Geophysical Journal International

*Geophys. J. Int.* (2022) **230,** 2073–2088
Advance Access publication 2022 May 6
GJI Seismology

https://doi.org/10.1093/gji/ggac160

# 3-D spatial cluster analysis of seismic sequences through density-based algorithms

Ester Piegari, Marcus Herrmann and Warner Marzocchi

*Dipartimento di Scienze della Terra, dell'Ambiente e delle Risorse, Università degli Studi di Napoli Federico II, 80126 Naples, Italy. E-mail: Piegari ester.piegari@unina.it*

Accepted 2022 April 22. Received 2022 April 3; in original form 2022 January 10

## SUMMARY

With seismic catalogues becoming progressively larger, extracting information becomes challenging and calls upon using sophisticated statistical analysis. Data are typically clustered by machine learning algorithms to find patterns or identify regions of interest that require further exploration. Here, we investigate two density-based clustering algorithms, DBSCAN and OPTICS, for their capability to analyse the spatial distribution of seismicity and their effectiveness in discovering highly active seismic volumes of arbitrary shapes in large data sets. In particular, we study the influence of varying input parameters on the cluster solutions. By exploring the parameter space, we identify a crossover region with optimal solutions in between two phases with opposite behaviours (i.e. only clustered and only unclustered data points). Using a synthetic case with various geometric structures, we find that solutions in the crossover region consistently have the largest clusters and best represent the individual structures. For identifying strong anisotropic structures, we illustrate the usefulness of data rescaling. Applying the clustering algorithms to seismic catalogues of recent earthquake sequences (2016 Central Italy and 2016 Kumamoto) confirms that cluster solutions in the crossover region are the best candidates to identify 3-D features of tectonic structures that were activated in a seismic sequence. Finally, we propose a list of recipes that generalizes our analyses to obtain such solutions for other seismic sequences.

**Key words:** Machine learning; Statistical methods; Seismicity and tectonics; Statistical seismology.

## 1 INTRODUCTION

In recent years, machine learning algorithms have been increasingly used in many different research fields due to the availability of large data sets and new software tools. Clustering is a type of unsupervised machine learning (Mehta *et al*. 2019; Bhattacharya 2021; Zhang *et al.* 2022) that groups data by means of a similarity measure. In the last decades, many clustering algorithms based on different similarity measures have been proposed (Kaufman & Rousseeuw 1990; Jain *et al.* 1999) and applied to a variety of scientific problems (Aggarwal & Reddy 2013; Lyra *et al.* 2014; Lindsey *et al.* 2018; Karpatne *et al.* 2019; Abdideh & Ameri 2020) with the aim of identifying hidden patterns in data. Regarding applications to seismicity, a fuzzy clustering algorithm was used to partition earthquake epicentres of Iranian seismic catalogues (Ansari *et al.* 2009), while approaches based on k-means (Ouillon *et al.* 2008), Gaussian Mixture models (Ouillon & Sornette 2011) and more recently agglomerative hierarchical clustering (Kamer *et al.* 2020) have been proposed for fault network reconstruction. Furthermore, Konstantaras *et al.* (2012), Schoenball & Ellsworth (2017) and Fan & Xu (2019) have applied the density-based (DB) algorithm DBSCAN for cluster analyses of earthquake epicentres, while Cesca *et al.* (2014), Cesca (2020) and Petersen *et al.* (2021) have developed a software tool based on DBSCAN for implementing multidimensional clustering that accounts for other properties (such as origin, times, focal mechanisms, moment tensors and waveform similarity).

The choice of the most appropriate clustering algorithm depends on the application at hand and is related to the definition of a cluster. Clusters are commonly identified either as groups of data that minimize the intracluster distance (and maximize intercluster distance) or as dense data regions separated by sparse regions. Here, we are interested in discovering spatial features of seismicity by density rather than distances between data points. This decision is crucial to identify clusters of arbitrary shapes and anisotropic structures in a 3-D space. Partitioning algorithms like k-means or Gaussian Mixture models instead minimize the distances between data points, which generally leads to identify convex (i.e. spherical) regions around denser groups of data points. Instead, DB connections among data points allow recognizing preferential alignments of anisotropic structures and provide information about their size (Ester *et al.* 1996). Another advantage of DB algorithms is their efficiency on large data sets compared to hierarchical clustering algorithms. Furthermore, DB clustering does not require every data point to be part of a cluster, which makes it possible to account for noise in data.

In the following, we will explore the two most popular DB clustering algorithms, DBSCAN (Ester *et al.* 1996) and its extension OPTICS (Ankerst *et al.* 1999). They are based on a simple set of instructions and require only two input parameters. The problem is that depending on the spatial distribution of earthquakes, even small changes of these parameters can lead to very different cluster solutions, ranging from many small to very few large clusters. For this reason, we explore the challenges in the calibration of these procedures to obtain stable cluster solutions. We deal with this sensitivity aspect by first exploring the whole parameter space and then discussing DB cluster solutions for different catalogues. Specifically, we perform cluster analyses of earthquake catalogues of the 2016 Kumamoto and 2016 Central Italy sequence and identify their main spatial features. Finally, on the basis of the findings from clustering, a tentative recipe with instructions to explore a seismic sequence and identify its main spatial features through DB algorithms is proposed. Then, an application to better characterize the region of the 2016 Kumamoto sequence where the main shocks occurred is illustrated. All the numerical analyses have been performed by using software packages available in the Statistics and Machine Learning Toolbox of MATLAB R2021a.

## 2 DB ALGORITHMS

### 2.1 DBSCAN

DBSCAN stands for *Density Based Spatial Clustering of Application with Noise* and was introduced by Ester *et al.* (1996) with the aim to discover clusters of arbitrary shapes in large spatial databases with noise. The algorithm is based on only two input parameters (see Fig. 1a): $\varepsilon$, the neighbourhood distance around a given point; and $Z$, the minimum number of points in a neighbourhood. Once the values of $\varepsilon$ and $Z$ are assigned, DBSCAN classifies data points, p, into three categories as follows:

(1)A a *core point*, if the number of points in its $\varepsilon$-neighbourhood, $N_\varepsilon(\mathrm{p})$, is greater than or equal to $Z$, that is $N_\varepsilon(\mathrm{p}) \geq Z$.
(2)As a *boundary point*, if two conditions are satisfied: (i) the number of points in its neighbourhood is less than $Z$, that is $N_\varepsilon(\mathrm{p}) < Z$, (ii) p is in the $\varepsilon$-neighbourhood of a core point.
(3)As a *noise point*, if it is neither a core point nor a boundary point, that is $N_\varepsilon\ (\mathrm{p}) < Z$.

Initially, DBSCAN searches for core points, assigns them a cluster index (hereafter called 'colour'), and gives the same colour to all core points that are in the $\varepsilon$-neighbourhood of each other. These points are called density connected core points (see Fig. 1a) and their spatial distribution determines the shape and the number of clusters. Boundary points take the colour of the nearest core point, while noise points are discarded. We notice that setting the values of $\varepsilon$ and $Z$ is equivalent to introducing a density threshold to influence which points become clustered. Thus, varying $\varepsilon$ and $Z$ corresponds to increase or decrease this threshold, that means clustering smaller or larger groups of points. Looking at the distribution of points in Fig. 1(a), for example, if $Z = 3$, all points belong to the same cluster except for one noise point; instead if $Z = 5$, the algorithm does not find any cluster because all points are noise points. One of the most striking features of this algorithm is that the cluster geometry is not predefined and clusters of any shape can be identified just grouping paths of density connected points. This is particular useful for cluster analyses of 3-D spatial distribution of earthquakes as it might be of help in discovering complex networks of fault systems.

Finally, we note that the number of clusters retrieved by DBSCAN does not depend on the order in which the data points are processed. Instead, boundary points might belong to adjacent clusters and the algorithm assigns them to the first discovered cluster.

### 2.2 OPTICS

OPTICS stands for *Ordering Points To Identify Clustering Structure* and is an extension of DBSCAN proposed by Ankerst *et al.* (1999). Actually, it is not a clustering algorithm but an ordering algorithm introduced to overcome the main drawback of DBSCAN, that is, not being able to distinguish regions with different densities. The basic idea is that for a given $Z$, denser clusters may be completely contained in clusters of lower density. Therefore, if higher density points are processed first, a clustering order can be obtained, which contains information about hierarchically nested clustering structures.

To identify the clustering structure, the algorithm computes for each point, p, two additional quantities called core distance, $d_{\mathrm{C}}$, and reachability distance, $d_{\mathrm{R}}$, as follows (see also Fig. 1b):

$$d_{\mathrm{C}}\,(\mathrm{p};\ \varepsilon,\ Z)\ = \begin{cases} \text{undefined} \ \ \text{if} \ \ N_\varepsilon\,(\mathrm{p}) < Z \\ \varepsilon' = \min\,(\varepsilon)\ \mid\ \ N_{\varepsilon'}\,(\mathrm{p}) \geq Z. \end{cases}$$

$$d_{\mathrm{R}}\,(\mathrm{p},\,\mathrm{q};\ \varepsilon,\ \ Z)\ = \begin{cases} \text{undefined if} \ \ N_\varepsilon\,(\mathrm{p}) < Z \\ \max\,(\varepsilon',\ \ \mathrm{dist}\,(\mathrm{p},\,\mathrm{q}))\ \ \text{otherwise.} \end{cases}$$

In other words, for a given $Z$, $d_{\mathrm{C}}$ is the minimum neighbourhood distance (i.e. minimum $\varepsilon$) to make the point p a core point, whereas $d_R$ between q and p is defined only if p is a core point, in which case $d_R$ equals the maximum of $d_{\mathrm{C}}$ and the Euclidean distance between p and q. It is worth noting that the algorithm does not necessarily need the parameter $\varepsilon$ because the search radius can span all possible values between zero and infinity, that is exploring all possible values for $d_{\mathrm{C}}$. Practically, to save computation time, $\varepsilon$ is set to a reasonably large value that serves as the maximum distance to consider.

The algorithm starts similar to DBSCAN with finding core points, but then explores new points in the order of lowest to highest $d_{\mathrm{C}}$. The result is a reachability plot that represents $d_{\mathrm{R}}$ of each point as a function of the cluster-ordered list of points and provides information about the clustering structure. An example reachability plot is shown in Fig. 2 for a data set with 300 data points and five clusters. Such a graph can be considered as a special type of dendrogram (Sander *et al*. 2003), since the obtained clustering structure is hierarchical and indicates the existence of nested clusters. In Fig. 2(b), the points belonging to clusters have very low $d_{\mathrm{R}}$ (<1.5), and correspond to apparent 'valleys'; the smaller $d_{\mathrm{R}}$, the denser are the corresponding clusters. The peaks represent points with larger $d_{\mathrm{R}}$ and separate individual clusters. The higher are the peaks, the more separated are the clusters. Clusters can be extracted from the reachability plot by selecting a threshold value of $\varepsilon$, that is drawing a horizontal line in Fig. 2(b). The number of valleys below such a threshold results in the exact same cluster solution as DBSCAN for the same $\varepsilon$ and $Z$.

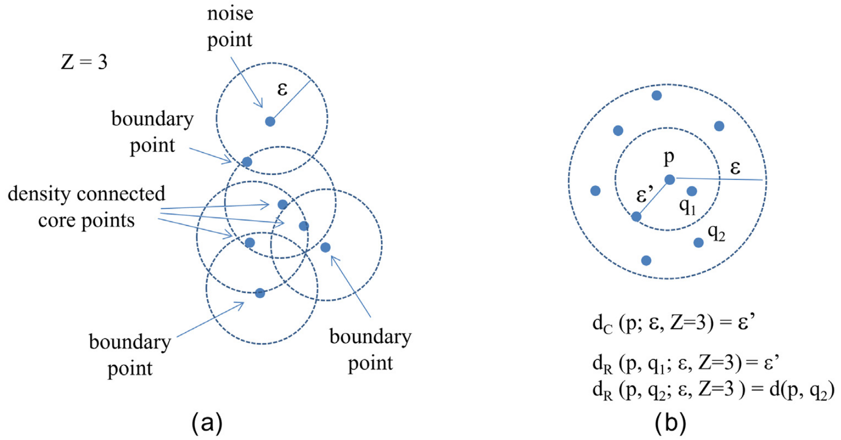

**Figure 1.** Graphical representation of (a) DBSCAN classification of data points basic concept and (b) OPTICS definitions of core and reachability distances.

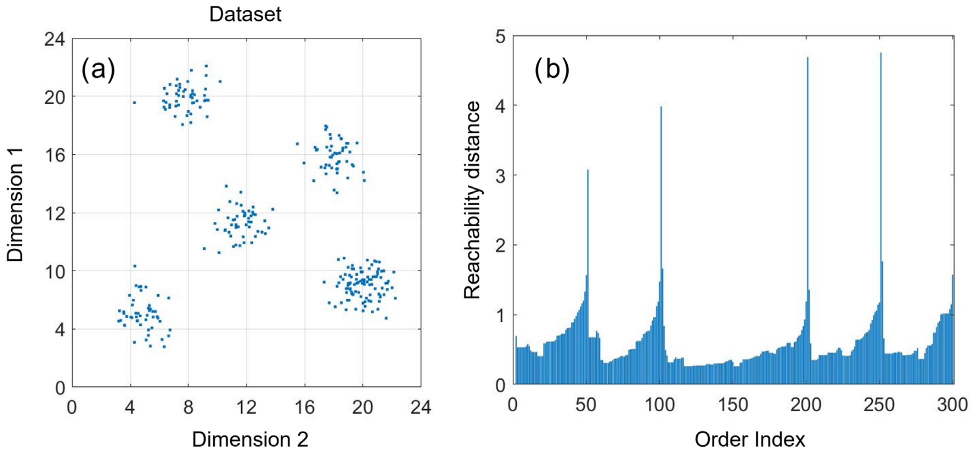

**Figure 2.** Example data set (a) and the corresponding reachability plot for $Z = 6$ (b).

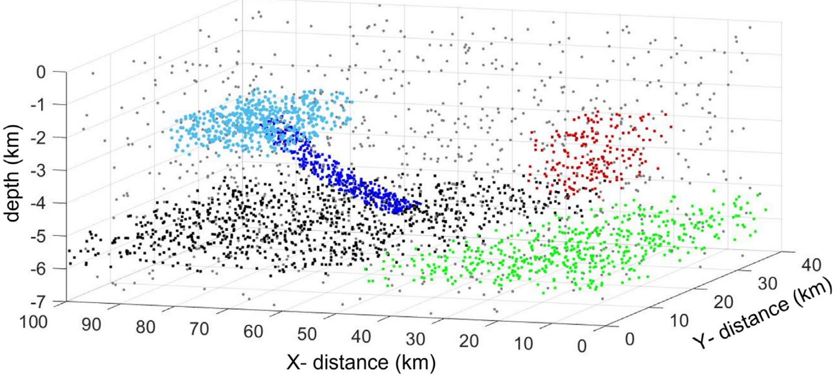

**Figure 3.** Spatial distribution of the synthetic data set consisting of five manually defined structures. The structures (coloured dots) and background activity in the whole volume (grey dots) are represented by uniformly distributed random points of varying density, totalling 3280 points.

## 3 APPLICATION TO A SYNTHETIC DATA SET

To illustrate how DB algorithms operate, we apply them to a synthetic data set of hypocentres. This analysis has multiple purposes summarized as follows:

(i) Illustrating how DB clustering works in principle by using an example with simple structures of known geometry.

(ii) Visualizing cluster solutions as function of the parameters.

(iii) Demonstrating that rescaling the data can help to recognize the largest structural features in presence of highly anisotropic structures.

### 3.1 Data set presentation

Fig. 3 shows the synthetic data set consisting of five manually defined large structures represented by uniformly distributed random

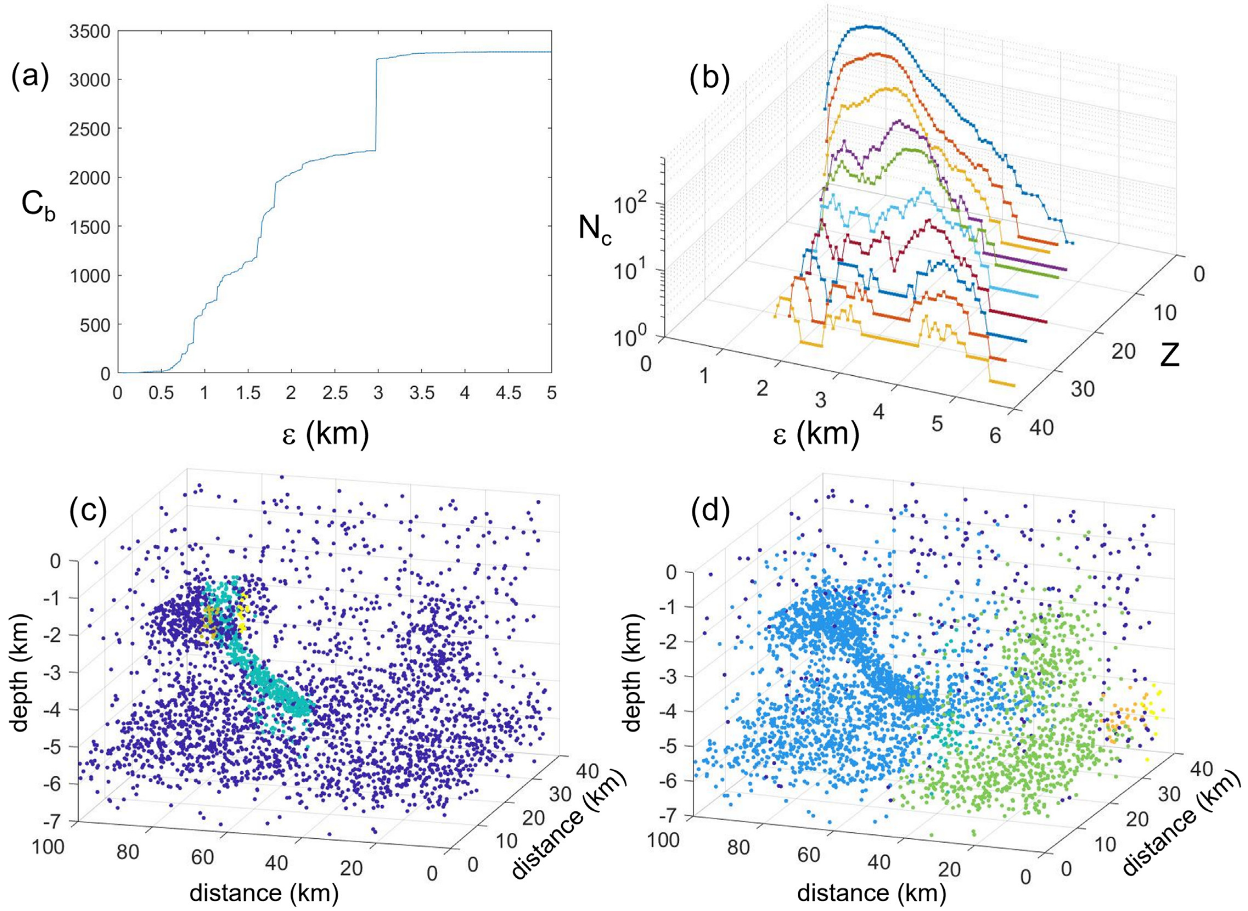

**Figure 4.** Exploring the influence of input parameters on DBSCAN solutions: (a) Number of points belonging to the biggest cluster, $C_b$, as a function of $\varepsilon$ for $Z = 1$; (b) Number of clusters, $N_c$, as function of input parameters $\varepsilon$ and $Z$; (c) DBSCAN solution for $\varepsilon = 1.4$ km and $Z = 15$; (d) DBSCAN solution for $\varepsilon = 3.4$ km and $Z = 15$. Dark blue points in (c) and (d) represent noise points and do not belong to any cluster.

points of varying density (2480 points in total). In addition, a uniform noise consisting of 800 uniformly distributed random points (about 25 per cent of the total points) was added to the whole volume (40 km × 100 km × 7 km) to represent background activity.

In particular, the synthetic geometric structures are polygonal regions representing (i) two planar structures at ∼6 km depth (black and green in Fig. 3, slightly shifted in depth), which extend up to 40 km and 60 km horizontally, respectively and about 1 km vertically; (ii) a shallow planar structure occupying a volume of about $10 \times 30 \times 1$ km$^3$ (cyan in Fig. 3); (iii) an inclined surface extending for about 2 km in depth and connecting two other structures (dark blue in Fig. 3) and (iv) a square prism-shaped volume of about 2 km height (red in Fig. 3). Choosing these structures has been motivated by the following reasons: (i) shallow and deep planar structures with different orientations mimic horizontal planes associated with thrust shear zones; (ii) intersections between structures mimic intersecting faults; (iii) strong anisotropy mimics larger sequences that propagate along a fault system and (iv) various orientations and overall 3-D interconnectedness mimics a fractured volume without preferential fault planes.

### 3.2 Cluster solutions in the parameter space

DBSCAN provides a wide range of solutions with clusters differing in number, shape and size depending on the value of $\varepsilon$ and $Z$. For $Z = 1$, all points become clustered (i.e. belong to one or more clusters). Hints about the number of the largest structures can be derived from the number of stepwise increases of the biggest cluster size, $C_b$, as a function of $\varepsilon$. The behaviour of $C_b(\varepsilon, Z = 1)$ for the synthetic data set is shown in Fig. 4(a). $C_b(\varepsilon, Z = 1)$ grows step-wise every time a clustered region joins the biggest cluster. By increasing $\varepsilon$, the density threshold, $Z/\varepsilon$, for identifying core and boundary points decreases, leading to the clustering of larger regions with lower density. Small jumps in $C_b$ indicate that small and dense regions are incorporated into the biggest cluster. Bigger jumps in $C_b$ instead indicate the presence of large and dense regions that are spatially distant, as the clusters they belong to must increase in size before joining the biggest cluster. This is more easily understood for a data set with two dense regions that are separated by a large gap. By increasing $\varepsilon$, two big clusters in each of the two regions will form and continue to increase in size (simultaneously and independently of each other) until they merge. At this point, the larger the spatial distance between the two dense regions, the larger the corresponding stepwise increase in $C_b$ will be (because the $\varepsilon$ range in which both clusters grow separately increases with the separation gap). Therefore, jumps in $C_b$ are controlled by the size and spatial distance of dense regions.

For small $Z$ ($< 5$), the number of clusters, $N_c$, typically becomes very large and then gradually decreases to 1 for increasing $\varepsilon$ (see Fig. 4b). For larger $Z$, $N_c(\varepsilon)$ becomes more complex including minor fluctuations before reaching 1 for large $\varepsilon$.

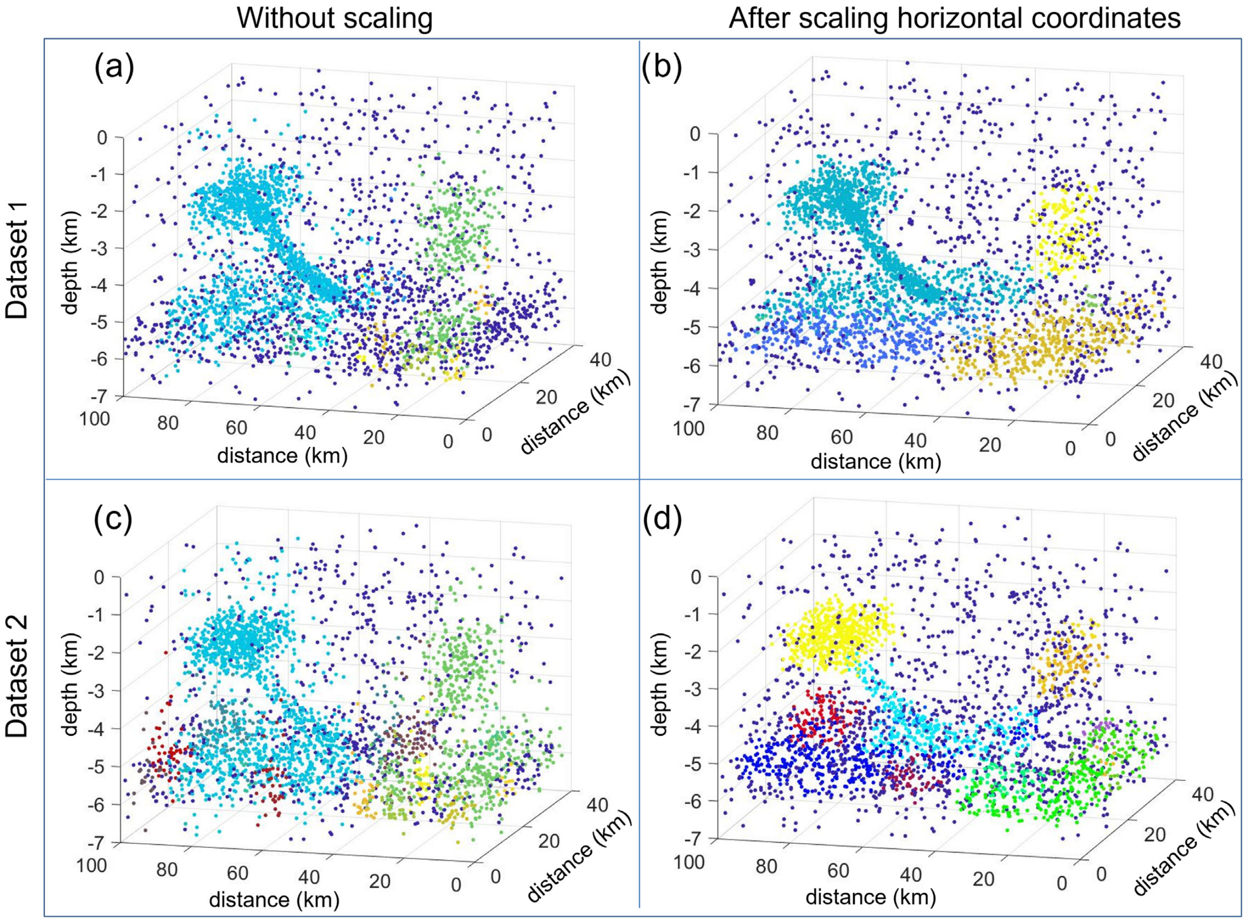

**Figure 5.** Influence of data scaling on DBSCAN solutions of two synthetic data sets for $Z = 15$. The two data sets differ only in the number of random points representing the inclined surface intersecting horizontal structures. (a) $\varepsilon = 2.6$ km; (b) $\varepsilon = 0.45$ km; (c) $\varepsilon = 2.8$ km and (d) $\varepsilon = 0.43$ km.

If $\varepsilon$ is small, $Z/\varepsilon$ becomes large, causing only regions with locally high densities to become clustered; most points are classified as noise. If $\varepsilon$ is large, $Z/\varepsilon$ becomes small, causing an inclusion of less dense regions into the clustering and most points to end up in the biggest cluster. Examples for these two extreme cluster solutions are shown in Figs 4(c) and (d), respectively: Fig. 4(c) illustrates that the region with the highest density becomes clustered, whereas Fig. 4(d) illustrates a separation of the large horizontal structures at depth, which resembles a characteristic feature of the synthetic data set.

Examples of cluster solutions for intermediate values of the threshold density $Z/\varepsilon$ are reported in Fig. 5. In particular, Fig. 5(a) shows that DBSCAN produces two large clusters that do not separate shallow and deep structures. This limitation is related to the isotropic neighbour searching, that is processing points by using spheres of radius $\varepsilon$, for which even a small increase in $\varepsilon$ leads to incorporate structures into the clusters that are outside the planar structures or linked to them. This can be more easily understood by focusing on the structures that form the big cyan cluster of Fig. 5(a). In presence of intersecting structures, like a planar structure and an inclined surface, DBSCAN is not able to distinguish them as individual structures even though the point density in the planar structure is high enough and the value of the neighbourhood search radius $\varepsilon$ does not exceed its thickness. This indiscernibility happens for two main reasons: (i) decreasing $\varepsilon$ while $Z$ is kept fixed leads to a considerable increase of noise points (see Fig. 4c) and (ii) DBSCAN gives the same colour to paths of density connected points of any shape, therefore making intersecting structures inseparable unless they are characterized by different densities.

In an attempt to overcome this limitation, we scaled the data by homogenizing horizontal and depth ranges before clustering, using the latter as a reference (here: 0–7 km). To translate each coordinate individually to a common range, we used the min–max scaling for each horizontal coordinate $x$:

$$x_{new} = \frac{(max_{new} - min_{new})}{(max_{old} - min_{old})}\ (x_{old} - min_{old}) + min_{new}, \qquad (1)$$

with $\mathrm{min}_{\mathrm{new}} = 0$ km and $\mathrm{max}_{\mathrm{new}} = 7$ km. For intermediate values of the threshold density $Z/\varepsilon$, Fig. 5(b) shows a cluster solution after applying this scaling, performing the clustering with DBSCAN and mapping the results back to the original space. In this case, clustering is more effective in resolving the shallow planar structure (yellow points) and one of the two horizontal structures (brown points). However, the shallow planar structure together with the inclined structure and a large part of a deep horizontal structure still belong to the same cluster (cyan cluster). The scaling-based cluster analysis fails in this part because the point density within the inclined structure is very high. To demonstrate the influence of this high density, we repeat the analysis for a subset of the synthetic data set in which the inclined structure has only 25 per cent of the points, that is a four times lower density (see Figs 5c and d).

Accordingly, clustering without data scaling is again not able to discriminate shallow and deep structures, whereas they can be

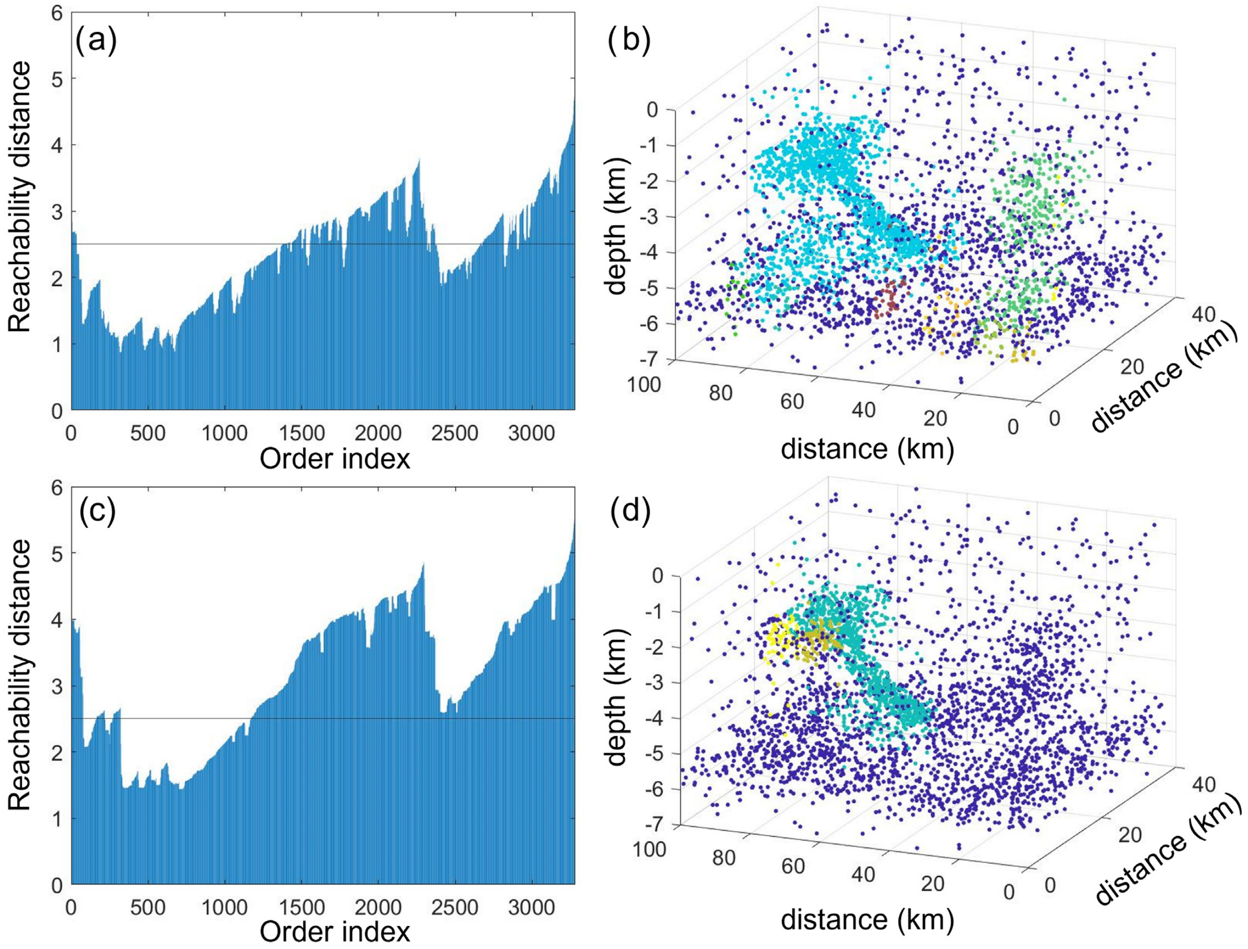

**Figure 6.** Reachability plots of the OPTICS algorithm for the synthetic data set in (a) and (c), and corresponding DBSCAN solutions for $\varepsilon = 2.5$ km in (b) and (d), respectively. Figures in the top row relate to $Z = 15$ and those in the bottom row to $Z = 30$. Dark blue points are noise points.

identified when data are scaled beforehand, even if the inclined structure is still not well defined. It is worth noting that the min–max scaling has some caveats. For instance, it may amplify the effect of local depth uncertainties and does not preserve the absolute distances among event pairs when the spatial boundary of the catalogue changes (but here we do not consider temporal changes of the catalogue). We suggest the min–max scaling only in the presence of strong anisotropies, that is when the horizontal extension of large dense regions is much larger than the vertical extension, $Lx,y/Lz >> 1$. Figs 5(b) and (d) show that the scaling-based cluster analysis fails in identifying intersecting structures as distinct objects if their contrast in point density values is not high enough. However, we think that such a scaling is useful to identifying planar structures that could be at least partially hidden by the uniform point distributions in depth caused for instance by uncertainties.

For two different choices of parameter $Z$ (15 and 30), Fig. 6 shows reachability plots (left-hand column) and the corresponding DBSCAN solutions for $\varepsilon = 2.5$ km (right-hand column). The comparison reveals that a small $Z$ produces more small-scale valleys in the reachability plot than a larger $Z$, which reduced their widths. Accordingly, a smaller $Z$ results in a larger number of clusters because the $\varepsilon$ threshold crosses more valleys horizontally than for larger $Z$. Although it is theoretically possible to get information about the number of characteristic structures by simply counting the number of the crossed valleys, practically this is not an easy task if $Z$ is too small, because the meaning of a valley may be ambiguous. Both reachability plots reveal two main valleys, which can be considered as the main features of the data set. Such valleys do not correspond to any of the five manually defined structures but contain them. In particular, the relative locations of the main valleys suggest a spatial separation that divides the investigated volume into two parts, which are illuminated in Fig. 5(d) by cyan and green dots, respectively. The fact that the $\varepsilon$ threshold cannot cross all the nested clusters (‘subvalleys’) inside the biggest clusters (main valleys) indicates that isotropic neighbour searching is not effective for our synthetic data set and that scaling improves its characterization in DB clustering.

## 4. REPRESENTING DB CLUSTER SOLUTIONS IN A PHASE DIAGRAM

DB algorithms provide cluster solutions that can vary greatly in size and shape depending on the values of $\varepsilon$ and $Z$. So, the question is how to choose these parameter values. A common strategy for estimating an appropriate $\varepsilon$ is to detect the ‘knee’ in a $k$-distance graph, which plots the distances of each point to its $k$th nearest point in sorted order (see Ester *et al.* 1996). However, this approach does not always return an optimal $\varepsilon$, especially when a certain number of large clusters is desired instead of a single big one. As noted by Cesca (2020), a general rule to determine the best value of $\varepsilon$ and

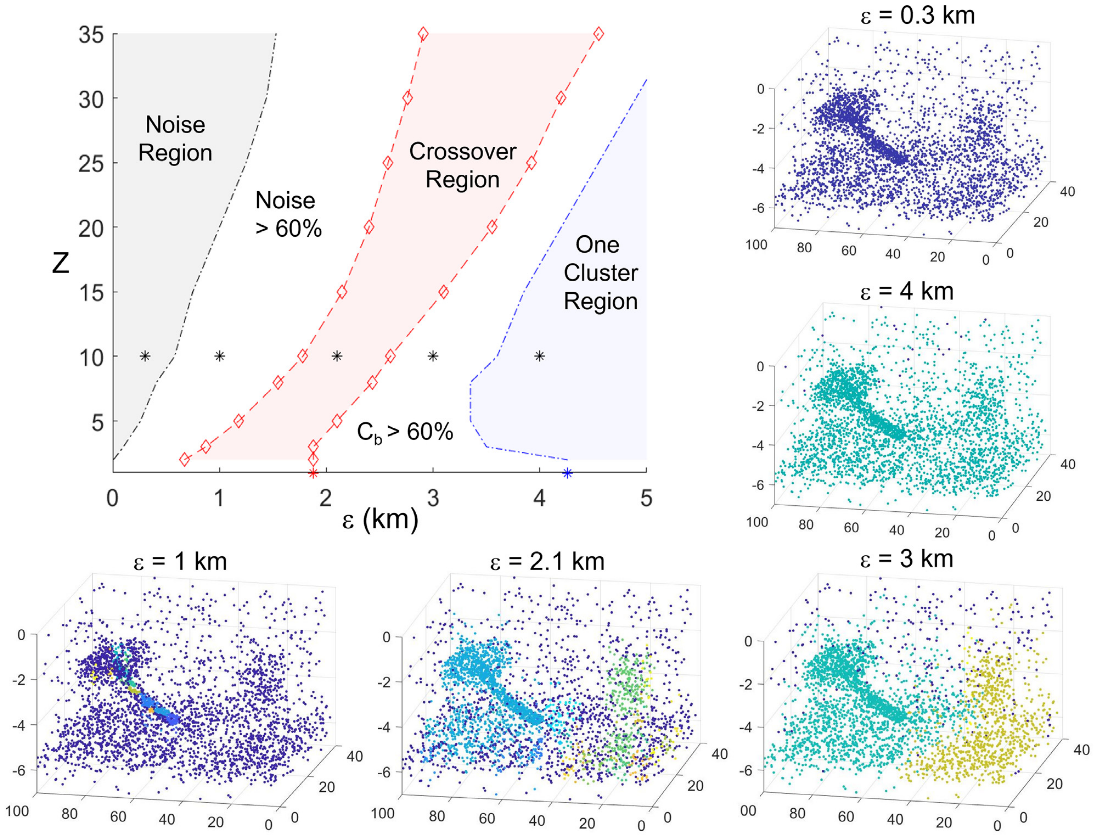

**Figure 7.** Phase diagram of DBSCAN solutions for the synthetic data set. The dotted–dashed and the red dashed lines divide the parameter space into five regions with different types of cluster solutions (see annotations and main text). Cluster solutions corresponding to the five points in the diagram for $Z = 10$ (marked by asterisks) are reported in separate subplots. In each subplot, dark blue points are noise points.

$Z$ cannot be provided because DB algorithms are used for different purposes.

To get a better understanding of what types of information can be retrieved from DB clustering algorithms by varying the input parameters, we explore the whole space of solutions for synthetic and real seismic catalogues. The numerical analysis has shown that the phase diagram in the parameter space can be divided into five areas, which represent different classes of cluster solutions. As an example, Fig. 7 shows the phase diagram of the analysed synthetic data set. At the opposite ends of $\varepsilon$ axis in the phase diagram, we find two extreme conditions: for very low $\varepsilon$ all data points become noise points, whereas for very high $\varepsilon$ all data points become connected to a single cluster. Moving horizontally from right to left in the phase diagram (i.e. decreasing $\varepsilon$ and increasing the density threshold $Z/\varepsilon$), the size of the biggest cluster, $C_b$, decreases and other clusters appear. The locations of the jumps in $C_b$ occur every time it splits into two or more clusters. Based on this behaviour, we obtain a first critical $\varepsilon$ value when $C_b$ contains 60 per cent of all points, that is for larger $\varepsilon$, cluster solutions are characterized by a big cluster that contains more than 60 per cent of the data. Similarly, by moving from left to right in the phase diagram (i.e. increasing $\varepsilon$ and decreasing the density threshold $Z/\varepsilon$), we obtain another critical $\varepsilon$ value when 60 per cent of the data are noise points, that is for lower $\varepsilon$, cluster solutions are characterized by more than 60 per cent of noise points. For the synthetic data set presented in the previous section, examples of cluster solutions for which Noise > 60 per cent and $C_b > 60$ per cent are shown in Figs 4(c) and (d), respectively, for $Z = 15$. The two critical $\varepsilon$ are determined for various $Z$ to construct the phase diagram (red markers in Fig. 7). The area between them is a transition zone named ‘crossover region’, which represents cluster solutions with many large clusters. Cluster solutions in Fig. 5 all belong to the crossover region. We are most interested in cluster solutions belonging to this region since they maximize the number of large clusters and help us to identify volumes with the highest density (*natural clustering*). Note that by using other jumps in $C_b$ (i.e. other splits of the biggest cluster), it is possible to reconstruct a cluster hierarchy and divide the crossover region into subregions that differ in the number of large, stable clusters—depending on the event density distribution.

For the special case $Z = 1$, there are no noise points, but it is still possible to define two critical $\varepsilon$ values, above which the biggest cluster contains more than 60 per cent of the data (red star) and all data (blue star), respectively.

Generally, decreasing $\varepsilon$ for a fixed $Z$ leads to more clusters, while increasing $Z$ for a fixed $\varepsilon$ leads to a fewer clusters. However, the number of clusters as a function of $\varepsilon$ and $Z$ does not behave monotonic (see Fig. 4b).

It is worth noting that increasing the height of the horizontal $\varepsilon$ threshold in the OPTICS’ reachability plot is equivalent to moving from left to right in the phase diagram. Obtaining cluster solutions in combination with the reachability plot has the advantage of accounting for the nested clustering structure—because the reachability plot visualizes, for a fixed $Z$, all cluster solutions of DBSCAN for a broad range of $\varepsilon$ values.

The phase diagram also shows that cluster solutions depend slightly on $Z$; an increase of $Z$ generally leads to an increase of noise points and to clusters that are more convex

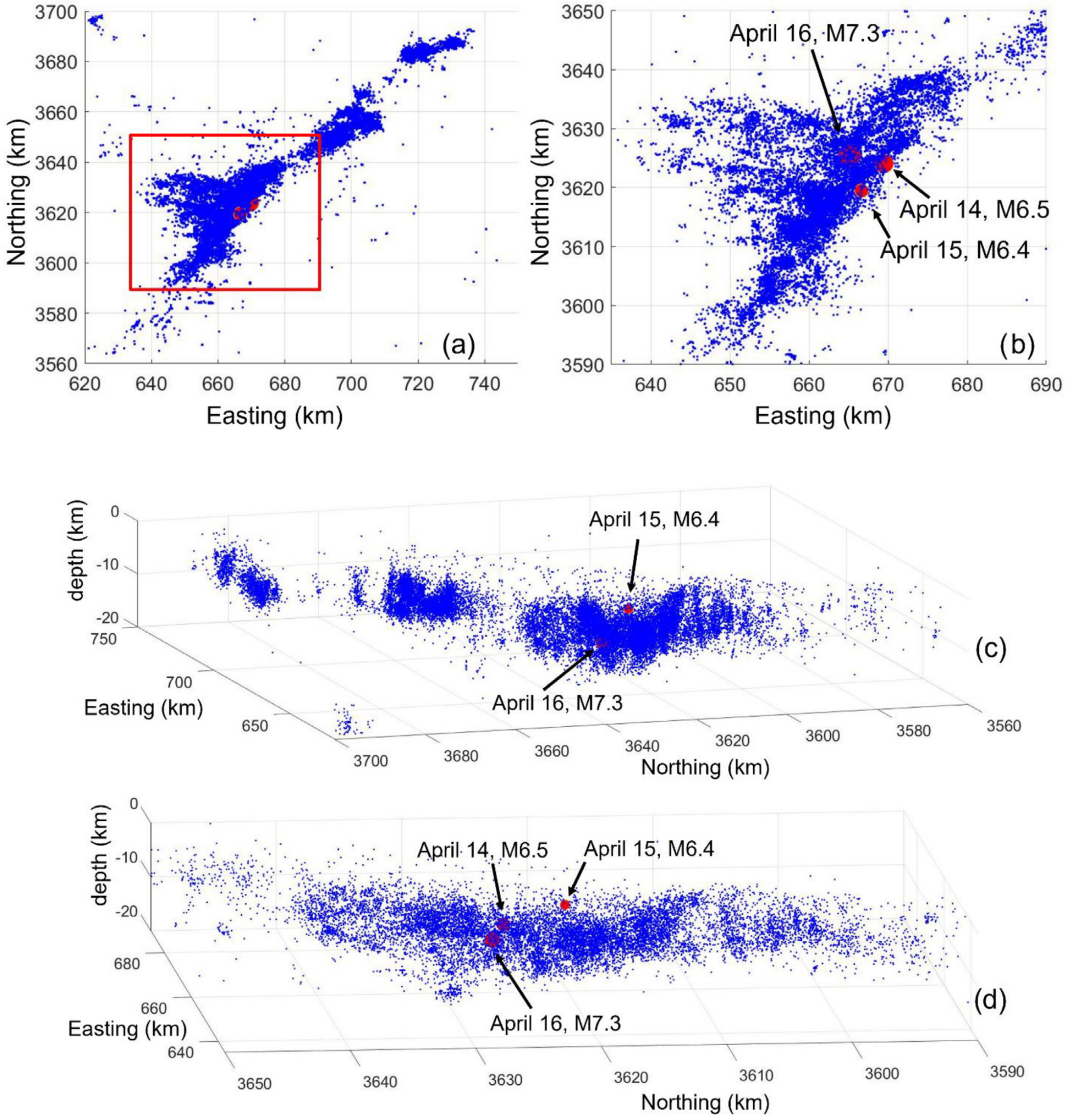

**Figure 8.** Overview of the 2016 Kumamoto sequence using catalogue extracts between 1 April 2016 and 31 August 2016. (a) Map view; (b) zoom into the area where the three largest earthquakes occurred [see red frame in '(a)']; (c) 3-D representation of (a); (d) 3-D representation of (b). The three largest earthquakes are represented with red markers and annotated with their magnitude and day of occurrence.

in shape. Recall that a larger $Z$ produces less small-scale valleys in the reachability plot. Thus, even if the reachability plot helps in identifying the number of large clusters in the data set, the choice of $Z$ still affects the characterization of the cluster hierarchy.

Finally, we want to point out that the areas covered by the five regions in the phase diagram of Fig. 7 depend on the spatial distribution of the data, which is an intrinsic property of a data set. Thus, changing the data distribution will change the width of the crossover region. Furthermore, exploring the whole parameter space is computationally expensive and practically unnecessary. Our analysis shows that only solutions in the crossover region are representative to extract meaningful information about the characteristic largest structures of a data set. In Section 6.1, we suggest a procedure for finding them, and selecting those with the desired level of nesting structure by using OPTICS.

## 5 APPLICATION TO REAL EARTHQUAKE CATALOGUES

### 5.1 The 2016 Kumamoto earthquake sequence

We performed DB cluster analysis of events that occurred between 1 April 2016 and 31 August 2016 (4 months) in the Kumamoto area, southwest of Japan. The earthquake catalogue was obtained from the Seismological Bulletin of Japan as provided by the Japan Meteorological Agency (JMA) and contains 163 988 events. We only use events with $M > 1$ and hypocentral depths shallower than 20 km within the spatial range of UTM coordinates from 621 to 745 km Easting and from 3564 to 3699 km Northing (WGS coordinates: 130.3–131.6°E, 32.2–33.4°N), totaling 20 887 events. Fig. 8 shows 2-D and 3-D representations of the hypocentral locations with the three largest earthquakes highlighted with a red marker ($M6.5$, $M6.4$ and $M7.3$).

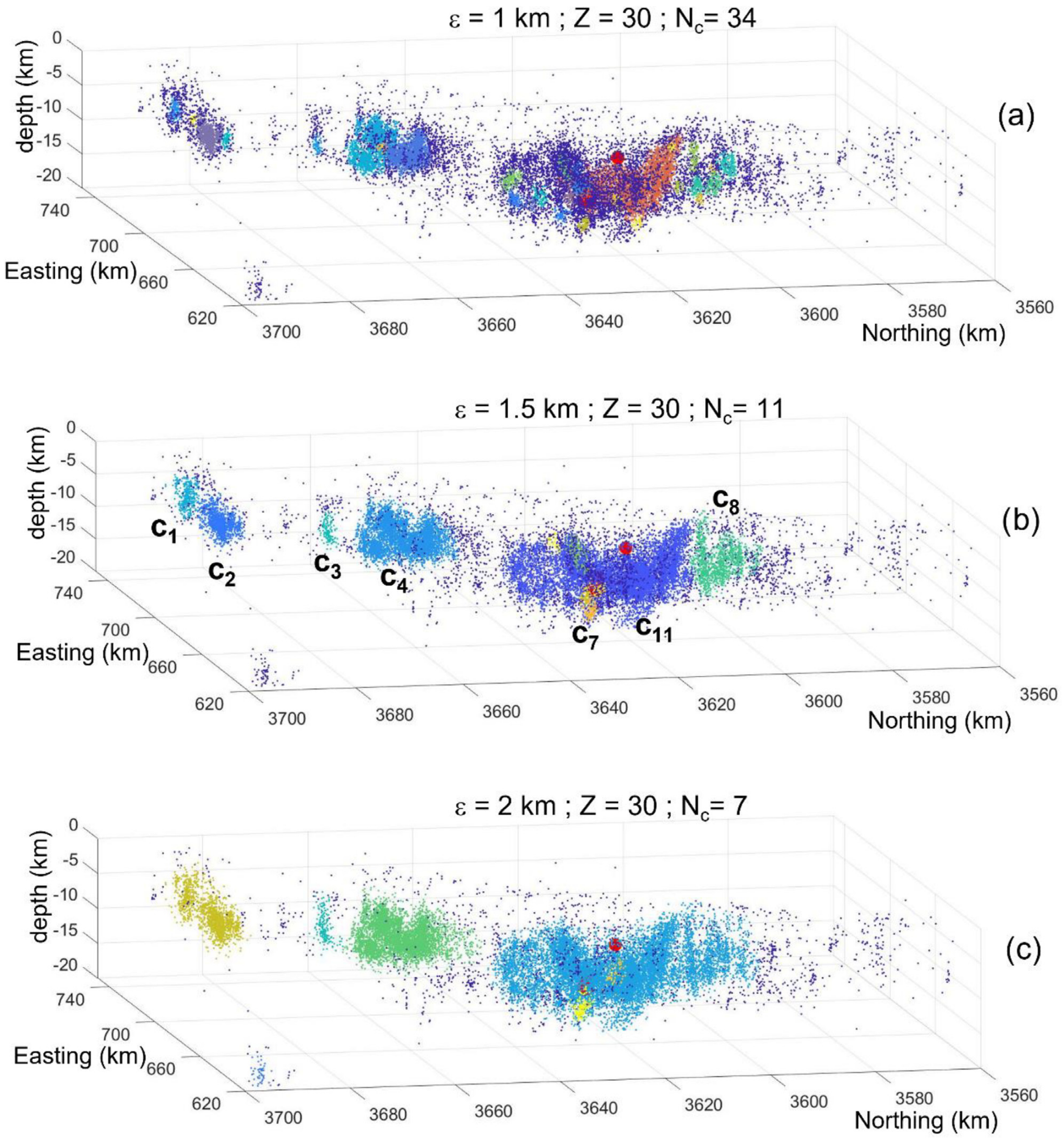

**Figure 9.** DBSCAN solutions of the 2016 Kumamoto sequence for $Z = 30$ and varying $\varepsilon$: (a) $\varepsilon = 1$ km, (b) $\varepsilon = 1.5$ km and (c) $\varepsilon = 2$ km. Red markers represent the location of the three largest earthquakes. Dark blue points represent noise points.

From Figs 8(a) and (c), we visually recognize a few big clusters as denser areas that are separated by areas with sparse seismicity. However, a zoom into the region where the largest earthquakes occurred (Figs 8b and d) blurs the sharp borders of the denser areas and reveals finer details, making a visual recognition of clusters ambiguous. With DB clustering, instead, we can divide the catalogue into natural groups in an exploratory way and identify patterns within it.

Fig. 9 shows three DBSCAN solutions for different choices of the input parameters inside the crossover region, that is for which both the number of noise points and $C_b$ are less than 60 per cent of the data. These choices for $\varepsilon = 1$, 1.5 and 2 km divide the seismic sequence into 34, 11 and 7 clusters, respectively. With an increasing $\varepsilon$, the number of clusters and the number of noise points decrease, whereas the largest clusters increase in size by incorporating more adjacent hypocentres. Even though the shapes of the clusters change by varying $\varepsilon$, the centres of the largest clusters remain the same; the clusters always represent the most active zones and the largest earthquakes always belong to the same cluster (coluored orange, purple and light blue in Figs 9a–c, respectively).

Fig. 10 shows three reachability plots for $Z = \{15, 30, 70\}$ and the DBSCAN solution for $Z = 30$ and $\varepsilon = 3$ km, which is characterized by the presence of three largest clusters. These three clusters are evident in each of the three reachability plots as the deepest and best-defined valleys, named C1, C2 and C3, which can be considered as the main features of the sequence. As indicated in Figs 10(b) and (c), this cluster solution can be obtained for a wide range of $\varepsilon$, that is many horizontal lines lead to a division into three big clusters. However, for a small $Z$ (Fig. 10a), the number of clusters increases significantly as revealed by the many narrow valleys inside the largest valleys. Consequently, a small variation of $\varepsilon$ can lead to new clusters that include very little data due to the narrowness of the valleys.

The height of the peaks in the reachability plots represents another feature of the seismic sequence, namely the spatial separation of the clusters. In particular, the highest peaks correspond to the

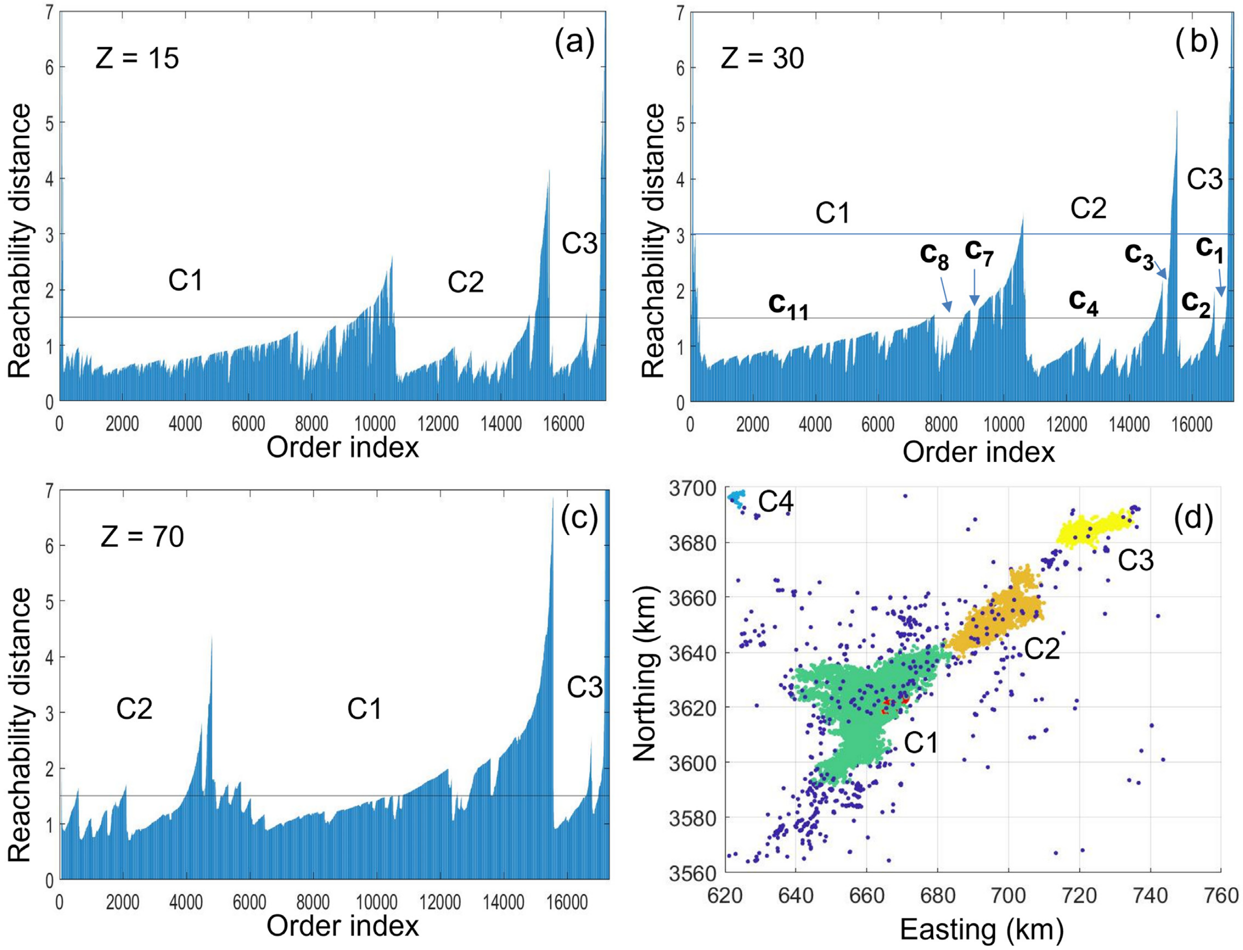

**Figure 10.** Cluster analysis of the 2016 Kumamoto sequence. (a–c) Reachability plots for different $Z$: (a) $Z = 15$, (b) $Z = 30$ and (c) $Z = 70$. Panel (d) shows a map view of the DBSCAN solution for $Z = 30$ and $\varepsilon = 3$ km. Red markers represent the location of the three largest earthquakes and dark blue points the noise points.

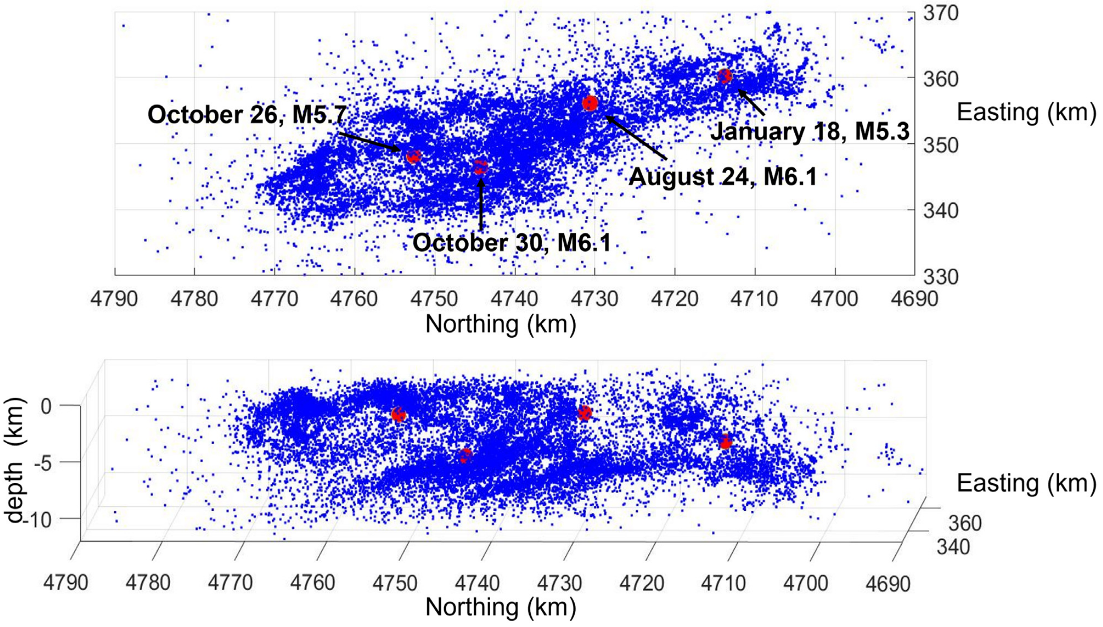

**Figure 11.** Overview of the 2016 Central Italy sequence between 15 August 2016 and 15 August 2017. Map view (top panel) and 3-D view (bottom panel). The four largest earthquakes are highlighted with a red marker and annotated with their magnitude and day of occurrence in the top panel.

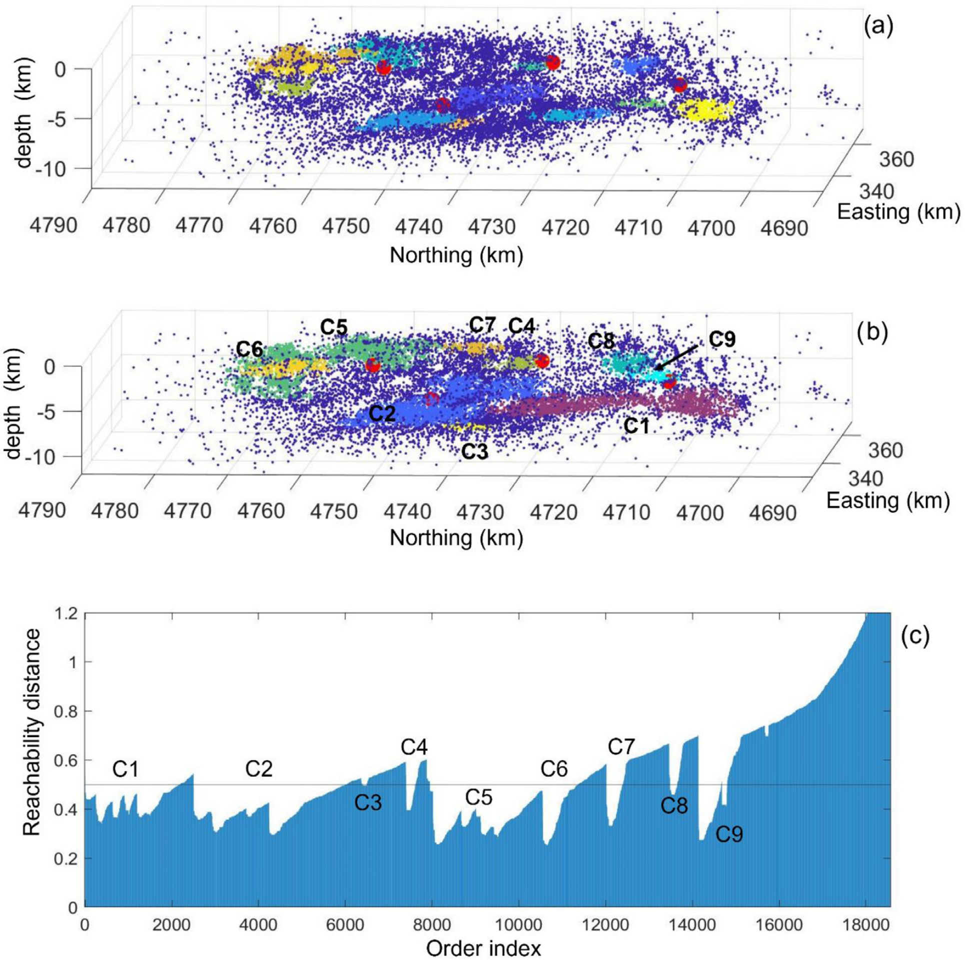

**Figure 12.** Cluster analysis of the 2016 Central Italy sequence by applying the DBSCAN algorithm for a fixed $Z = 100$. (a) and (b) Cluster solutions for $\varepsilon = 0.4$ km (a) and $\varepsilon = 0.5$ km (b). Red markers represent the location of the four largest earthquakes. Dark blue points represent noise points. (c) Reachability plot with the black horizontal line corresponding to the $\varepsilon$ threshold shown in (b). The annotated cluster names in (c) correspond to the ones in (b).

points with the largest $d_R$, which indicate the most separated clusters (see Section 2.2). The difference in height of the three main peaks in Figs 10(a)–(c) indicates that C1 and C2 are less spatially separated with respect to C3. In addition, Fig. 10(b) ($\varepsilon = 1.5$ km and $Z = 30$) lets us identify two well-defined nested structures related to the smaller valleys inside both C2 and C3 (named $c_1$, $c_2$, $c_3$ and $c_4$), which are visible as light blue and green coloured clusters in Fig. 9(b). Note that C4, which corresponds to a small event group in the northwestern sector of the area (see Fig. 10d), is not easily recognizable in the reachability plots due to its very deep and narrow valley on the left-hand side.

Finally, the horizontal $\varepsilon = 1.5$ km thresholds in the reachability plots produce a different number of clusters for different $Z$: 14 for $Z = 15$ (Fig. 10a), 11 for $Z = 30$ (Fig. 10b) and 12 for $Z = 70$ (Fig. 10c).

### 5.2 The 2016 Central Italy seismic sequence

For the 2016 Central Italy sequence, we used the high-resolution earthquake catalogue of Tan *et al*. 2021 spanning from 2016-08-15 to 2017-08-15. We only considered events with $M_w > 2$ and hypocentral depths shallower than 12 km within the spatial range of UTM coordinates from 330 to 370 km Easting and from 4690 to 4790 km Northing (12.9°–13.4°E, 42.3°–43.2°N), totalling 18 595 events (see Fig. 11). The locations of the four largest earthquakes are indicated with a red marker ($M_w$6.1 Amatrice event on 24 August 2016, $M_w$5.7 Visso event on 26 October 2016, $M_w$6.1 Norcia event on 30 October 2016 and $M_w$5.3 Campotosto event on 18 January 2017).

Since the data set is characterized by horizontally extended structures within a limited vertical range, the cluster analysis was applied after scaling the horizontal coordinates to the depth range (0–12 km) as discussed in Section 3. After remapping into the original coordinate system, Fig. 12 visualizes the obtained clusters for two different $\varepsilon$, but a fixed $Z = 100$. The first parameter set ($\varepsilon = 0.4$ km, $Z = 100$, see Fig. 12a) represents a cluster solution whose proportion of noise points is larger than 60 per cent, that is left of the crossover region. As expected, many small clusters are returned (13 clusters, maximally about 1000 events each). Instead, the second parameter set ($\varepsilon = 0.5$ km, $Z = 100$, see Fig. 12b) is located inside the crossover region and produces a balance between the amount of noise points and density-connected points, maximizing the number of large clusters.

Fig. 12(c) shows the reachability plot for the same $Z = 100$ and reveals several well-defined valleys corresponding to many high-density zones. The threshold $\varepsilon = 0.5$ km (black horizontal line in Fig. 12c) crosses nine valleys, which correspond to the DBSCAN

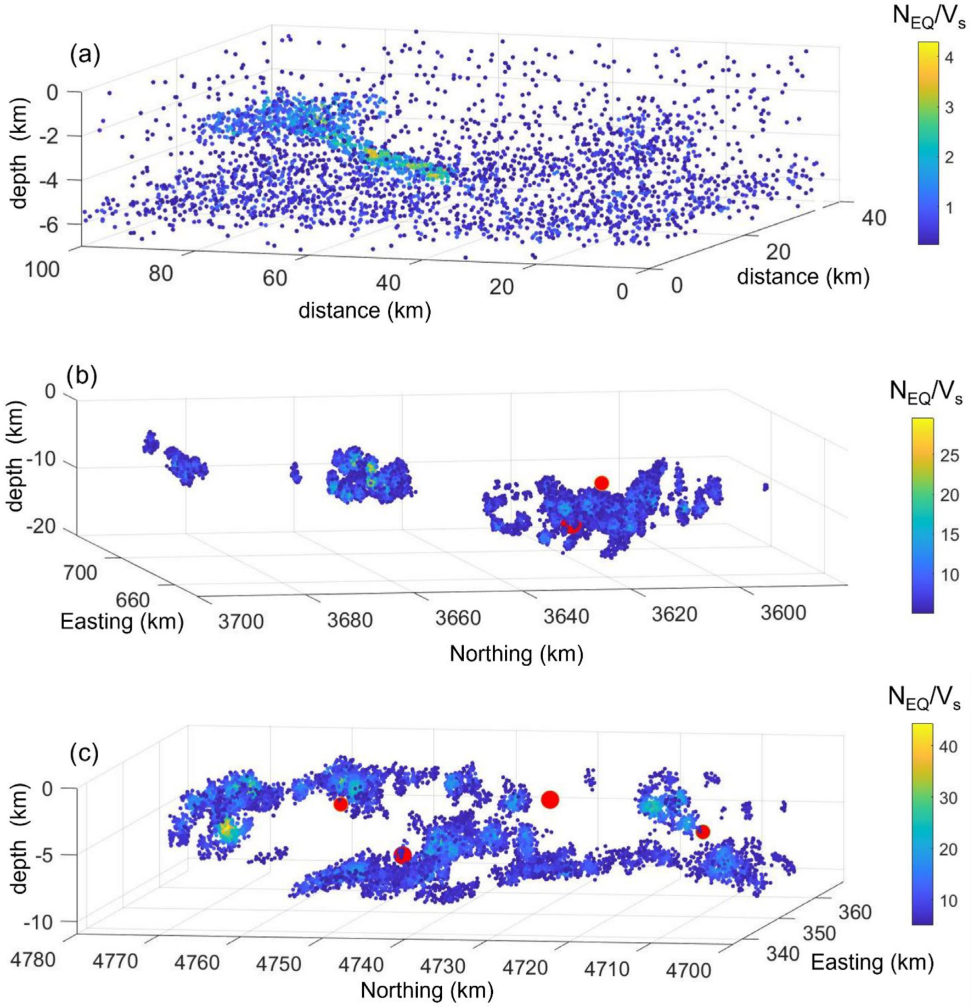

**Figure 13.** Hypocentre density of the (a) synthetic data set, (b) 2016 Kumamoto sequence and (c) 2016 Central Italy sequence. The hypocentre density (see colour bar) is represented at each event, but for the real cases only if it is larger than 5. Red markers represent the location of the largest earthquakes.

solution shown in Fig. 12(b). The reachability plot provides information not only on the presence of nested structures but also on the size and the number of the largest clusters. The main features of the catalogue are the three largest earthquake clusters, named C1, C2 and C5 in Figs 12(b) and (c). C1, which represents the extended structure at depth in the south, contains four smaller valleys. These four valleys were identified as individual clusters for a smaller $\varepsilon = 0.4$ km and are visible in Fig. 12(a) as clusters of different colour in this region. C2, which represents the extended structure at depth in the centre of the sequence and includes the Norcia main shock, is characterized by two larger and one smaller valleys—three substructures also visible in Fig. 12(a). C5, which represents a shallower structure in the north and contains the Visso event, contains five valleys of which three correspond to structures identified with $\varepsilon = 0.4$ km in this region (Fig. 12a). The spatial volumes illuminated by C1, C2 and C5 are also the main features of this catalogue with a lower magnitude cut-off ($M_w > 1.5$, totalling 76 055 events), which have been statistically analysed to characterize the behaviour of the magnitude distribution during and within this complex sequence (Herrmann *et al*. 2021). The remaining clusters in Fig. 12(b) either did not change significantly between the two parameter sets (e.g. C4, C6 and C8), or were added for the higher $\varepsilon$ (e.g. C7).

## 6 A GENERALIZED APPROACH TO DB CLUSTER ANALYSIS AND A FURTHER APPLICATION

DB clustering algorithms undoubtedly facilitate the analysis of large catalogues by only using two input parameters. Yet, these two parameters can lead to a variety of cluster solutions, making their choice difficult. Ultimately, the preferred clustering solution depends on the purpose (i.e. the desired grouping of the data set), because a single best clustering solution does not exist. The cluster hierarchy of the catalogue can serve as key information for choosing the preferred solution and can be retrieved from the reachability plot of the OPTICS algorithm. Our analyses showed that parameter $Z$ is crucial when the interest is in finding not only regions with the highest hypocentre density but also large clusters that represent the main structures. We have shown that parameter sets lying in the crossover region of the phase diagram are good candidates for exploring the catalogue in a meaningful way. However, finding all cluster solutions in the crossover region by exploring the entire parameter space is impractical (and needless) especially for large catalogues. Based on our findings and some general considerations, we describe below a recipe for finding a representative cluster solution in the crossover region.

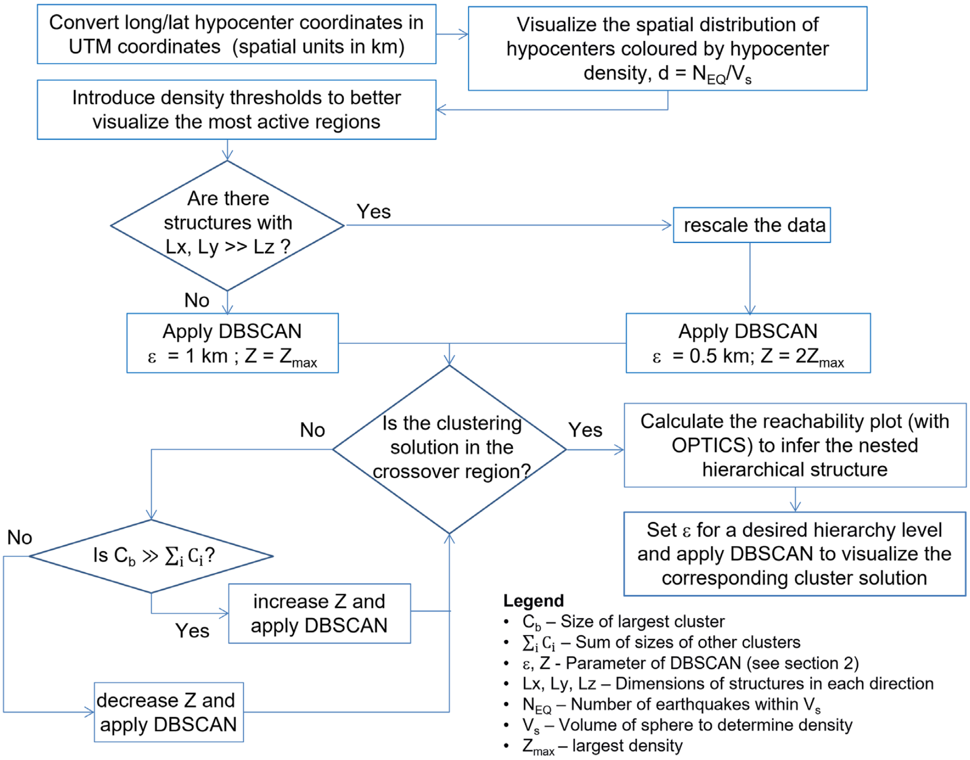
Convert long/lat hypocenter coordinates in UTM coordinates (spatial units in km)
Visualize the spatial distribution of hypocenters coloured by hypocenter density, d = $N_{EQ}/V_s$
Introduce density thresholds to better visualize the most active regions
Are there structures with Lx, Ly >> Lz ?
Yes
rescale the data
No
Apply DBSCAN ε = 1 km ; Z = $Z_{max}$
Apply DBSCAN ε = 0.5 km; Z = $2Z_{max}$
No
Is the clustering solution in the crossover region?
Yes
Calculate the reachability plot (with OPTICS) to infer the nested hierarchical structure
Set ε for a desired hierarchy level and apply DBSCAN to visualize the corresponding cluster solution
No
Is $C_b \gg \sum_i C_i$?
Yes
increase Z and apply DBSCAN
decrease Z and apply DBSCAN
Legend
• $C_b$ – Size of largest cluster
• $\sum_i C_i$ – Sum of sizes of other clusters
• ε, Z - Parameter of DBSCAN (see section 2)
• Lx, Ly, Lz – Dimensions of structures in each direction
• $N_{EQ}$ – Number of earthquakes within $V_s$
• $V_s$ – Volume of sphere to determine density
• $Z_{max}$ – largest density

**Figure 14.** Flow diagram of our proposed DB cluster analysis of a seismic sequence.

### 6.1 A tentative recipe for finding cluster solutions in the crossover region

Because DB algorithms identify clusters as dense regions separated by sparse regions, their main drawback relates to the identification of cluster boundaries, that is if density drops are absent, cluster boundaries are not well defined. Therefore, we suggest their use with a foregoing inspection of the spatial distribution of earthquake hypocentres, which add information about their density and can help in the cluster analysis. Fig. 13 illustrates this for the investigated catalogues. The colour scale represents the hypocentre density defined as the number of earthquakes, $N_{\mathrm{EQ},}$ in a sphere of radius 1 km, $V_s$. The number of events displayed in Figs 13(b) and (c) differs from Figs 8 and 11 because we have only visualized hypocentres for which the corresponding density is above a threshold of five events per $V_s$. This threshold simply avoids that many irrelevant events (in low-density areas) prevent the view of areas of interest (those that have high density). Fig. 13 highlights that density information is fundamental not only to locate the most active regions, but also to quantify the intensity of seismicity. In particular, the maximum value of the density, $Z_{\mathrm{max}}$, is approximately equal to 5, 30 and 45 for the synthetic data set and our extracted catalogues of the 2016 Kumamoto and Central Italy sequences, respectively. Interestingly, the most active regions in both investigated real cases do not include the largest earthquakes. We can use $Z_{\mathrm{max}}$ to find solutions in the crossover region. In Fig. 14, we propose a diagram that shows the main steps to obtain such solutions without performing an extensive exploration of the phase diagram.

Given an earthquake catalogue for a region of interest, the first step consists in converting the map units into km units (e.g. UTM coordinates), because an orthogonal coordinate system is required to correctly measure Euclidean distances between hypocentres. Then, the density of hypocentres needs to be computed and visualized for every event to infer the spatial distribution of hypocentres and obtain the $Z_{\mathrm{max}}$ and the geometry of dense regions. The cluster analysis starts with $\varepsilon = 1$ km and $Z = Z_{\mathrm{max}}$. If strong anisotropic structures characterize the data set, data scaling is suggested and to start cluster analysis with $\varepsilon = 0.5$ km and $Z = 2Z_{\mathrm{max}}$. We note that the evaluation of anisotropic structures is done retrospectively considering the spatial distribution of the whole seismic sequence and when the depth range of hypocentres significantly differs from their horizontal range. Such an evaluation becomes more feasible when the catalogue increases in size, and cannot be done at the beginning of a sequence.

These initial choices for $\varepsilon$ and $Z$ were motivated by investigating several catalogues (also catalogues not discussed here), because they proved effective in providing solutions in the crossover region or its proximity. Regarding the earthquake density definition, changing the sphere size $V_s$ does not change the spatial distribution of points in Fig. 13, but only their colour. Generally, for an increasing sphere radius, the earthquake density decreases because the volume increases faster than $N_{\mathrm{EQ}}$. Thus, if density values decrease, also $Z_{\mathrm{max}}$ decreases. Consequently, small values of $Z$ in the initial configuration require small values of $\varepsilon$ to find solutions in the crossover region, which however are associated with the undesired feature of a big amount of noise points. In addition, uncertainties of hypocentral depths are typically of the order of 1 km, so that smaller values of $V_s$ might not be useful. Regarding the initial choice for $\varepsilon$, we set the radius of the spherical neighbourhood to 1 km because earthquakes usually occur at a depth of about 0–15 km. If $\varepsilon$ is larger than 1 km, DBSCAN likely returns solutions with vertical extensions of the clusters spanning the entire depth range (see Fig. 9), not allowing to distinguish shallow and deep structures. Note that larger $\varepsilon$ require larger $Z$ to avoid solutions with only one or two huge clusters but instead remain in the crossover region. Besides, larger $\varepsilon$ and $Z$ result in more convex cluster shapes.

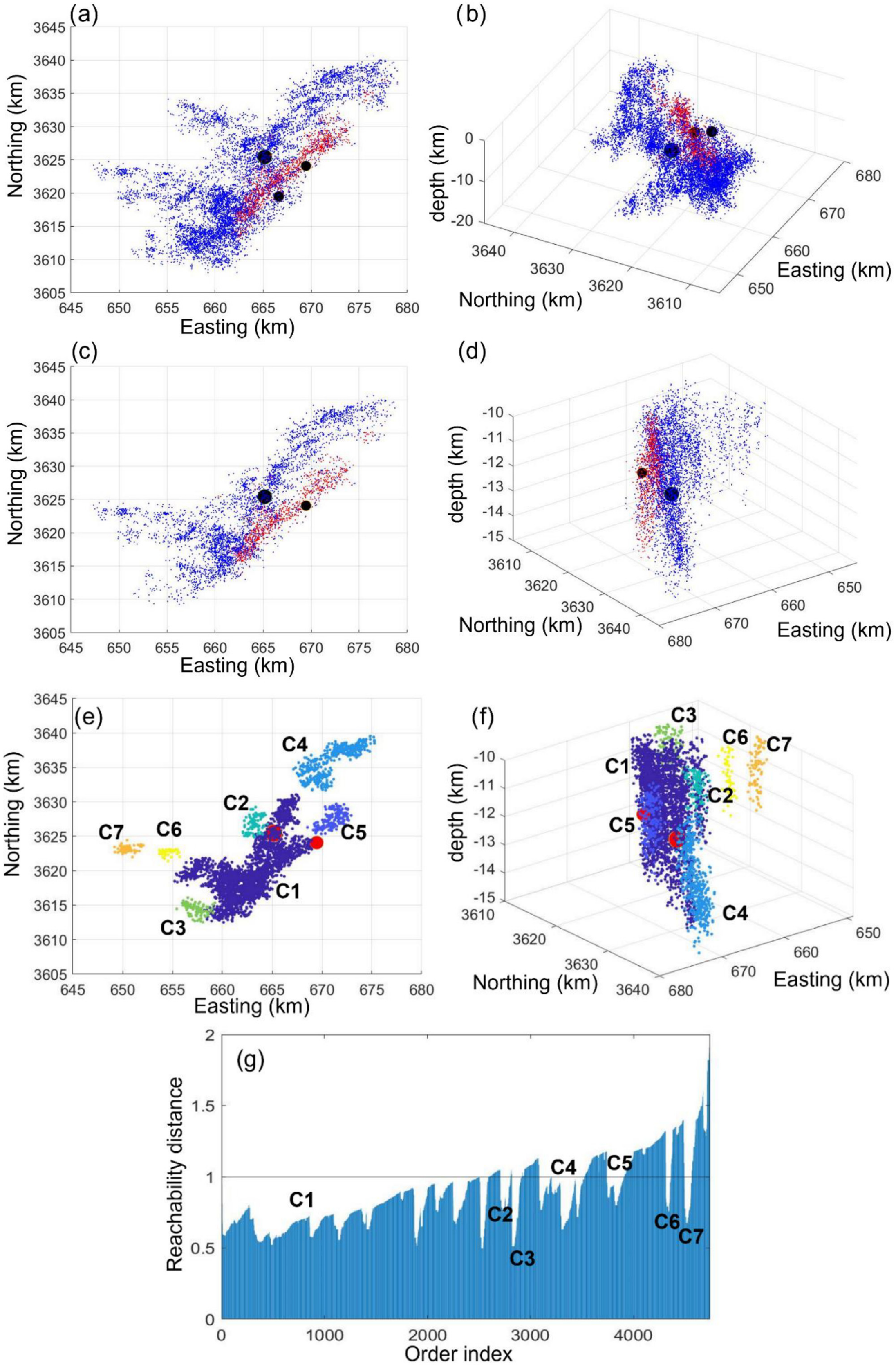

**Figure 15.** Performing cluster analysis for the largest cluster of the 2016 Kumamoto sequence shown in Fig. 9(b), totalling 9414 events. (a) Map view and (b) 3-D view. (c) Map view and (d) 3-D view using data in the depth range 10–15 km, totalling 4742 events. Events occurring between the two largest events (*M*6.5 and *M*7.3) are shown in red, the rest in blue (16 April 2016 to 31 August 2016). Black markers represent the three largest events of the sequence. (e–g) Cluster analysis for $Z = 25$ applied to the depth-constrained subset shown in panels (c) and (d). The map view and 3-D view in panels (e) and (f) relate to a DBSCAN solution using $\varepsilon = 1$ km, which is indicated by a horizontal line in the reachability plot in panel (g). Noise points are not shown.

Iteratively, the test condition of the cluster solution belonging to the crossover region is checked by computing the number of noise points and the size of the biggest cluster $C_b$. If this condition is not satisfied, comparing $C_b$ to the size of the other clusters defines how to change the $Z$ value: if $C_b$ is much larger than the sum of sizes of other clusters, it must be decreased, otherwise decreased. Once a solution in the crossover region has been obtained, the reachability plot is computed for this $Z$ with the OPTICS algorithm. Given this visualization of the nested hierarchical structure, the $\varepsilon$ value is determined by the desired hierarchy level. This $\varepsilon$ completes the parameter set for DBSCAN to obtain the final cluster solution.

### 6.2 Application to a real case

The proposed recipe (Fig. 14) is applied to the largest cluster of the Kumamoto sequence obtained from the cluster solution shown in Fig. 9(b). Since this group of earthquakes contains the three largest earthquakes of the sequence, we want to investigate if they may belong to different partitions. Fig. 15 emphasizes two periods of the data set: between the two largest events, a $M$6.5 and a $M$7.3 (14 and 15 April 2016, show in red) and everything after (16 April to 31 August 2016, show in blue). The earlier events represent a well-known, preferred alignment (Yano & Matsubara 2017), which also persists in the depth range of 10–15 km (Figs 15c and d), and are characterized by a spatial distribution that resembles a branched structure. The two largest events initiated at similar depths and belong to two different branches. From our cluster analysis, we find that both events belong to the same cluster. By applying our proposed procedure only using hypocentres in the depth range of 10–15 km, we again find that the two largest events belong to the same cluster (see Figs 15e and f), supporting the findings of previous studies (Sugito *et al.* 2016; Yue *et al.* 2017). The reachability plot nicely reflects the hierarchy of the data set and its characteristic structures (see Fig. 15g). In particular, a horizontal cut at $\varepsilon = 1$ km crosses seven valleys corresponding to the seven clusters shown in Figs 15(e) and (f) as retrieved by DBSCAN. From the reachability plot, we can infer the density and size of each cluster and already presume what happens when we change the $\varepsilon$ threshold: a small increase in $\varepsilon$ will cause C2 and C3 to be included in C1, whereas a decrease in $\varepsilon$ leads to a splitting of C1 into smaller clusters due to several smaller valleys contained in it.

## 7 CONCLUSIONS

We performed 3-D spatial cluster analyses of seismic sequences by applying the popular density-based clustering algorithms DBSCAN in combination with the reachability plot of the OPTICS algorithm to synthetic and real hypocentre catalogues. Our analyses address the influence of the input parameters on cluster solutions and provide suggestions for exploring earthquake catalogues more appropriately.

Several studies that applied DBSCAN to earthquake catalogues using hypocentre locations, occurrence times, and/or focal mechanisms all remain vague about the choice of input parameters. Here we showed that such choices are crucial to discover regions of interest for a subsequent analysis and to identify meaningful tectonic structures that were activated in a seismic sequence.

We showed that varying the DBSCAN parameters leads to a variety of cluster solutions that can be classified into five different regions of the phase diagram. Cluster solutions inside the so-called crossover region are the most representative candidates for characterizing 3-D spatial features of seismic sequences, because they represent the individual structures as large clusters. To identify these solutions, we proposed a tentative recipe that includes a density representation of earthquakes and investigating the nested clustering structure.

We draw the following conclusions from our analyses: (i) using DB algorithms for cluster analysis requires utmost care in the selection of input parameters and the type to which the considered solution belongs to; (ii) graphically representing the spatial distribution of hypocentres and their density helps to select the input parameters and (iii) only cluster solutions in the crossover region represent information about the largest characteristic structures of a data set. Investigating such solutions can provide insight into the main features of a seismic sequence (e.g. its 3-D fault geometry) and open new perspectives for studying the spatiotemporal evolution of fault systems.

## ACKNOWLEDGMENTS

We thank editor Andrew Valentine, reviewer Simone Cesca and an anonymous reviewer for their careful reading and valuable feedback, which helped us to considerably improve the quality of the manuscript. This study was partially supported by the ‘Real-time Earthquake Risk Reduction for a Resilient Europe’ (RISE) project, funded by the European Union’s Horizon 2020 research and innovation program (Grant Agreement No. 821115).

## DATA AVAILABILITY

The synthetic data for this paper are available by contacting the corresponding author at ’ester.piegari@unina.it’.

## REFERENCES

Abdideh, M. & Ameri, A., 2020. Cluster analysis of petrophysical and geological parameters for separating the electrofacies of a gas carbonate reservoir sequence, *Nat. Resour. Res.,* **29,** 1843–1856.

Aggarwal, C.C. & Reddy, C.K., 2013. *Data Clustering: Algorithms and Applications,* 1st edn, Chapman & Hall/CRC.

Ankerst, M., Breunig, M.M., Kriegel, H.P. & Sander, J., 1999. OPTICS: ordering points to identify the clustering structure, *ACM SIGMOD Record,* **28**(2), 49–60.

Ansari, A., Noorzad, A. & Zafarani, H., 2009. Clustering analysis of the seismic catalog of Iran, *Comput. Geosci.,* **35,** 475–486.

Bhattacharya, S., 2021. *A Primer on Machine Learning in Subsurface Geosciences,* 1st edn., Vol., **1,** pp. 1–172, Springer.

Cesca, S., Sen, A.T. & Dahm, T., 2014. Seismicity monitoring by cluster analysis of moment tensors, *Geophys J. Int.,* **196,** 1813–1826.

Cesca, S., 2020. Seiscloud, a tool for density-based seismicity clustering and visualization, *J. Seismol.,* **24,** 443–457.

Ester, M., Kriegel, H. P., Sander, J. & Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise, *KDD-96 Proc.,* **34,** 226–231.

Fan, Z. & Xu, X., 2019. Application and visualization of typical clustering algorithms in seismic data analysis, *Proc. Comp. Sci.,* **151,** 171–178.

Herrmann, M., Piegari, E. & Marzocchi, W., 2021. b-Value of what? Complex behavior of the magnitude distribution during and within the 2016-2017 central Italy sequence, *Nature Communications,* (accepted, preprint). DOI:10.21203/rs.3.rs-1210699/v1.

Jain, A.K., Murty, M.N. & Flynn, P.J., 1999. Data clustering: a review, *ACM Comput. Surv.,* **31,** 264–323.

Kamer, Y., Ouillon, G. & Sornette, D., 2020. Fault network reconstruction using agglomerative clustering: applications to southern Californian seismicity, *Nat. Hazards Earth Syst. Sci.,* **20,** 3611–3625.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Ali Babaie, H. & Kumar, V., 2019. Machine learning for the geosciences: challenges and opportunities, *IEEE Trans. Knowled. Data Eng.,* **31,** 1544.

Kaufman, L. & Rousseeuw, P.J., 1990. *Finding Groups in Data: An Introduction to Cluster Analysis,* Wiley Series in Probability and Statistics, Wiley.

Konstantaras, A.J., Katsifarakis, E., Maravelakis, E., Skounakis, E., Kokkinos, E. & Karapidakis, E., 2012. Intelligent spatial-clustering of seismicity in the vicinity of the hellenic seismic arc, *Earth Sci. Res.,* **1,** 2, E-ISSN 1927-0550.

Lindsey, C.R., Neupaneb, G., Spycher, N., Fairley, J.P., Dobson, P., Wood, T., McLing, T. & Conrad, M., 2018. Cluster analysis as a tool for evaluating the exploration potential of known geothermal resource areas, *Geothermics,* **72,** 358–370.

Lyra, G.B., Oliveira-Júnior, J.F. & Zeri, M., 2014. Cluster analysis applied to the spatial and temporal variability of monthly rainfall in Alagoas state, Northeast of Brazil, *Int. J. Climatol.,* **34,** 3546–3558.

Mehta, P., Bukov, M., Wang, C.H., Day, A.G.R., Richardson, C., Fisher, C.K. & Schwab, D.J., 2019. A high-bias, low-variance introduction to Machine Learning for physicists, *Phys. Rep.,* **810,** 1–124.

Ouillon, G., Ducorbier, C. & Sornette, D., 2008. Automatic reconstruction of fault networks from seismicity catalogs: three-dimensional optimal anisotropic dynamic clustering, *J. geophys. Res.,* **113**(B1), doi:10.1029/2007JB005032.

Ouillon, G. & Sournette, D., 2011. Segmentation of fault networks determined from spatial clustering of earthquakes, *J. geophys. Res.,* **116**(B2), doi:10.1029/2010JB007752.

Petersen, G.M., Niemz, P., Cesca, S., Mouslopoulou, V. & Bocchini, G.M., 2021. Clusty, the waveform-based network similarity clustering toolbox: concept and application to image complex faulting offshore Zakynthos (Greece), *Geophys. J. Int.,* **224,** 2044–2059.

Schoenball, M. & Ellsworth, W.L., 2017. A systematic assessment of the spatiotemporal evolution of fault activation through induced seismicity in Oklahoma and Southern Kansas, *J. geophys. Res.,* **122,** 10 189–10 206.

Sugito, N., Goto, H., Kumahara, Y., Tsutsumi, H., Nakata, T., Kagohara, K., Matsuta, N. & Yoshida, H., 2016. Surface fault ruptures associated with the 14 April foreshock (Mj 6.5) of the 2016 Kumamoto earthquake sequence, southwest Japan, *Earth, Planet Space,* **68,** 170.

Yano, T.E. & Matsubara, M., 2017. Effect of newly refined hypocenter locations on the seismic activity recorded during the 2016 Kumamoto Earthquake sequence, *Earth, Planets Space,* **69,** 74.

Yue, H., Ross, Z.E., Liang, C., Michel, S., Fattahi, H., Fielding, E., Moore, A., Liu, Z. & Jia, B., 2017. The 2016 Kumamoto Mw = 7.0 earthquake: a significant event in a fault–volcano system, *J. geophys. Res.,* **122,** 9166–9183.

Zhang, W., Zhang, Y., Gu, X., Wu, C. & Han, L., 2022. *Application of Soft Computing, Machine Learning, Deep Learning and Optimizations in Geoengineering and Geoscience,* 1st edn, Vol., **1,** pp. 1–138, Springer.

Sander, J., Qin, X., Lu, Z., Niu, N. & Kovarsky, A., 2003. *Automatic extraction of clusters from hierarchical clustering representations, in Pacific-Asia Conference on Knowledge Discovery and Data Mining,* Springer, pp.75–87.