corresponding author.
E-mail addresses: utpalkumar@berkeley.edu (U. Kumar), legendre@ig.cas.cz (C.P. Legendre), jclee@earth.sinica.edu.tw (J.-C. Lee), lizhaopku@pku.edu.cn (L. Zhao), bfchao@earth.sinica.edu.tw (B.F. Chao).

https://doi.org/10.1016/j.cageo.2022.105243
Received 25 August 2021; Received in revised form 26 September 2022; Accepted 27 September 2022
Available online 1 October 2022
0098-3004/

Mikesell et al., 2015). Hence, it can also be used to estimate the similarity between two different data series. The DTW distance, a measure for the dissimilarity between two time series, is estimated by finding the optimal match in the time series data by compressing and extending the time axis (Berndt and Clifford, 1994).

The DTW technique has been extensively explored in the field of speech and image recognition, data exploration, finances, medicine, engineering, environmental science, and other fields (Itakura, 1975; Sakoe and Chiba, 1978). In particular, Kumar et al. (2022) employed DTW for several seismological applications, including the clustering of seismic time series. It has also been successfully combined with the Hierarchical Agglomerative Clustering (HAC) analysis to cluster similar time series together and identify anomalous behavior (Huang and Jansen, 1985; Niennattrakul and Ratanamahatana, 2007; Kumar et al., 2022). However, combining DTW with HAC analysis has some difficulties as the time complexity of the DTW technique is quadratic (Salvador and Chan, 2007; Izakian et al., 2015). Nevertheless, several recent studies have employed an iterative approach in implementing the DTW technique that offers speed and other improvements (e.g., Shen et al., 2017). The present study aims to demonstrate the efficacy of the DTW-based HAC method on the spatiotemporal GNSS displacement field data in the Taiwan region.

2. GNSS data in Taiwan

Continuous GNSS networks in Taiwan, comprising over 400 stations operated by the Central Weather Bureau, Academia Sinica, and the Central Geological Survey, routinely collect data which are compiled and processed by the GPS Lab at the Institute of Earth Sciences of Academia Sinica (Chen et al., 2013) using the Bernese Global Navigation Satellite System software (Dach et al., 2015). The solutions are solved, relative to the tectonically stable station (S01R) of Penghu off the west coast, for the east (E), north (N), and up (U) components in the International Terrestrial Reference Frame (ITRF) Cartesian coordinates (see http://gps.earth.sinica.edu.tw for details).

We select 345 time records from 115 stations based on the consistency, quality, and data lengths (see Figs. S1 and S2 in SI), and discard those not characterizing common observational epochs (for details, see Kumar et al., 2020). We remove the outlier points exceeding 2σ of the mean variance and linearly-interpolate over the data gaps, and average them into daily solutions. Covering the period of 2007–2018 (4017 days), these GNSS data sometimes show spatially incoherent motions over relatively short time spans due to local secular changes. The long-duration data feature large degrees of freedom and contain valuable information about long-term behavior of the active tectonics.

3. Method: DTW-based HAC analysis

We developed a Python software package named *dtwhaclustering*, to ease the execution of the steps in the DTW distance-based HAC analysis. The workflow is as follows (see Fig. 1): After removing the least-squares fit of the seasonal and tidal signals (section 3.1), we apply our DTW-based HAC method to find the relative similarity and hence anomalous variability between the GNSS residual displacements in the target region along the eastern coast of Taiwan (section 3.2). Then we cluster and select the optimal number of clusters based on the relative changes in the DTW distance (section 4.3), and cross-validate the results with iterative spatiotemporal stability tests (section 4.4).

3.1. Least-squares model GNSS time series

We first model each GNSS time series of 3-D displacements (N, E, and U) as a superposition of the following terms:

$$f(t) = a + bt + \sum_k c_k H(t - t_0) + \sum_{k=1}^{6}\left[A_k \cos\left(\frac{2\pi T}{P_k}\right) + B_k \sin\left(\frac{2\pi T}{P_k}\right)\right] + Residual \quad (1)$$

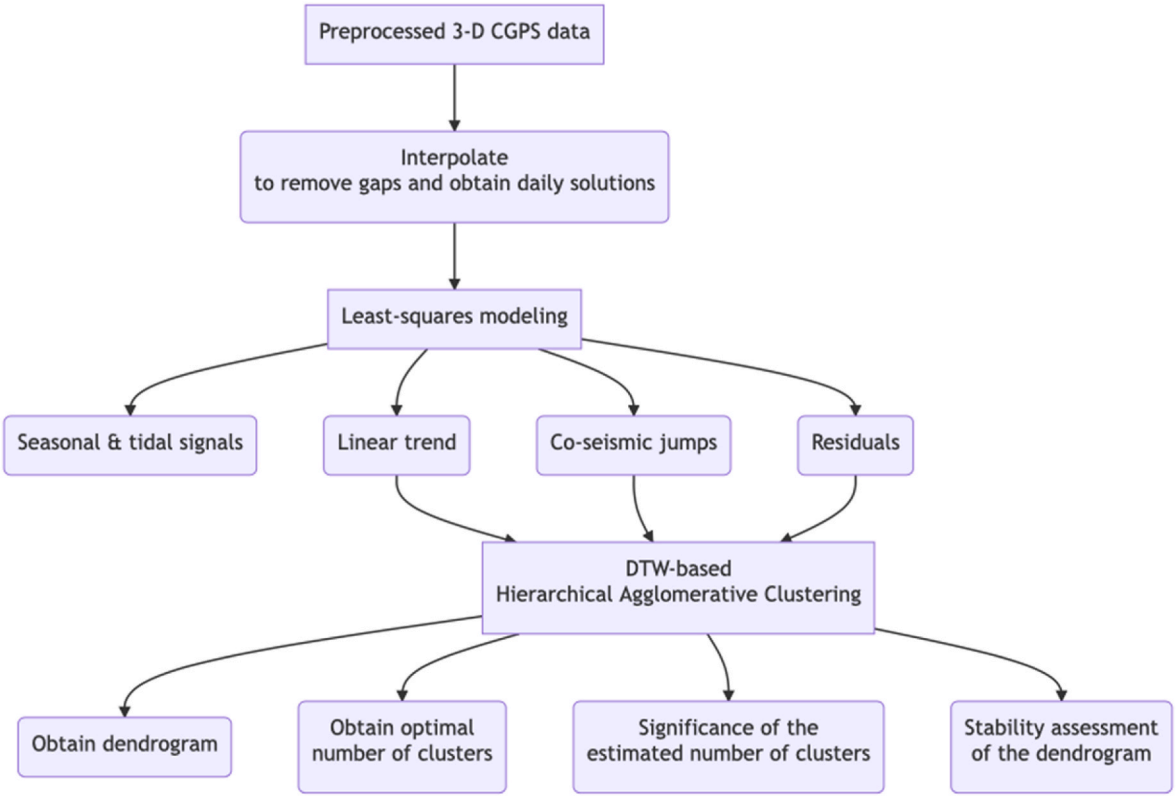


Fig. 1. Workflow of Hierarchical Agglomerative Clustering based on DTW distance of continuous GNSS displacements for tectonic investigation.

where, $a + bt$ accounts for the mean + linear trend in the time series, the Heaviside function$H(t - t_0)$ account for the coseismic and other unmodeled data jumps, and the six periodic variations accounts for the annual (365.26 days), semiannual (182.62 days), four tidal (with periods 13.6608 days, Mf; 14.7653 days, Msf; 27.5546 days, Mm; 18.6 years). Then the linear coefficients are least-squares estimated by minimizing the variance of the *Residual*. We mention in passing that some studies have suggested a more sophisticated approach of using Maximum Likelihood Estimation, singular spectrum analysis, and Monte Carlo Markov Chain (e.g., He et al., 2017). We used the least-squares regression in the simple tasks here for its inherent simplicity and flexibility in its implementation.

For example Fig. 2 displays the least-squares fits and the resultant residuals (by subtraction out the fits) at station ERPN. In addition, we have also removed the common-mode errors (CME) from the 3-D GNSS displacements via the estimation using Empirical Orthogonal Function technique (Weare and Nasstrom, 1982; Chang and Chao, 2014; Kumar et al., 2020). CME are dominant systematic errors exhibiting strong spatial coherence in the regional networks, which, if not treated properly may bias the final result.

Fig. 3 shows the geospatial distribution pattern of the linear trend estimated as above from the 3-D GNSS displacements over the whole Taiwan; the values are interpolated using the nearest interpolation algorithm to obtain the spatial pattern for Taiwan. The results are consistent with previous studies (Yu et al., 1999; Ching et al., 2011; Huang et al., 2015). The strong anthropogenic subsidence in the western coastal region (Liu et al., 2004; Tung and Hu, 2012) is clearly captured by the four stations: FUNY, VR02, PKGM, and HUWE. The stations in the eastern coastal region of Taiwan, particularly in the Longitudinal Valley, show significant northwestward motion relative to the S01R station. The northeastern part of Taiwan shows clockwise rotation, presumably due to the ongoing extensional crustal deformation related to the back-arc extension of the Okinawa Trough (e.g., Tsai et al., 2015).

3.2. The HAC analysis

For investigating the spatiotemporal behavior of the GNSS displacements in Taiwan, we build upon the HAC approach proposed by Simpson et al. (2012) for GNSS datasets. The HAC analysis is an unsupervised technique that is commonly used to lump together similar waveforms with bounding thresholds. The traditional HAC algorithm first projects the horizontal velocity data into a Euclidean velocity space (Savage and Simpson, 2013a, 2013b; Takahashi et al., 2019). It begins by assuming each time series as a cluster in itself, hence there are N clusters for N stations to begin with. Then it iteratively merges each pair of clusters by introducing the new cluster at the pair's centroid position in velocity space. Traditionally, the clusters are merged based on their measure of the Euclidean distance (the geometrical distance in velocity space), which represents the similarity between them. The graphical representation of the tree formed by the iterative merging of data subsets is called a dendrogram. The dendrogram is essential in investigating the relationships between the linkages of the clusters. We can observe a dendrogram from the bottom with N initial clusters at the bottom up to one big cluster at the top. Various clustering schemes share the above procedure as a common definition but differ in the way the similarity matrix is computed (Huang and Jansen, 1985; Bar-Joseph et al., 2001; Müllner, 2011).

The concrete choice of the similarity measure has a large influence on the outcome of the dendrogram of the HAC analysis. It can significantly affect the performance and effectiveness of the analysis. Although Euclidean distance is the most commonly used distance measure, there are many distance metrics available such as the Manhattan distance, Minkowski distance, Mahalanobis distance, etc., serving different purposes (Defays, 1977; Awasthi et al., 2012).

The HAC method based on the Euclidean distance in the horizontal velocity space takes into account only the GNSS horizontal velocity at each station while ignoring the vertical motion. It also leaves out temporal information in localized waveforms (Simpson et al., 2012). To study a tectonically complex region where the 2-D velocity vector may be inadequate to provide an accurate representation of the tectonics, and to incorporate the full waveform information cluster analysis, we now replace the Euclidean distance with the DTW distance measure.

3.3. DTW distance

The DTW distance quantifies the similarity between two time series with improved accuracy and efficiency by shifting and warping the time axis of the two series to obtain an optimum alignment (Berndt and Clifford, 1994; Senin, 2008). Traditionally, DTW analysis uses the dynamical programming technique by dividing a problem into subproblems and solving it recursively (Bellman and Dreyfus, 1962; Bellman, 1966). Recent studies implemented iterative approach and other upgrades that improved the computation speed and accuracy (Shen et al., 2017). For details on the DTW implementation, the readers are referred to previous relevant literature (Izakian et al., 2015; Mikesell et al., 2015).

The DTW distance is zero for two time series with completely similar

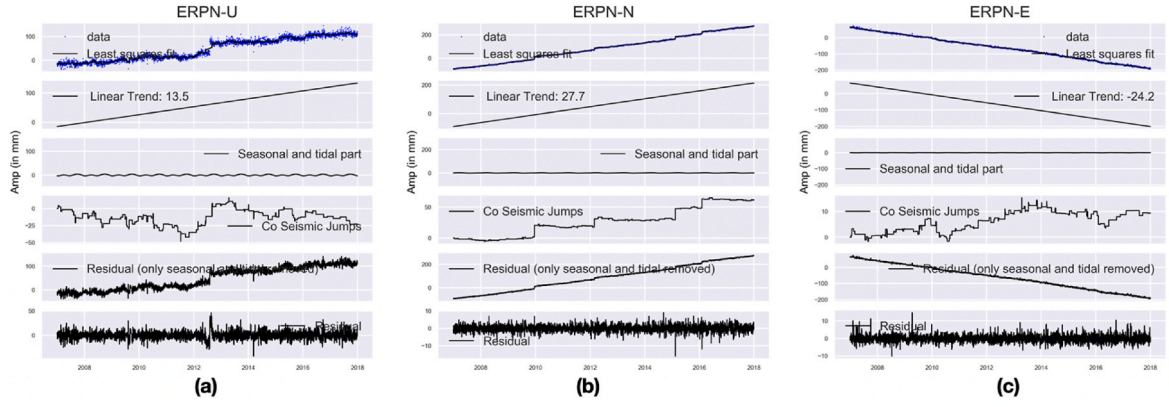


Fig. 2. Least-squares modeling for the daily GNSS time series (2007–2018) at station ERPN for (a) Vertical, (b) North, and (c) East components. The panels from the top down: the GNSS data points (blue dots) and the modeled least-squares fit (black solid line); the linear trend; estimated seasonal + tidal signals; modeled co-seismic jumps (note the different scales); GNSS residuals after removing the seasonal and tidal signals (to be used in the tectonics study after removing common-mode errors); GNSS residuals after subtracting linear trend, co-seismic jumps and seasonal + tidal signals. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

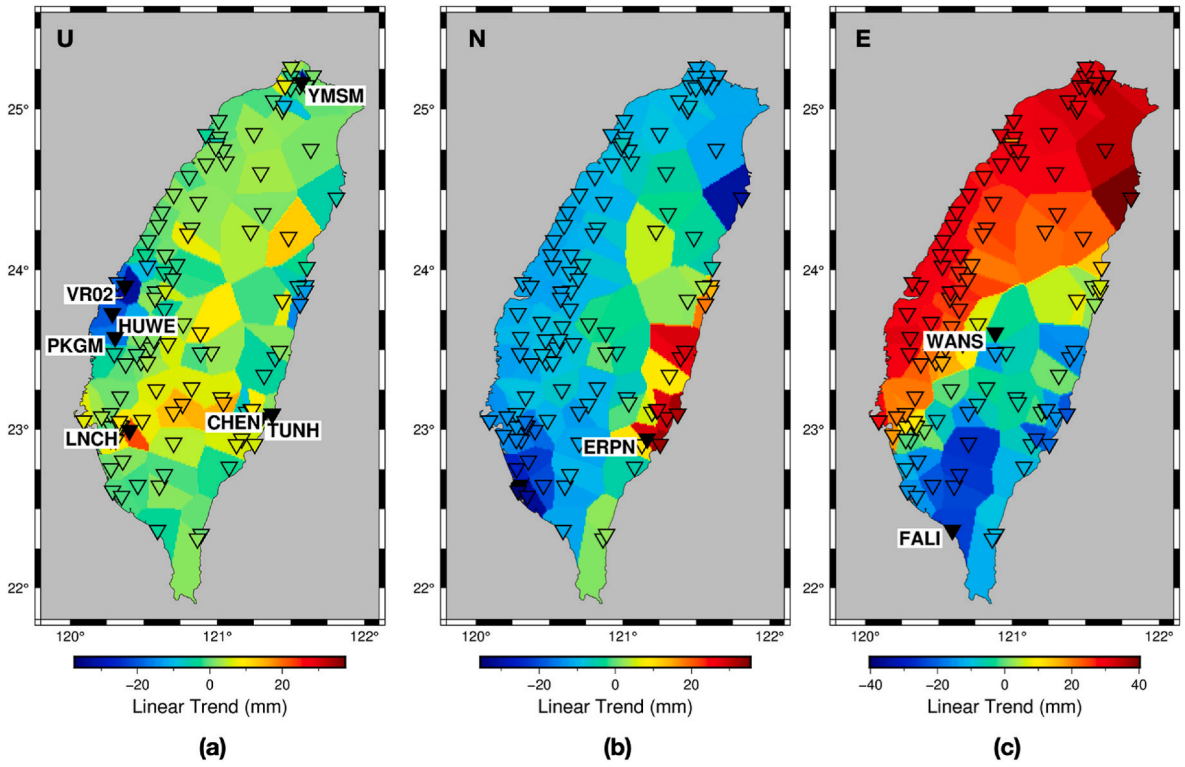


Fig. 3. Linear trends (background colors) obtained by least-squares estimation of the 11 years of GNSS displacements for (a) Vertical, (b) North, and (c) East components, respectively. The inverted triangles show the locations of GNSS stations. The figure is auto-generated using the *dtwhaclustering* package. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

patterns, while the higher DTW distance indicates an increase in the dissimilarity. Note that the DTW distance measures the cumulative similarity of the full length of the two time series, therefore for two time series having similar recording time, DTW distance can be used as a direct comparison measure. Given the similarity in the two approaches, the Euclidean distance can be easily replaced by the DTW distance, keeping other parts of the HAC algorithm unchanged. In the new HAC approach, we begin by N stations, and iteratively merge them based on their waveform similarity obtained by the DTW distance.

3.4. Synthetic tests

We validate our clustering algorithm by performing tests using synthetic data. As shown in Fig. S6 in SI, we first generate five time series based on known basic functions. Then, 15 time series are derived by making three copies each of the five synthetic signals and adding random noises (Fig. 4a) and assigned to arbitrarily defined locations in space (Fig. S7). The DTW-based clustering correctly identifies the base patterns present in the synthetic data (Fig. 4b), along with their correct spatial locations. The clustering algorithm successfully retrieves the 5 clusters as expected (Fig. 4c). The results from the DTW-based HAC are comparable to the Euclidean-based HAC analysis. However, the dendrogram of the former accounts for more of the short-term temporal variations in the waveforms.

4. Results

4.1. Clustering of the continuous GNSS displacements in Taiwan

We apply the DTW-based HAC analysis independently to each component of the 115 selected continuous GNSS displacements of 11 years (2007–2018). Fig. 5 displays the polar chart of the dendrogram with the similarity relations between the clusters given in terms of the DTW distance. Radial lines and concentric arcs in the dendrogram correspond to the vertical and lateral lines, respectively, in the conventional dendrogram (see SI). Given around the circular edge of each dendrogram are the station names, and the numbers in the dendrograms are the DTW distance values. The colors of the lines in the dendrogram are set to emphasize the optimal number of parameters obtained using the elbow method (see next section). Notice that the dendrogram of the vertical component is dominated by a few stations. This is mostly due to the anomalous subsidence in the western region of Taiwan (Fig. 3).

4.2. Estimating the optimal number of clusters

Clustering analysis consists both of grouping the data based on some criterion and cross-validating the resulting groups. There are several approaches in data science that are commonly used to extract the optimal number of clusters, such as the elbow method, the gap-statistic method, silhouette analysis, etc. (Breckenridge, 2000; Tibshirani et al., 2001; Wang, 2010; Kawamoto and Kabashima, 2017; Fu and Perry, 2020). The elbow method is the most well-known, in which the percentage of variance (or DTW similarity measure in this study) as a function of the number of clusters is calculated and graphed, and the maxima of the first derivative of the function (elbows) are looked for to determine the optimal number of clusters. The elbow method is based on the idea that adding a greater number of clusters than the selected optimal one does not significantly improve the modelling of the data.

In this study, we implement the elbow method to obtain the optimal number of clusters given by the maximum curvature of the DTW distance. Based on the maximum relative difference in DTW distance (computed using the elbow method) between the clusters in the dendrogram in Fig. 5, we have identified five clusters for the U, five clusters for the N, and three clusters for the E as the optimal number of clusters.

It is also important to note that a few anomalous waveforms in the analysis with the least similarity (hence large DTW distance) in relation to other waveforms can alter the estimated optimal number of clusters

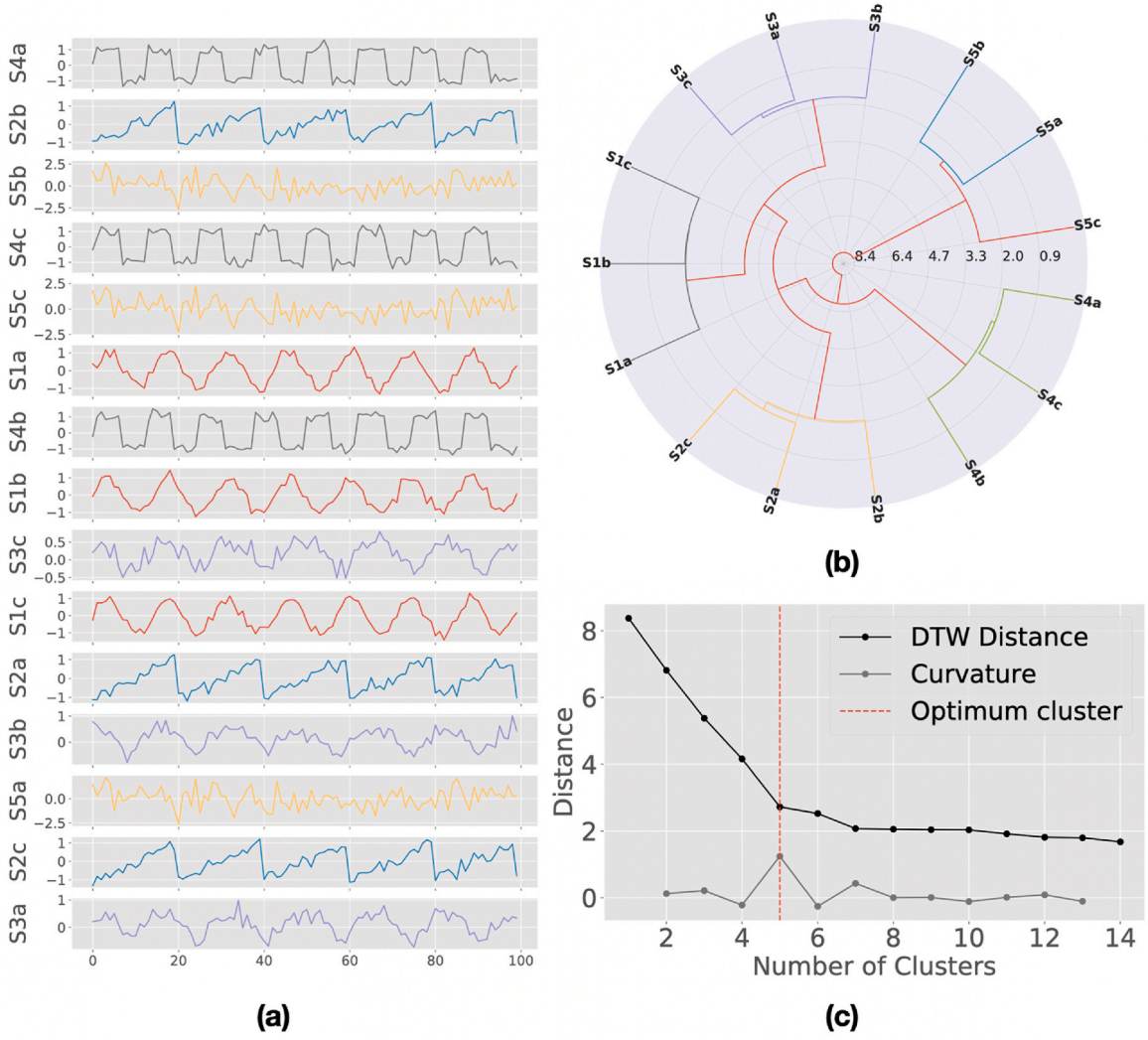


Fig. 4. (a) Shuffled signals for the clustering obtained by making three copies each (a, b, c) of the original five signals and adding random noise. (b) Polar dendrogram showing the results of the DTW-based HAC clustering. (c) Estimation of the number of optimum clusters in the dataset through the Elbow method.

using the elbow method. This seems apparent in Figs. 5 and 6. For example, the vertical component at stations PKGM, YMSM, LNCH and the north component at ERPN feature anomalous waveform fluctuations and hence are assigned separate clusters by the HAC analysis. Many times, we would like to identify the anomalous stations using the clustering analysis and isolate them from other clusters to further examine the relation between other waveforms in more detail. Alternatively, we can use higher-order curvature in the analysis for the optimal number of clusters.

4.3. Significance assessment of the estimated number of clusters

We implement the non-parametric approach where we perform the Monte-Carlo simulation-based hypothesis testing to assess the statistical significance of the obtained optimal number of clusters. This approach is similar to that of the gap-statistic test (Tibshirani et al., 2001). We assume the null hypothesis that the waveforms at each station are obtained from a random distribution, and it does not lead to any clusters of interest. Then, using the Monte-Carlo simulations of DTW-based HAC on randomly shuffled original waveforms for N times (where N is large), we compute the likelihood (the p-value) of obtaining k clusters given by the elbow method. If the p-value for k clusters is less than the decided threshold (e.g., $p < 0.05$), then we reject the null hypothesis and accept the number k computed using the elbow method as the optimal number of clusters.

Fig. 7 shows the results of the Monte-Carlo simulations of the HAC analysis for $N = 100$ to assess the significance of the estimated number of clusters for each component. The error bar shows the 95% confidence interval estimated by 100 Monte-Carlo simulations of the HAC analysis on the randomly shuffled original waveforms for the 115 stations. In each plot, the black line shows the curvature of the relative DTW distance at an increasing number of clusters (or hierarchy in the dendrogram). We can see that for the randomly shuffled waveforms, the optimal number of clusters tends to be the same as the number of stations to begin with. Hence, the obtained optimal clusters (indicated by the green dashed line) for the three components are statistically significant.

4.4. Stability assessment of the dendrogram in the presence of noise

The GNSS data can contain observational noise due to global and local disturbances, including orbit information errors and water vapor anomalies. Hence, it is essential to examine whether our clustering analysis results are trustworthy with the effects of noise taken into consideration. Following Takahashi et al. (2019), we begin by adding synthetic noise according to the observed noise levels to the 115 sets of three-component GNSS data as follows:

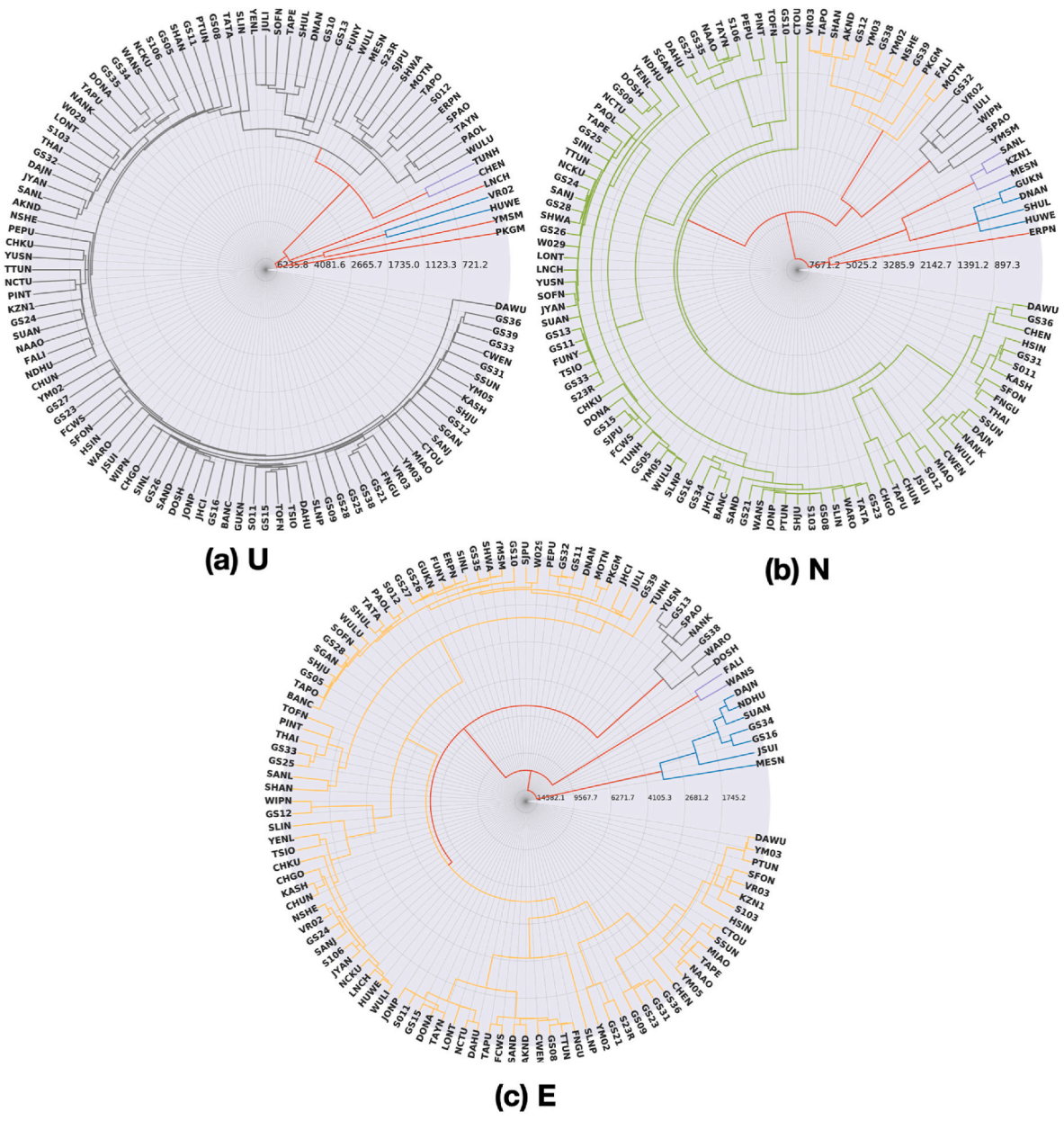


Fig. 5. Resulting dendrogram for the HAC method as a polar chart for the (a) vertical, (b) north, and (c) east component GNSS residuals of Taiwan. The figure is auto-generated using the *dtwhaclustering* package.

$$d_{i,m} = d_{i,ref} + \varepsilon(\sigma_i)_m \tag{2}$$

where $d_{i,m}$ denotes synthesized GNSS data at a station i ($i = 1,\ldots,115$) for a given component, m denotes the synthesized data index, and $d_{i,ref}$ is the original daily fluctuations at the same station. σ_i is the standard deviation of GNSS velocity, and $\varepsilon(\sigma_i)_m$ is the synthetic error obtained from the Gaussian distribution with zero mean and the standard deviation σ_i. We feed the synthetic GNSS data for the 115 stations used in our original analysis to our clustering algorithm and examine the effects of noise. The relational structure among the clusters in the output dendrogram changes by less than 10% for the standard deviation of up to 2σ. When the noise levels are further increased, the relational structure among the cluster starts to break down significantly and the results become unreliable (see Fig. S5 in the SI). This shows that the DTW-based HAC analysis is robust in the presence of random noise in the data.

5. Discussion and conclusions

In this study, we have combined the DTW distance measure and HAC to develop a new approach to clustering analysis. A Python software package, *dtwhaclustering*, has been designed to ease the execution of the procedures in the DTW distance-based HAC analysis. The algorithm was validated using synthetic datasets to ensure better performance in comparison to traditional cluster approaches.

We have applied our DTW distance-based HAC approach to 11-year (2007–2018) records of three-component GNSS displacements at 115 stations in Taiwan. We first conducted least-squares modeling of GNSS daily solutions to inspect the dominant linear trend and remove seasonal variations. Then, we applied the DTW-based clustering to obtain the optimal clusters for the 115 stations in Taiwan. These clusters are consistent with the patterns of the linear trend and the known geology in the region. We also cross-validated the results for spatial and temporal stability.

The hierarchical order obtained from waveform-based objective

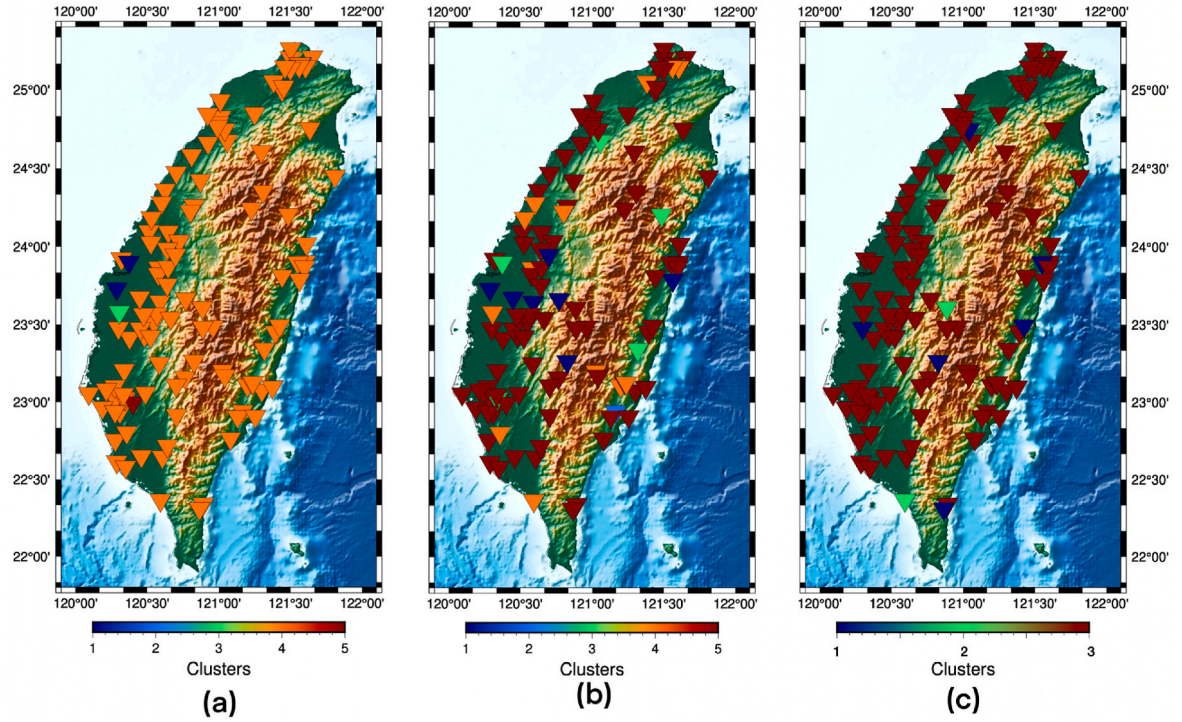


Fig. 6. Clustering results with geographical locations for the (a) vertical, (b) north, and (c) east components showing the optimal number of parameters estimated by the elbow method. The figure is auto-generated using the *dtwhaclustering* package. The color scale is optimized to show the maximum differences for each component. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

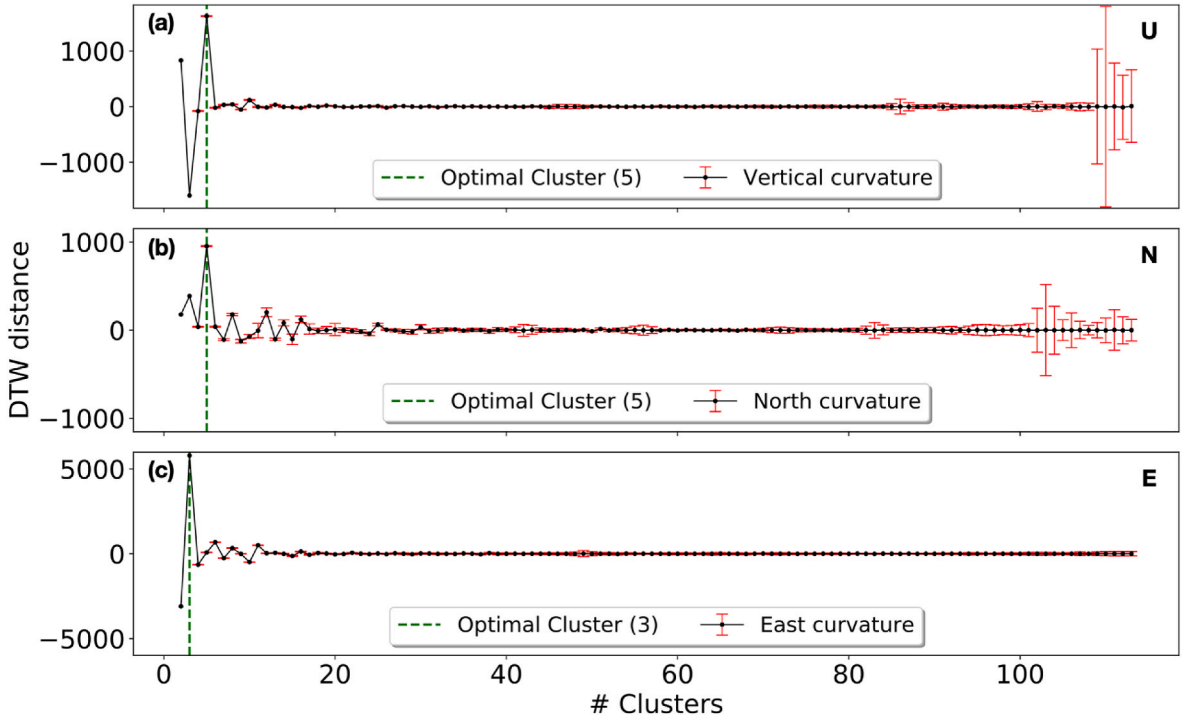


Fig. 7. Estimation of the optimal number of clusters for the HAC analysis based on DTW for (a) vertical, (b) north, and (c) east components. The maximum change in DTW distance (or the similarity between stations) is indicated by the dashed green line. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

clustering methods, such as the DTW-based HAC analysis developed in this study, is likely to be associated with the characteristics of crustal deformation. The resulting clusters are consistent with the geospatial patterns obtained from the linear trend analysis on the same dataset (see Fig. 3). However, linear trend analysis only extracts the first-order variations from the geodetic time series. In contrast, the DTW analysis gives more insight into the short-term temporal behavior of the GNSS stations.

The DTW distance computation allows multidimensional and three-component datasets to be bound together at each station using the N x 3 array, where N is the number of points for each component. Binding the three components together can be helpful in detecting stations with anomalous three-component motion. In this manuscript, for the sake of simplicity, we only analyze the clustering for each component independently.

The hierarchical representation obtained from the DTW-based HAC provides quantitative distance values among individual stations and the clusters, which can be physically interpreted as their mutual dissimilarity. We have cross-validated our clustering results for both stability and reliability, as it is necessary to ensure that the obtained hierarchical relationships among the stations are robust in the presence of noise. The most significant level of the hierarchy, the optimum clusters, contains 5 clusters for both vertical and north components, and three clusters for the east component. The obtained clusters are in good agreement with the geographical/geological locations of the stations, and the stations relatively close to each other tend to have higher similarity and fall in the same cluster. There are some lower levels of hierarchy that also exhibit significant changes in similarity, which can be explored to further study more detailed tectonics in the region.

The most anomalous waveform behavior can be seen in the vertical component at stations in the western coastal region near the city of Taichung (PKGM, YMSM, LNCH, VR02, and HUWE). These stations are known to have anomalous subsidence, which is also reflected in the estimated linear trend (Fig. 3). Although these stations have similar overall subsidence behavior, it is interesting to see that the waveform dissimilarities among those stations do not allow them to be grouped into the same cluster.

The clustering analysis results reveal strongly incoherent crustal motions at various points in the Longitudinal Valley (Fig. 6), exhibiting a strong effect of the collision between the Philippine Sea Plate and the Eurasian plate. Several GNSS measurements along with tide-gauge/altimetry studies (Shui-Beih et al., 1990; Nikolaidis and Bock, 2002; Shyu et al., 2006; Kuo-Chen et al., 2012; Chen et al., 2013; Shin et al., 2013) have found rapid anomalous elevation changes in the Longitudinal Valley.

The clusters at higher hierarchical levels of the HAC analysis (except for the anomalous outlier stations) represent major crustal velocity discontinuities, presumably correspond to major crustal tectonic/geological block boundaries or large local deformation sources (Simpson et al., 2012; Takahashi et al., 2019). However, the DTW-based HAC also considers the localized temporal fluctuations in the GNSS data, and hence the resultant clusters are dependent on the cumulative tectonic motion beneath each station for the selected period of the waveforms (2007–2018). The full waveform approach of the DTW-based HAC method is effective in the identification of anomalous stations, outliers, or abnormal behavior in the region, which would be difficult to achieve using the traditional Euclidean distance-based HAC.

Although the DTW distance-based HAC results can distinguish the subtle crustal differences at a pair of stations and can ameliorate certain problems in the traditional clustering approach, it is sensitive to the length and quality of the GNSS data. In this study, we have assumed that the GNSS displacement variations is only influenced by the geological settings. However, there may be other effects such as random errors from instruments, etc.

DTW can only distinguish between time series slightly different in frequency, amplitude, or initial phase (Huang and Jansen, 1985), which is the case in most geophysical applications. In this study, we have used longer GNSS data to moderate the data quality further, as DTW distance tends to give more weight to the dominant patterns.

The DTW-based clustering analysis can be effective in inspecting the anomalies, outliers, or abnormal behaviors of the GNSS displacements that takes into account the fluctuations of the full-waveforms.

Authorship contribution statement

Utpal Kumar: Code development, numerical calculations, and manuscript writing.

Cédric. P. Legendre, Jian-Cheng Lee, Li Zhao, Benjamin Fong Chao: Technical and data support and manuscript revision.

Computer code availability

dtwhaclustering is an open-source Python package developed under the Apache License, Version 2.0 (the "License"). The package can be download from PyPi: *pip install dtwhaclustering*. For the documentation on the *dtwhaclustering* package, visit: https://dtwhaclustering.readthedocs.io/en/latest/. Jupyter Notebooks for the complete analysis of this study can be downloaded from https://github.com/earthinversion/Dynamic-Time-Warping-based-Hierarchical-Agglomerative-Clustering.git.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The GNSS data used for this study is obtained from the GPS lab of the Institute of Earth Sciences (http://gps.earth.sinica.edu.tw). The DTW-based clustering implementation is performed after the modification of the DTW analysis package provided by DTAI Research Group (Github Repository: https://github.com/wannesm/dtaidistance). We thank Dr. Keogh for his constructive comments on the manuscript. This work was funded by the National Science Council of Taiwan under grants: MOST 109-2116-M-001-028, 108-2116-M-001-010-MY3.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cageo.2022.105243.

References

Angeles-Yreta, A., Solís-Estrella, H., Landassuri-Moreno, V., Figueroa-Nazuno, J., Solís-Estrella, H., Landassuri-Moreno, V., Figueroa-Nazuno, J., 2004. Similarity search in seismological signals. In: Proceedings of the Fifth Mexican International Conference in Computer Science. ENC 2004 KW -, IEEE, pp. 50–56. https://doi.org/10.1109/ENC.2004.1342588, 2004.

Angelier, J., Chang, T.-Y., Hu, J.-C., Chang, C.-P., Siame, L., Lee, J.-C., Deffontaines, B., Chu, H.-T., Lu, C.-Y., 2009. Does extrusion occur at both tips of the Taiwan collision belt? Insights from active deformation studies in the Ilan Plain and Pingtung Plain regions. Tectonophysics 466 (3–4), 356–376. https://doi.org/10.1016/j.tecto.2007.11.015.

Awasthi, P., Blum, A., Sheffet, O., 2012. Center-based clustering under perturbation stability. Inf. Process. Lett. 112 (1–2), 49–54. https://doi.org/10.1016/j.ipl.2011.10.006.

Bar-Joseph, Z., Gifford, D.K., Jaakkola, T.S., 2001. Fast optimal leaf ordering for hierarchical clustering. Bioinformatics 17 (1). https://doi.org/10.1093/bioinformatics/17.suppl_1.S22. S22–S29.

Bellman, R., 1966. Dynamic programming. Science 153 (3731), 34–37. https://doi.org/10.1126/science.153.3731.34, 80-.

Bellman, R.E., Dreyfus, S.E., 1962. Applied Dynamic Programming. Princeton University Press. https://doi.org/10.1515/9781400874651.

Berndt, D., Clifford, J., 1994. Using dynamic time warping to find patterns in time series. In: Workshop on Knowledge Knowledge Discovery in Databases, pp. 359–370.

Breckenridge, J.N., 2000. Validating cluster Analysis: consistent replication and symmetry. Multivariate Behav. Res. 35 (2), 261–285. https://doi.org/10.1207/S15327906MBR3502_5.

Chang, E.T.Y., Chao, B.F., 2014. Analysis of coseismic deformation using EOF method on dense, continuous GPS data in Taiwan. Tectonophysics 637 (C), 106–115. https://doi.org/10.1016/j.tecto.2014.09.011.

Chang, E.T.Y.Y., Chao, B.F., Chiang, C.-C.C., Hwang, C., 2012. Vertical crustal motion of active plate convergence in Taiwan derived from tide gauge, altimetry, and GPS data. Tectonophysics 578 (C), 98–106. https://doi.org/10.1016/j.tecto.2011.10.002.

Chen, H.-Y.Y., Lee, J.-C.C., Tung, H., Yu, S.-B.B., Hsu, Y.-J.J., Lee, H., 2013. A new velocity field from a dense GPS array in the southernmost Longitudinal Valley, southeastern taiwan. Terr. Atmos. Ocean Sci. 24 (5), 837. https://doi.org/10.3319/TAO.2013.06.18.01 (T).

Ching, K.-E., Rau, R.-J., Johnson, K.M., Lee, J.-C., Hu, J.-C., 2011. Present-day kinematics of active mountain building in Taiwan from GPS observations during 1995-2005. J. Geophys. Res. Solid Earth 116. https://doi.org/10.1029/2010JB008058. B9.

Dach, R., Lutz, S., Walser, P., Fridez, P., 2015. Bernese GNSS Software Version 5.2. University of Bern, Bern Open Publishing.

Defays, D., 1977. An efficient algorithm for a complete link method. Comput. J. 20 (4), 364–366. https://doi.org/10.1093/comjnl/20.4.364.

Fu, W., Perry, P.O., 2020. Estimating the number of clusters using cross-validation. J. Comput. Graph Stat. 29 (1), 162–173. https://doi.org/10.1080/10618600.2019.1647846.

Hale, D., 2013. Dynamic warping of seismic images. Geophysics 78 (2), S105–S115. https://doi.org/10.1190/geo2012-0327.1.

He, X., Montillet, J.P., Fernandes, R., Bos, M., Yu, K., Hua, X., Jiang, W., 2017. Review of current GPS methodologies for producing accurate time series and their error sources. J. Geodyn. 106, 12–29. https://doi.org/10.1016/j.jog.2017.01.004.

Huang, H.-C., Jansen, B.H., 1985. EEG waveform analysis by means of dynamic time-warping. Int. J. Bio Med. Comput. 17 (2), 135–144. https://doi.org/10.1016/0020-7101(85)90084-4.

Huang, H.H., Lin, F.C., Schmandt, B., Farrell, J., Smith, R.B., Tsai, V.C., 2015. The Yellowstone magmatic system from the mantle plume to the upper crust. Science 348 (6236), 773–776. https://doi.org/10.1126/science.aaa5648, 80-.

Itakura, F., 1975. Minimum prediction residual principle applied to speech recognition. IEEE Trans. Acoust. 23 (1), 67–72. https://doi.org/10.1109/TASSP.1975.1162641.

Iwata, T., Umeno, K., 2017. Preseismic ionospheric anomalies detected before the 2016 Kumamoto earthquake. J. Geophys. Res. Sp. Phys. 122 (3), 3602–3616. https://doi.org/10.1002/2017JA023921.

Izakian, H., Pedrycz, W., 2013. Anomaly detection in time series data using a fuzzy c-means clustering. In: 2013 Joint IFSA World Congress and NAFIPS Annual Meeting. IFSA/NAFIPS, pp. 1513–1518.

Izakian, H., Pedrycz, W., Jamal, I., 2015. Fuzzy clustering of time series data using dynamic time warping distance. Eng. Appl. Artif. Intell. 39, 235–244. https://doi.org/10.1016/j.engappai.2014.12.015.

Kawamoto, T., Kabashima, Y., 2017. Cross-validation estimate of the number of clusters in a network. Sci. Rep. 7 (1), 3327. https://doi.org/10.1038/s41598-017-03623-x.

Kumar, U., Chao, B.F., Chang, E.T.-Y.Y., 2020. What causes the common-mode error in array GPS displacement fields: case study for taiwan in relation to atmospheric mass loading. Earth Space Sci. https://doi.org/10.1029/2020ea001159, 0–2.

Kumar, U., Legendre, C.P., Zhao, L., Chao, B.F., 2022. Dynamic time warping as an alternative to windowed cross correlation in seismological applications. Seismol Res. Lett. https://doi.org/10.1785/0220210288.

Kuo-Chen, H., Wu, F.T., Roecker, S.W., 2012. Three-dimensional P velocity structures of the lithosphere beneath Taiwan from the analysis of TAIGER and related seismic data sets. J. Geophys. Res. Solid Earth 117 (B6). https://doi.org/10.1029/2011JB009108.

Liu, C.-H., Pan, Y.-W., Liao, J.-J., Huang, C.-T., Ouyang, S., 2004. Characterization of land subsidence in the Choshui River alluvial fan, Taiwan. Environ. Geol. 45 (8), 1154–1166. https://doi.org/10.1007/s00254-004-0983-6.

Mikesell, T.D., Malcolm, A.E., Yang, D., Haney, M.M., 2015. A comparison of methods to estimate seismic phase delays: numerical examples for coda wave interferometry. Geophys. J. Int. 202 (1), 347–360. https://doi.org/10.1093/gji/ggv138.

Mohapatra, S.S., Bhuyan, P.K., V Rao, K., 2012. Genetic algorithm fuzzy clustering using GPS data for defining level of service criteria of urban streets. Eur. Transp. Eur. 49 (52), 1–19.

Müllner, D., 2011. Modern Hierarchical, Agglomerative Clustering Algorithms. *arXiv Prepr. arXiv1109.2378*.

Niennattrakul, V., Ratanamahatana, C.A., 2007. On clustering multimedia time series data using k-means and dynamic time warping. In: Proceedings - 2007 International Conference on Multimedia and Ubiquitous Engineering. MUE 2007, IEEE, pp. 733–738. https://doi.org/10.1109/MUE.2007.165.

Nikolaidis, R.M., Bock, Y., 2002. Observation of geodetic and seismic deformation with the global positioning system. Earth Sci. 265 https://doi.org/10.1029/2001JB000329.

Rau, R.-J., Ching, K.-E., Hu, J.-C., Lee, J.-C., 2008. Crustal deformation and block kinematics in transition from collision to subduction: global positioning system measurements in northern Taiwan, 1995-2005. J. Geophys. Res. 113, B09404 https://doi.org/10.1029/2007JB005414. B9.

Sakoe, H., Chiba, S., 1978. Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust. 26 (1), 43–49. https://doi.org/10.1109/TASSP.1978.1163055.

Salvador, S., Chan, P., 2007. Toward accurate dynamic time warping in linear time and space. Intell. Data Anal. 11 (5), 561–580. https://doi.org/10.3233/ida-2007-11508.

Savage, J.C., Simpson, R.W., 2013a. Clustering of GPS velocities in the mojave block, southeastern California. J. Geophys. Res. Solid Earth 118 (4), 1747–1759. https://doi.org/10.1029/2012JB009699.

Savage, J.C., Simpson, R.W., 2013b. Clustering of velocities in a GPS network spanning the sierra Nevada block, the northern walker lane belt, and the central Nevada seismic belt, California-Nevada. J. Geophys. Res. Solid Earth 118 (9), 4937–4947. https://doi.org/10.1002/jgrb.50340.

Senin, P., 2008. Dynamic time warping algorithm review. Inf. Comput. Sci. Dep. Univ. Hawaii Manoa Honolulu, USA 855 (1–23), 40.

Shen, J., Huang, W., Zhu, D., Liang, J., 2017. A novel similarity measure model for multivariate time series based on LMNN and DTW. Neural Process. Lett. 45 (3), 925–937. https://doi.org/10.1007/s11063-016-9555-5.

Shin, T.C., Chang, C.H., Pu, H.C., Lin, H.W., Leu, P.L., 2013. The geophysical database management system in Taiwan. Terr. Atmos. Ocean Sci. 24 (1), 11–18. https://doi.org/10.3319/TAO.2012.09.20.01(T).

Shui-Beih, Y., Jackson, D.D., Guey-Kuen, Y., Chi-Ching, L., 1990. Dislocation model for crustal deformation in the Longitudinal Valley area, eastern Taiwan. Tectonophysics 183 (1–4), 97–109. https://doi.org/10.1016/0040-1951(90)90190-J.

Shyu, J.B.H., Sieh, K., Avouac, J.-P., Chen, W.-S., Chen, Y.-G., 2006. Millennial slip rate of the Longitudinal Valley fault from river terraces: implications for convergence across the active suture of eastern Taiwan. J. Geophys. Res. 111, B08403 https://doi.org/10.1029/2005JB003971. B8.

Simpson, R.W., Thatcher, W., Savage, J.C., 2012. Using cluster analysis to organize and explore regional GPS velocities. Geophys. Res. Lett. 39 (18), 1–5. https://doi.org/10.1029/2012GL052755.

Strle, B., Mozina, M., Bratko, I., 2009. Qualitative approximation to Dynamic Time Warping similarity between time series data. Proc. QR. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.611.3754&rep=rep1&type=pdf.

Takahashi, A., Hashimoto, M., Hu, J., Takeuchi, K., Tsai, M., Fukahata, Y., 2019. Hierarchical cluster Analysis of dense GPS data and examination of the nature of the clusters associated with regional tectonics in taiwan. J. Geophys. Res. Solid Earth 124 (5), 5174–5191. https://doi.org/10.1029/2018JB016995.

Thatcher, W., 2009. How the continents deform: the evidence from tectonic geodesy. Annu. Rev. Earth Planet Sci. 37 (1), 237–262. https://doi.org/10.1146/annurev.earth.031208.100035.

Tibshirani, R., Walther, G., Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic. J. R. Stat. Soc. Ser. B (Statistical Methodol. 63 (2), 411–423. https://doi.org/10.1111/1467-9868.00293.

Tsai, M.-C., Yu, S.-B., Shin, T.-C., Kuo, K.-W., Leu, P.-L., Chang, C.-H., Ho, M.-Y., 2015. Velocity field derived from taiwan continuous GPS array (2007-2013). Terr. Atmos. Ocean Sci. 26 (5), 527. https://doi.org/10.3319/TAO.2015.05.21.01 (T).

Tung, H., Hu, J.-C., 2012. Assessments of serious anthropogenic land subsidence in yunlin county of central taiwan from 1996 to 1999 by persistent scatterers InSAR. Tectonophysics 578, 126–135. https://doi.org/10.1016/j.tecto.2012.08.009.

Venstad, J.M., 2013. Dynamic time warping - an improved method for 4D and tomography time shift estimation? Geophysics 79 (5), R209–R220. https://doi.org/10.1190/GEO2013-0239.1.

Wang, J., 2010. Consistent selection of the number of clusters via crossvalidation. Biometrika 97 (4), 893–904. https://doi.org/10.1093/biomet/asq061.

Weare, B.C., Nasstrom, J.S., 1982. Examples of extended empirical orthogonal function analyses. Mon. Weather Rev. 110 (6), 481–485. https://doi.org/10.1175/1520-0493(1982)110<0481:EOEEOF>2.0.CO;2.

Wu, F.T., Liang, W.-T., Lee, J.-C., Benz, H., Villasenor, A., 2009. A model for the termination of the Ryukyu subduction zone against Taiwan: a junction of collision, subduction/separation, and subduction boundaries. J. Geophys. Res. 114, B07404 https://doi.org/10.1029/2008JB005950. B7.

Xia, Y., Pan, S., Meng, X., Gao, W., Ye, F., Zhao, Q., Zhao, X., 2020. Anomaly detection for urban vehicle GNSS observation with a hybrid machine learning system. Rem. Sens. 12 (6), 971. https://doi.org/10.3390/rs12060971.

Yu, S.-B.B., Chen, H.-Y.Y., Kuo, L.-C.C., 1997. Velocity field of GPS stations in the Taiwan area. Tectonophysics 274 (1–3), 41–59. https://doi.org/10.1016/S0040-1951(96)00297-1.

Yu, S.B., Kuo, L.C., Punongbayan, R.S., Ramos, E.G., 1999. GPS observation of crustal deformation in the Taiwan-Luzon region. Geophys. Res. Lett. 26 (7), 923–926. https://doi.org/10.1029/1999GL900148.