Topological comparison between the stochastic and the nearest-neighbour declustering methods through network analysis

Antonella Peresan[1], Elisa Varini[2], and Jiancang Zhuang[3]

[1]National Institute of Oceanography and Experimental Geophysics
[2]Istituto di Matematica Applicata e Tecnologie Informatiche
[3]Institute of Statistical Mathematics

November 26, 2022

Abstract

Earthquake clustering is a relevant feature of seismic catalogs, both in time and space. Several methodologies for earthquake cluster identification have been proposed in the literature in order to characterise clustering properties and to analyse background seismicity. We consider two recent data-driven declustering techniques, one is based on nearest-neighbor distance and the other on a stochastic point process. These two methods use different underlying assumptions and lead to different classifications of earthquakes into background events and secondary events. We investigated the classification similarities by exploiting graph representations of earthquake clusters and tools from network analysis. We found that the two declustering algorithms produce similar partitions of the earthquake catalog into background events and earthquake clusters, but they may differ in the identified topological structure of the clusters. Especially the clusters obtained from the stochastic method have a deeper complexity than the clusters from the nearest-neighbor method. All of these similarities and differences can be robustly recognised and quantified by the outdegree centrality and closeness centrality measures from network analysis.

Topological comparison between the stochastic and the nearest-neighbour declustering methods through network analysis

E. Varini[1], A. Peresan[2], and J. Zhuang [3]

[1]C.N.R. - Institute of Applied Mathematics and Information Technology
via Bassini 15, 20133 Milano, Italy
[2]National Institute of Oceanography and Experimental Geophysics
CRS-OGS, via Treviso 55, 33100 Udine, Italy
[3]Institute of Statistical Mathematics, Research Organization of Information and Systems
10-3 Midori-Cho, Tachikawa, Tokyo 190-8562, Japan

Key Points:

- Two recent data-driven declustering methods are compared, one based on nearest-neighbor distance and one on the ETAS model
- Similarities in classification and in earthquake clusters are investigated by tree graphs and tools from network analysis
- Obtained clusters are consistent, though nearest-neighbor method usually provides simpler structures than stochastic declustering method

Corresponding author: Antonella Peresan, `aperesan@inogs.it`

Abstract

Earthquake clustering is a relevant feature of seismic catalogs, both in time and space. Several methodologies for earthquake cluster identification have been proposed in the literature in order to characterize clustering properties and to analyze background seismicity. We consider two recent data-driven declustering techniques, one is based on nearest-neighbor distance and the other on a stochastic point process. These two methods use different underlying assumptions and lead to different classifications of earthquakes into background events and secondary events. We investigated the classification similarities by exploiting graph representations of earthquake clusters and tools from network analysis. We found that the two declustering algorithms produce similar partitions of the earthquake catalog into background events and earthquake clusters, but they may differ in the identified topological structure of the clusters. Especially the clusters obtained from the stochastic method have a deeper complexity than the clusters from the nearest-neighbor method. All of these similarities and differences can be robustly recognized and quantified by the outdegree centrality and closeness centrality measures from network analysis.

Plain Language Summary

Clustering, in both space and time, is a widely recognised feature of seismicity. An adequate identification of earthquake clusters allows splitting seismicity into background and clustered events (e.g. aftershocks), and is an essential step in several studies, ranging from seismic hazard assessment to long- and short-term earthquake forecasting. Also, the space-time patterns of identified clusters may provide useful insights on the structural and dynamic tectonic features of a region. Among the several methods proposed so far to identify and characterise seismic clusters, we consider two recent data-driven declustering techniques, one based on nearest-neighbor distance and the other on a stochastic point process. These two methods use different underlying assumptions and may lead to different classifications of earthquakes into background events and clustered events. Therefore this study aims to compare their performances, including clusters structure characterisation, by exploiting tree graph representations and tools from network analysis. We found that: (1) the two declustering algorithms produce similar partitions of the earthquake catalog; (2) they may differ in the internal structure outlined for individual clusters, with the nearest-neighbor method usually providing simpler structures than stochastic declustering method; and (3) these features can be robustly quantified by centrality measures widely used in network analysis.

1 Introduction

Short-term earthquake clustering is a widely recognised feature of seismic activity, which eventually complicates the analysis of seismicity, especially when we evaluate long-term earthquake risks. An ideal partition of an earthquake catalog is into two subsets of events, referred as background seismicity and secondary seismicity, respectively. Background events are intended as spontaneous or independent earthquakes; secondary events are considered as triggered by other earthquakes, therefore manifestly dependent events, generally forming spatio-temporal clusters and producing a significant increase of the seismicity rate. It is often supposed that background events are representative of the long-term spatio-temporal behaviour of seismicity in a region. Poisson model, renewal model, and stress release model are typically assumed as suitable stochastic processes to describe background events (Vere-Jones, 1978; Rotondi, 2010; Rotondi & Varini, 2019). On the other hand, the identification of earthquake clusters is important to understand and to forecast the spatio-temporal evolution of a seismic sequence on short time scales; the Omori-Utsu formula, the Epidemic-Type Aftershock-Sequence model and its extensions are typically used to model earthquake clusters, such as swarms or aftershock sequences (Ogata, 1998).

However, an objective and commonly agreed method for separating earthquake clusters from each other and from the background seismicity is critical. There are several declustering algorithms in the literature (van Stiphout et al. (2012) and references therein), which are likely to identify different earthquake clusters and, accordingly, different declustered versions of a catalog.

The most used declustering algorithms are the mainshock-window method by Gardner and Knopoff (1974) and the linked-window method by Reasenberg (1985), due to their simplicity and software availability: the former removes all earthquakes in a certain space-time window around each suitably defined mainshock; the latter performs scans within certain space-time windows of each event in the catalog in order to form clusters of events and then replace each cluster with a single event (e.g. the first, or the larger). The drawback of window methods is that they require some subjective choices, such as the definition of mainshock or the dimensions of the space-time windows, which might seriously influence the results.

Among the valid alternatives to window-based methods, we focus on two recently proposed declustering algorithms: the nearest-neighbor method by Zaliapin and Ben-Zion (2013, 2016) and the stochastic declustering method by Zhuang et al. (2002, 2004) and Zhuang (2006). They have been the subject of several recent papers to which the readers can refer for additional details (e.g. Peresan and Gentili (2018), Zhang and Shearer (2016), Nandan et al. (2019) for the nearest-neighbor method and Davoudi et al. (2018), Zhuang et al. (2005), Talbi et al. (2013) for the stochastic declustering method). Both methods are data-driven and can be satisfactorily applied to decompose the seismic catalog into background seismicity and sequences of clustered earthquakes.

In addition, both methods allow studying the internal structure of the identified sequences (or several probable realizations of it, in the case of stochastic declustering method) since they provide the connections between events forming each cluster.

For example Wang et al. (2010) compared the Reasenberg's, Kagan's, and Zhuang's methods; Talbi et al. (2013) dealt with the methods of Gardner and Knopoff, Reasenberg, and stochastic declustering. However, in-depth comparison was carried out so far between these more recent methods.

This study focuses on the nearest-neighbor and the stochastic declustering algorithms because they can be used not only to identify background seismicity, but also to investigate the properties and internal structure of seismic clusters (Zhuang et al., 2004; Guo et al., 2015, 2017). The aim is to compare the features of clusters identified by the two algorithms exploiting tools and measurements from network analysis. Moreover the research aims to improve our understanding of the role of moderate earthquakes in the region, providing in the meanwhile a characterization of seismicity patterns and their variations at short-term space-time scales.

This article is organised as follows: a short description of both declustering methods is given in Section 2; the seismicity of Northeastern Italy and the related earthquake data sets, to be used as a case study, are introduced in Section 3. Section 4 gives the computational details to fit the declustering algorithms to the data and then it provides a global comparison of the background seismicity and earthquakes clusters obtained from the two methods. Section 5 deals with the analysis of the clusters structure by exploiting graphical tools and quantitative methods from network theory. Conclusions are drawn in Section 6.

2 Declustering Algorithms Under Examination

Given a catalog $\{(t_i, x_i, y_i, m_i) \ : \ i \ = \ 1, \cdots n\}$, where n is the total number of earthquakes, and t_i, $(x_i,\, y_i)$, and m_i are the occurrence time, epicentral location, and magnitude, respectively, the numerical algorithms of these two declustering methods are given in following subsections.

2.1 Nearest-neighbor algorithm (NN)

This approach is based on the NN-distance (nearest-neighbor distance) between two earthquakes in the space-time-energy domain, as defined by Baiesi and Paczuski (2004):

$$\eta_{ij} = (t_j - t_i) r_{ij}^{d_f} 10^{-bm_i} \tag{1}$$

where $t_i \ < \ t_j$ and r_{ij} is the spatial distance between events i and j. This metric exploits the following statistical properties of seismicity to quantify the correlation between earthquakes: the inter-occurrence time, the fractal dimension of the hypocentres distribution, and the Gutenberg–Richter law. There are only two unknown parameters, namely fractal dimension d_f and b-value, which are jointly and robustly estimated by the Unified Scaling Law for Earthquakes (USLE) method (Nekrasova et al., 2011); a separation distance η_0 is also estimated in order to identify clusters of events (details in Peresan and Gentili (2018)).

The nearest-neighbor distance η_{ij} can be equivalently decomposed into the corresponding rescaled space (R_{ij}) and rescaled time (T_{ij}) distances from the parent to its offspring event (Zaliapin et al., 2008), namely η_{ij}=T_{ij} R_{ij}, where: $T_{ij} \ = \ t_i j 10^{-bm_i/2}$ and $R_{ij} = r_{ij}^{d_f} 10^{-bm_i/2}$.

Accordingly each event j is connected to its nearest-neighbor $i = \arg\min_{k:k<j} \ \eta_{kj}$. Then, by removing all connections η_{ij} such that $\eta_{ij} > \eta_0$, the earthquake catalog is unambiguously partitioned on distinct clusters, each containing at least one event (Zaliapin & Ben-Zion, 2013, 2016). The maximum magnitude event of each cluster is labelled as background event and the remaining events of the clusters are included in the secondary seismicity.

2.2 Stochastic declustering algorithm (SD)

This approach is based on the space-time ETAS (epidemic-type aftershock sequence) model (Ogata, 1998), a branching point process defined by its intensity function conditional on the observation history $\mathcal{H}_t$:

$$\lambda(t, x, y \mid \mathcal{H}_t) = \mu(x, y) + \sum_{k:t_k<t} g(t - t_k, x - x_k, y - y_k; m_k) \tag{2}$$

where $\mu(x, y)$ is the spatial background rate of a time-homogeneous Poisson process and, at time t, $g(t-t_k, x-x_k, y-y_k;\, m_k)$ is the contribution to seismic hazard due to triggering effects of the k-th earthquake. The explicit functional forms in Eq. (1) are the following:

$$\begin{aligned} \mu(x,y) &= \nu \cdot u(x,y) \\ g(t,x,y;m) &= Ae^{\alpha(m-m_0)} \cdot (p-1)c^{p-1}\,(t+c)^{-p} \cdot \\ &\cdot \ \frac{1}{2\pi d e^{\alpha(m-m_0)}} \exp\left\{-\frac{1}{2}\frac{x^2+y^2}{de^{\alpha(m-m_0)}}\right\} \end{aligned} \tag{3}$$

where $\nu, A, c, \alpha, p, d, q, \gamma$ are positive parameters and $u(x, y)$ is an unknown spatial function (Zhuang et al., 2002). An iterative algorithm simultaneously provides the maximum likelihood estimates of the eight model parameters and a non parametric kernel estimate of the spatial background rate.

According to point process theory, the probability that event j is generated by the background process is $\varphi_j = \mu(x_j, y_j)/\lambda(t_j, x_j, y_j \mid \mathcal{H}_{t_j})$, and the probability that it is triggered from previous event i is $\rho_{ij} = g(t_j - t_i, x_j - x_i, y_j - y_i;\, m_i)/\lambda(t_j, x_j, y_j \mid \mathcal{H}_{t_j})$. Thinning (sampling) the process according to these probabilities allows splitting the catalog into background events and triggered events, and also setting connections between triggering and triggered events (Zhuang et al., 2002, 2004; Zhuang, 2006). The first event of each cluster is labelled as background event, which may not be the maximum magnitude event within the cluster; it is named ancestor because it represent the earthquake that triggers others in the cluster. The remaining events of the clusters are included in the secondary seismicity and are called descendants. Unlike NN method, SD algorithm can provide many declustered catalogs by simulation.

2.3 Differences and connections between the NN and SD methods

Notably the two methods have a different definition of background events: while NN assigns to the background seismicity the largest event from each cluster (i.e. the mainshock), SD assigns to it the first event of the cluster (not necessarily the mainshock); therefore the declustered catalogs may differ, particularly when foreshocks are identified.

The NN declustering method has some connections with the stochastic declustering method. Firstly, the NN-distance η_{ij} takes a similar form as $1/g(t_j - t_i, x_j - x_i, y_j - y_i; m_i)$. If we consider an ETAS-like model with the conditional intensity

$$\lambda_0(t, x, y \mid \mathcal{H}_t) = \mu_0 + A \sum_{i: t_i < t} (t - t_i)^{-1}\, r(x_i, y_i; x, y)^{-d_f}\, 10^{b m_i}, \tag{4}$$

where $r(x, y; x', y')$ is the Euclidean distance between (x, y) and (x', y'), the quantity $\rho_{ij}^{(0)} = A(t_j - t_i)^{-1}\, r(x_i, y_i; x_j, y_j)^{-d_f}\, 10^{b m_i}/\lambda_0(t_j, x_j, y_j \mid \mathcal{H}_{t_j})$ is proportional to the reciprocal of η_{ij}. In this new model the background rate μ_0 is an unknown constant and A is also unknown, which are in fact connected to the NN method through $\eta_0 = A/\mu_0$.

The basic differences between these two methods are clear.

1. The NN method classifies the clusters based on the minimum distance η_{ij}, which corresponds, for each event, to the largest probability ρ_{ij}, among the probabilities that the event is from background seismicity or triggered by one of the previous events, according to the model in (4). The SD method, on the other side, makes use of the full probability distribution of ρ_{ij}, leading to several possible cluster classifications. As a rule, a probabilistic-manner resampling is recommended to reflect the uncertainty in the classification of the family tree; however, SD can also classify the clusters based on the maximum probability ρ_{ij}, in the same manner as the NN method.
2. The NN method implicitly estimates the classification parameter η_0, approximately according to the separation between two modes of the NN-distance distribution; the two remaining parameters, namely the b-value and the fractal dimension of epicenters, are estimated independently, and used as a priori input information. No explicit assumption is made about the background seismicity, which can be inhomogeneous in space (Zaliapin et al., 2008) and possibly also in time. The SD method is based on the ETAS model, where the model parameters and the optimal non-homogeneous background rate are estimated through MLE procedure, thus providing a summary description of the considered data set. Accordingly, the NN method allows for a rather fast and robust identification of clusters, with less stringent requirements about the catalog completeness and homogeneity, while the SD provides a more detailed, specific and sophisticated data description and classification, requiring high-quality catalogs.

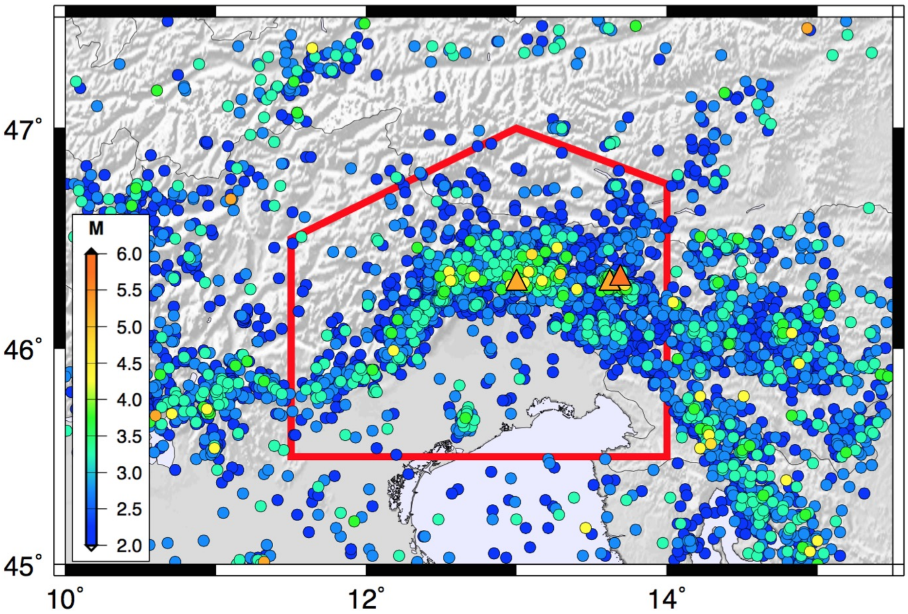

Figure 1. The study region (red polygon) and the epicentres of the earthquakes occurred since 1977. The strongest earthquakes, with magnitude larger than 5.0, are marked by triangles.

3 Study Region and Data

The study region, which comprises North-Eastern Italy and Western Slovenia, is located along the northern edge of Adria micro-plate, at the transition between Alpine and Dinaric fault systems. Earthquakes are mostly shallow (up to 12 km), and are prevalently of thrust type to the west and strike-slip to the east. The instrumental seismicity recorded during about 40 years, prevalently consists of low to moderate earthquakes, only occasionally exceeding magnitude 4.0; the largest earthquake was recorded in 1998 ($M5.6$), nearby the border between Italy and Slovenia. Despite the moderate seismic activity that has recently affected this region, the historical seismicity testifies to its high seismic hazard and high vulnerability. According to the Italian Parametric Earthquake Catalogue CPTI15 (Rovida et al., 2014), at least six destructive earthquakes with magnitude larger that 6.0 hit that area in the past millennium, the most recent one being the M6.4 1976 Friuli earthquake (Slejko et al., 1999).

To investigate the clustering features in the study region, we consider the earthquake bulletins compiled at the National Institute of Oceanography and Experimental Geophysics, which include 27353 earthquakes occurred in the time span from 7 May 1977 to 30 April 2018, and with duration magnitude up to $M_d5.6$. Fig. 1 shows the distribution of earthquake epicentres, as well as the study region, which is a polygonal area delimited by the following five vertices: $(11.5, 45.5)$; $(11.5, 46.5)$; $(13.0, 47.0)$; $(14.0, 46.75)$; $(14.0, 45.5)$. A detailed analysis of the data completeness in space and time, including delineation of the study region and estimation of the scaling parameters of seismicity, was carried out by Peresan and Gentili (2018). Within the identified area (red polygon in Fig. 1), the bulletins can be considered fairly complete for magnitudes $M \geq 2.0$ during the whole time span 1977-2018 (Fig. 2), except for a time interval between December 1990 and May 1991, when data acquisition was interrupted due to a fire accident (Fig. 3, bottom panel).

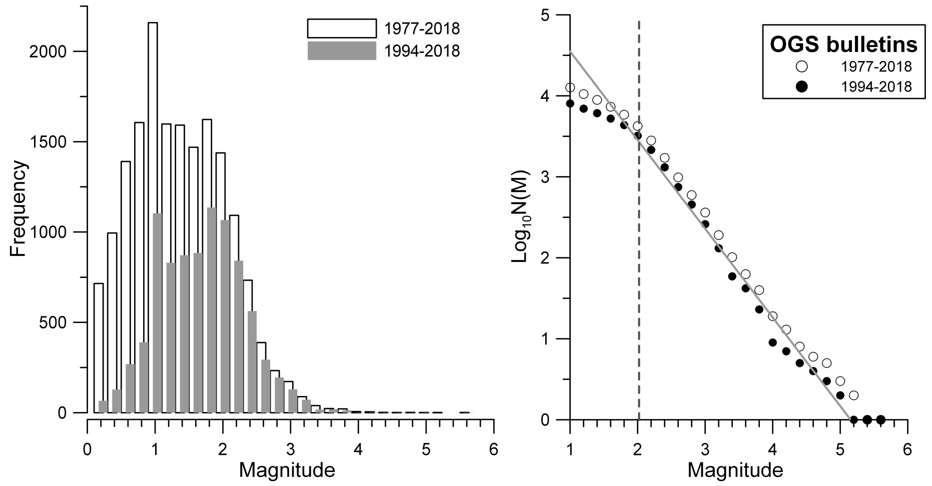

Figure 2. Histogram on magnitude (left) and estimated Gutenberg–Richter law (right) for the full (1977-2018) and the complete (1994-2018) data sets.

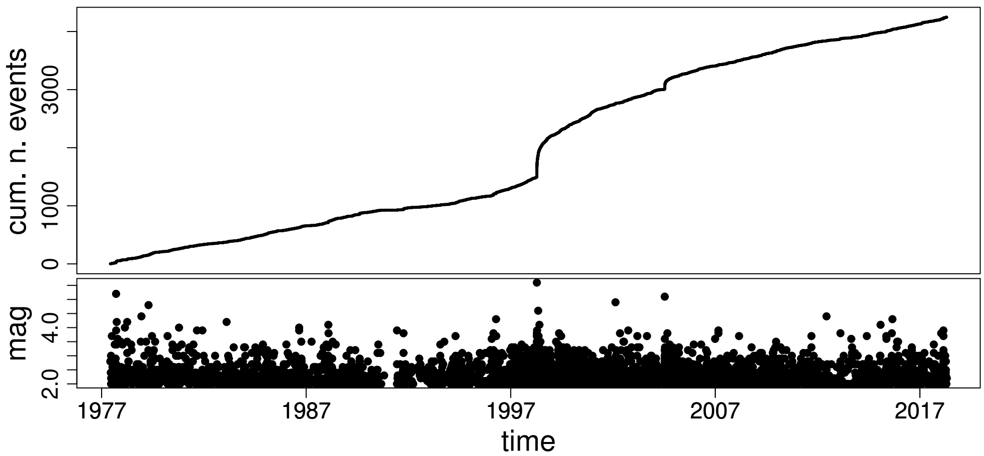

Figure 3. Full data set (1977-2018, $M \geq 2.0$): cumulative number of events versus time (top) and magnitude versus time (bottom).

Since the data are certainly incomplete in the early 1990s, two subsets of the catalog are considered hereinafter. The former, referred to as the complete data set, includes all the 3219 earthquakes having magnitude at least 2.0 and occurred since 1994; the statistical completeness and the b-value of the Gutenberg-Richter law have been estimated using only this part of the data (Fig. 2). The latter subset, named the full data set, is obtained from the catalog by setting a minimum threshold magnitude equal to 2.0; therefore, it covers the entire time span from 1977 to 2018 and it includes 4247 earthquakes (Fig. 3).

4 Declustering Outputs

4.1 Declustering settings and global features of the two declustered catalogs

Both NN and SD algorithms are applied in order to obtain declustered versions of the full data set.

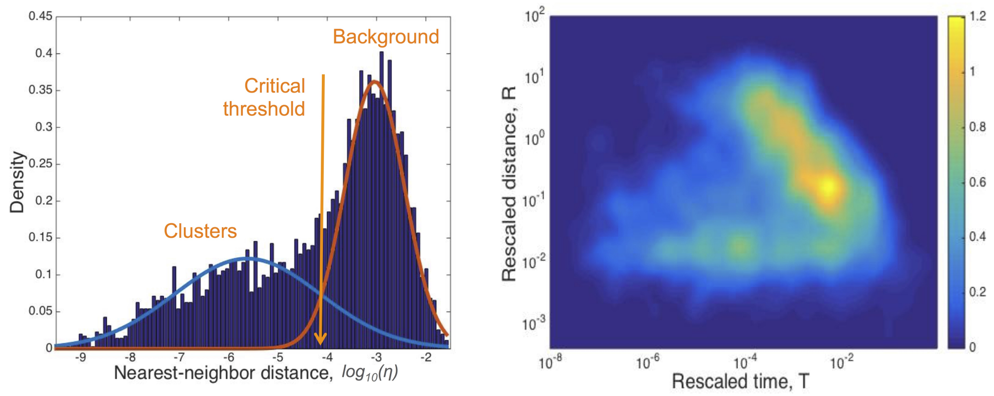

Figure 4. Distributions of NN-distances, between each event and its nearest neighbour, estimated for earthquakes with M≥ 2.0 in 1977-2018. Left column: 1D density distribution of $\log \eta$, with estimated Gaussian densities for clustered (blue) and background (red) components. Right column: 2D joint distribution of rescaled space and time distances (R,T).

The scaling parameters of NN-algorithm are simultaneously estimated by the USLE method and their values are $b = 0.9$ and $d_f = 1.1$ as defined in Peresan and Gentili (2018); the logarithm of the separation distance is automatically set equal to $\log \eta_0 = -4.1$ (Fig. 4).

Based on these parameters, the NN-algorithm delivers its partition of the data set, which is hereinafter referred to as the NN-catalog. The background seismicity turns out to be composed by the isolated events (singles) and the largest event of each cluster (i.e. the mainshocks, the number of which equals the number of clusters); all other events belong to the secondary seismicity. Table 1 (top) summarizes the NN-catalog by providing the number of events assigned to background seismicity and to secondary seismicity, as well as the number of isolated events (singles), the number of identified earthquake clusters, and the total number of events that temporally precede/follow the strongest earthquake that occurred in their own cluster (here conventionally referred to as foreshocks and aftershocks).

As for the SD-algorithm, the complete data set (which ranges from 1994 to 2018) has been used for the maximum likelihood estimation of ETAS parameters, by assuming that the past history $\mathcal{H}_t$ of the process is given by the full data set (which ranges from 1977 to 2018). The following estimates of the ETAS parameters are thus given: $\nu = 0.6772$, $A = 0.6656$, $c = 0.0146$, $\alpha = 1.5407$, $p = 1.0378$, $d = 0.00007$, $q = 2.2527$, and $\gamma = 0.6239$.

Fig. 5 shows the estimated total rate $\hat{\lambda}(t, x, y \mid \mathcal{H}_t)$ in the region, the ratio between estimated cluster rate and total rate, and the histogram of the estimated background probabilities $\hat{\varphi}_j$ of each event j in the catalog ($j = 1, ..., n$). According to the SD-method, several declustered catalogs can be obtained by simulating the connections between events based on both the estimated background probabilities $\{\hat{\varphi}_j : j = 1, ..., n\}$ and the estimated triggering probabilities $\{\hat{\rho}_{ij} : i, j = 1, ..., n, i < j\}$. To make the comparison between the two declustering methods feasible, we decided to select only one of those simulated catalogs. A reasonable choice is to select the "most probable declustered catalog", which is obtained by retaining the most probable connections between any pair of events according to the estimated background and triggering probabilities; the resulting partition of the full data set is hereinafter referred to as the SD-catalog. Table 1 (bot-

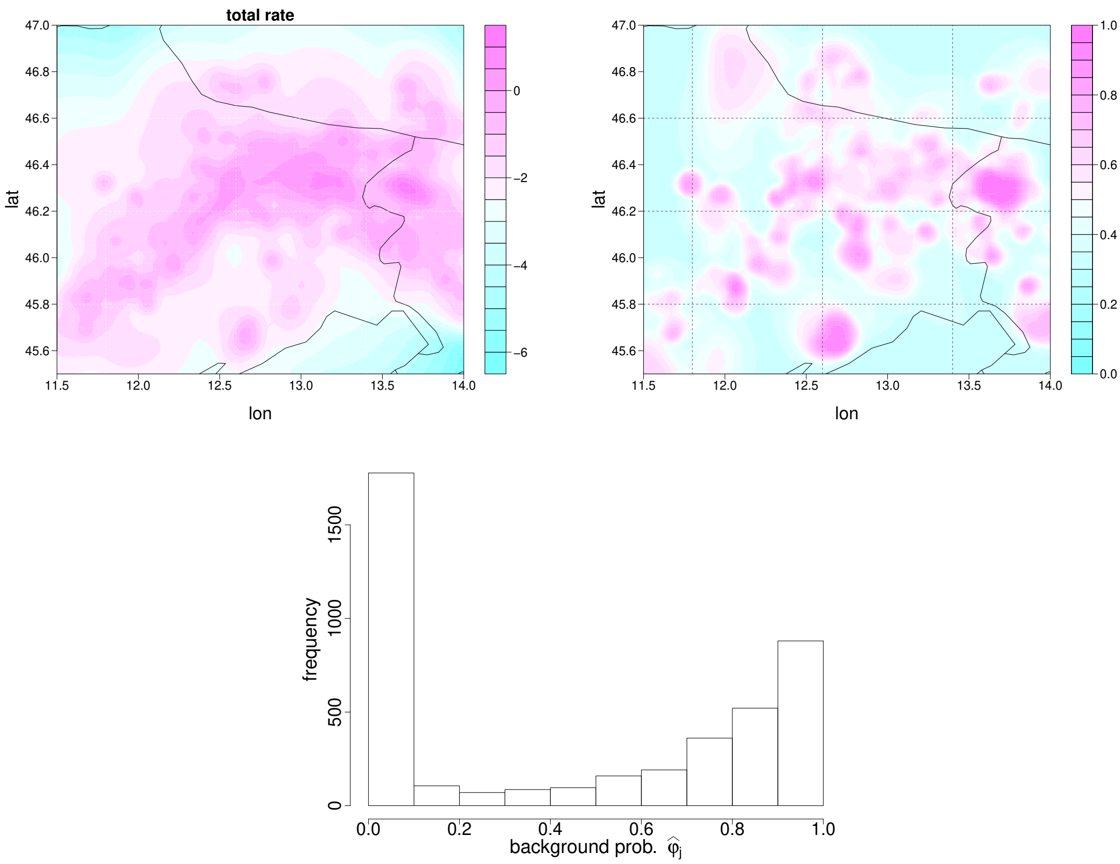

Figure 5. Some results from the SD-algorithm: map of the estimated logarithm of the total rate (top left), map of the ratio between estimated cluster rate and total rate (top right), histogram of the estimated background probabilities for each earthquake in the data set (bottom).

tom) summarizes some counts on the SD-catalog, which turn out fairly consistent with those obtained from NN-method (top of Table 1).

4.2 Comparison of clusters size

The clusters identified by the NN and SD methods are first of all compared in terms of cluster size (i.e. number of events composing the cluster), by assuming clusters are formed by at least two events. The cluster size distributions of NN-catalog and SD-catalog are shown in Fig. 6; in both cases about 95% of the clusters are composed by less than 10 events and about 85% of the identified clusters has even less than 5 events. This means that, for both methods, the number of relevant clusters is quite limited, less that 15% of identified clusters.

It is not obvious to establish a one-to-one correspondence between NN-clusters and SD-clusters, because events from one NN cluster may be separated into different SD clusters. To facilitate the comparison of individual clusters identified by the two declustering methods, we consider the largest earthquake in each cluster as the representative event of the cluster. If a NN-cluster and a SD-cluster have the same representative event, we

Table 1. Summaries of the NN-catalog (top) and the SD-catalog (bottom). Tables report the number of events classified as background/secondary seismicity, the number of single events, the number of clusters, the total number of secondary events that occur before/after the maximum magnitude event in their own cluster (foreshock/aftershock). Percentages with respect to the total number of data are also reported.

NN-catalog				
n.background 2468 (58.11%)		*n.secondary* 1779 (41.89%)		*n.events* 4247 (100%)
n.singles 2123 (49.99%)	*n.clusters* 345 (8.12%)	*n.aftershocks* 1548 (36.45%)	*n.foreshocks* 231 (5.44%)	

SD-catalog				
n.background 2255 (53.10%)		*n.secondary* 1992 (46.90%)		*n.events* 4247 (100%)
n.singles 1884 (44.36%)	*n.clusters* 371 (8.74%)	*n.aftershocks* 1685 (39.67%)	*n.foreshocks* 307 (7.23%)	

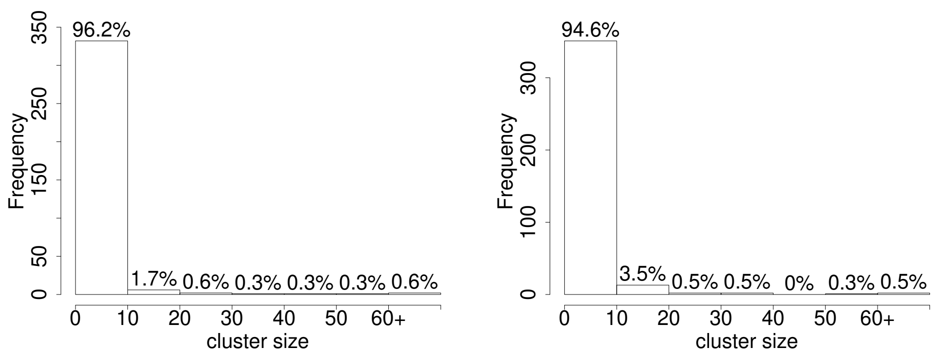

Figure 6. Distribution of the cluster size for the NN-catalog (left) and the SD catalog (right).

Table 2. Selection of large earthquake clusters identified by both declustering methods. The table lists: date and magnitude of the largest event in the cluster; cluster size based on the NN-method and the SD-method; number of events identified by both methods.

largest event	*cluster size* NN	*cluster size* SD	matched events	*largest event*	*cluster size* NN	*cluster size* SD	matched events
12 April 1998 M5.6	720	757	682	20 April 1994 M3.7	21	27	21
12 July 2004 M5.1	201	238	196	14 February 2002 M4.9	19	14	14
13 April 1996 M4.3	52	52	48	5 October 1991 M3.8	18	19	18
16 September 1977 M5.2	41	38	36	12 February 2013 M3.8	15	12	11
1 February 1988 M4.1	34	39	34	25 February 2018 M3.9	15	15	15
18 April 1979 M4.8	28	12	12	29 August 2015 M4.3	5	14	5

say that they are matched clusters. In our application we found exactly 241 pairs of matched clusters.

Table 2 lists some significant clusters, reporting their cluster size according to NN-method and SD-method, as well as the number of events associated by both methods, i.e. the matched events. We notice that, in general, the number of matching events between NN-clusters and SD-clusters is sizable compared to the total cluster size; therefore we can state that the two declustering methods roughly identify the same earthquake clusters. However, this comparison neglects the links between the events, which are established by each declustering method. In section 5 we deepen the comparison between NN-clusters and SD-clusters by analyzing also their internal structure.

5 Topological Structure of Earthquake Clusters

Connections between events of a cluster, as established by the considered declustering methods, allow us to represent the cluster as a network graph. In this section we focus on some centrality measures developed in network theory, which should quantitatively express the way earthquakes get organized within clusters.

5.1 Tree graph representation of clusters

By construction, the identified clusters are organized in rooted time-oriented tree graphs, where each tree root represents the triggering event and the other nodes are the triggered secondary events. For example, Fig. 7 illustrates the tree graph representation of the earthquake cluster occurred in 1988, according to NN-algorithm (left) and SD-algorithm (right). Nodes are joined by edges, which represent the connections between pairs of events. Each node (event), other than the root, is directly connected to its only parent (which triggers the event); in other words, that node is a direct descendant of its parent. The nodes along the path between the root and a node v are named ancestors of node v. The descendants of node v are those nodes of which v is an ancestor.

It is worth noting that the triggering earthquake of the sequence (tree root) is not necessarily the strongest event of the cluster. Let us consider, for instance, the 1988 cluster (Fig. 7): both declustering methods recognised that the 1 February 1988 11:12:41.28 earthquake, with magnitude M3.0, is the triggering earthquake of the sequence; therefore, this event turns out to be an ancestor of the largest event within the cluster, an earthquake with magnitude M4.1 that occurred on 1 February 1988 14:21:38.29.

As for 1988 cluster, there is little difference between NN-cluster and SD-cluster in terms of cluster size, tree graphs, and spatio-temporal distribution of the cluster events

(Tab. 2 and Fig. 7). But this is not always the case. Indeed we noticed that NN-method is prone to cluster some events relatively distant in space and, conversely, SD-method tends to cluster events close in space, but quite far in time, as for the clusters occurred in 1996 and 1998, respectively (e.g. Figs. 8-9). Moreover, SD-method may provide a more complex structure for clusters, reflecting the multilevel triggering property of the ETAS model (Fig. 9).

5.2 Some centrality measures

We have chosen some tools from network theory in order to study the structural properties of clusters through their network representations (tree graphs).

We focus hereafter on the concept of *centrality measure*, which is strictly related to the topology (structural properties) of the network (Freeman, 1978). A centrality value is attributed to each node according to its importance (centrality) within the network. Since "importance" has a relative meaning and appropriate interpretation with respect to circumstances, several centrality measures have been proposed in the literature (Wasserman and Faust (1994), Freeman (1978), Bonacich (1987), Bonacich and Lloyd (2001), Borgatti (2005), and references therein). A brief overview of two centrality measures we considered as relevant for our analysis, is provided hereinafter.

Outdegree centrality. The simplest centrality measures are based on the degree, indegree, and outdegree of a node v, which are respectively defined as the number of edges (links) that are connected to v, the number of incoming edges to v, and the number of outgoing edges from v. We notice that, by construction, each event of a declustered catalog has indegree equal to 0 or 1 (corresponding to background events or secondary events, respectively), and we expect that high outdegrees are especially associated with mainshocks within a cluster. Therefore outdegree turns out to be more suitable than indegree in our application. Let $\delta(v)$ be the outdegree of node v in tree T,

$$\delta(v|T) = \text{number of edges in tree } T \text{ that go down from } v. \tag{5}$$

Since the outdegree of a node is at most $\#T-1$, where $\#T$ is the total number of nodes in T, the outdegree centrality of v is defined as the proportion of direct offsprings from v in the entire tree T:

$$c_\delta(v|T) = \frac{\delta(v|T)}{\#T-1} \quad , \tag{6}$$

so as to obtain a measure independent on network size. Outdegree centrality ranges in $[0, 1]$, where high degree values denote the most important nodes, to which most of the events are connected.

Closeness centrality. The most important node according to closeness centrality has minimum distance from every other nodes. Closeness centrality of a node v is defined as

$$c_c(v|T) = \frac{\#T-1}{\sum_{w\in T} d(v,w)} \quad , \tag{7}$$

where $d(v, w)$ is the shortest distance in T from v to w (i.e., the number of edges in the shortest path from v to w); the numerator $\#T-1$ is the minimum value that the sum in the denominator can take. If there is no path from v to w (e.g. from a node to the root), then $d(v, w)$ is set equal to the total number of nodes in T. Closeness centrality ranges in $[0, 1]$ and, in analogy with outdegree centrality, high degree values denote the most important nodes.

Finally, a global index, named *centralization*, is introduced in order to summarize the centrality measures of all the nodes in the network: Centralization quantifies the differences between the centrality of the most central node v^* and that of all other nodes.

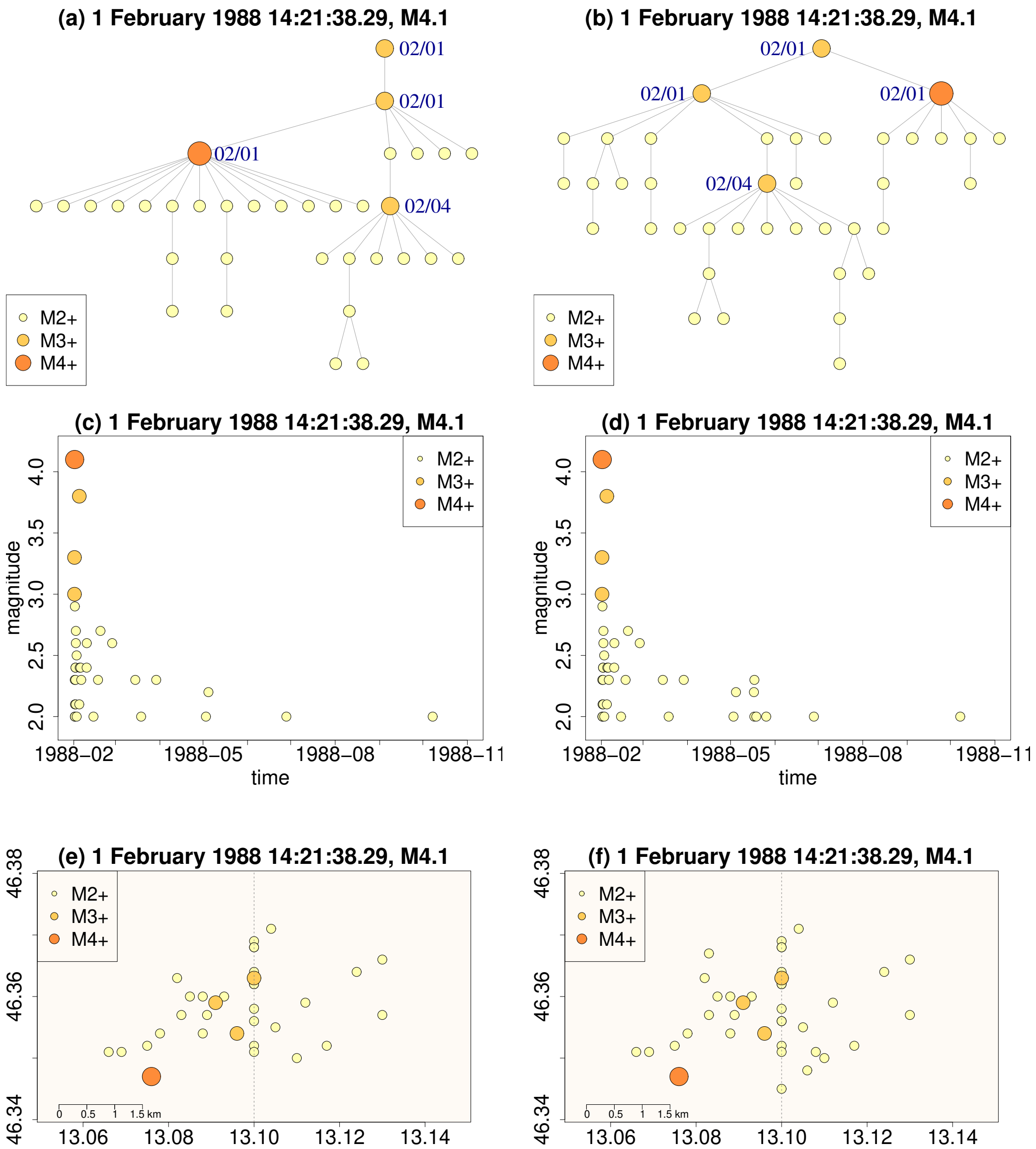

Figure 7. NN-cluster (left) and SD-cluster (right) of the seismic sequence occurred in 1988: (a-b) tree graph representation, (c-d) magnitude versus occurrence times, (e-f) map of the epicentres. Date and magnitude of the largest event is also reported.

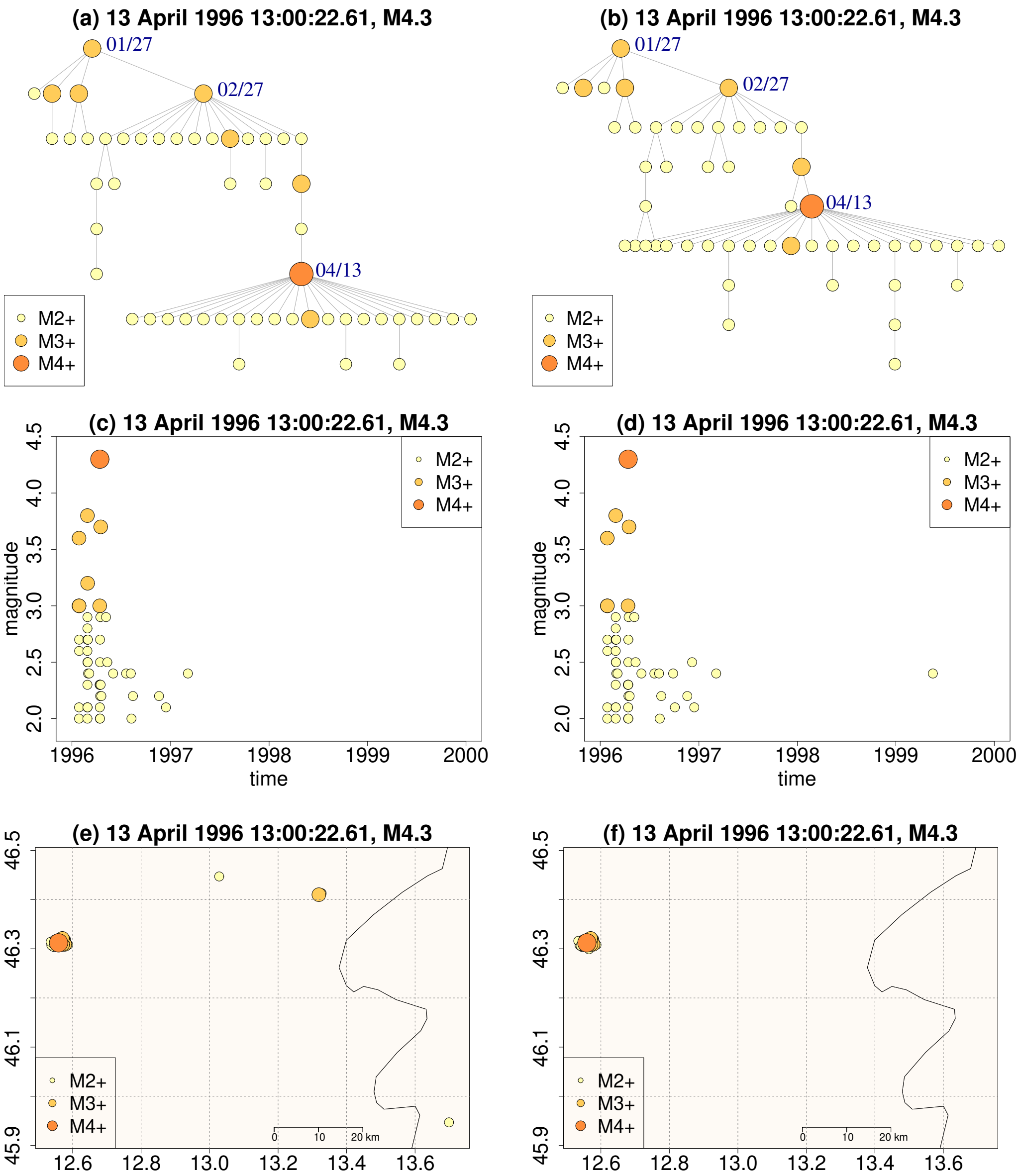

Figure 8. NN-cluster (left) and SD-cluster (right) of the seismic sequence occurred in 1996: (a-b) tree graph representation, (c-d) magnitude versus occurrence times, (e-f) map of the epicentres. Date and magnitude of the largest event is also reported.

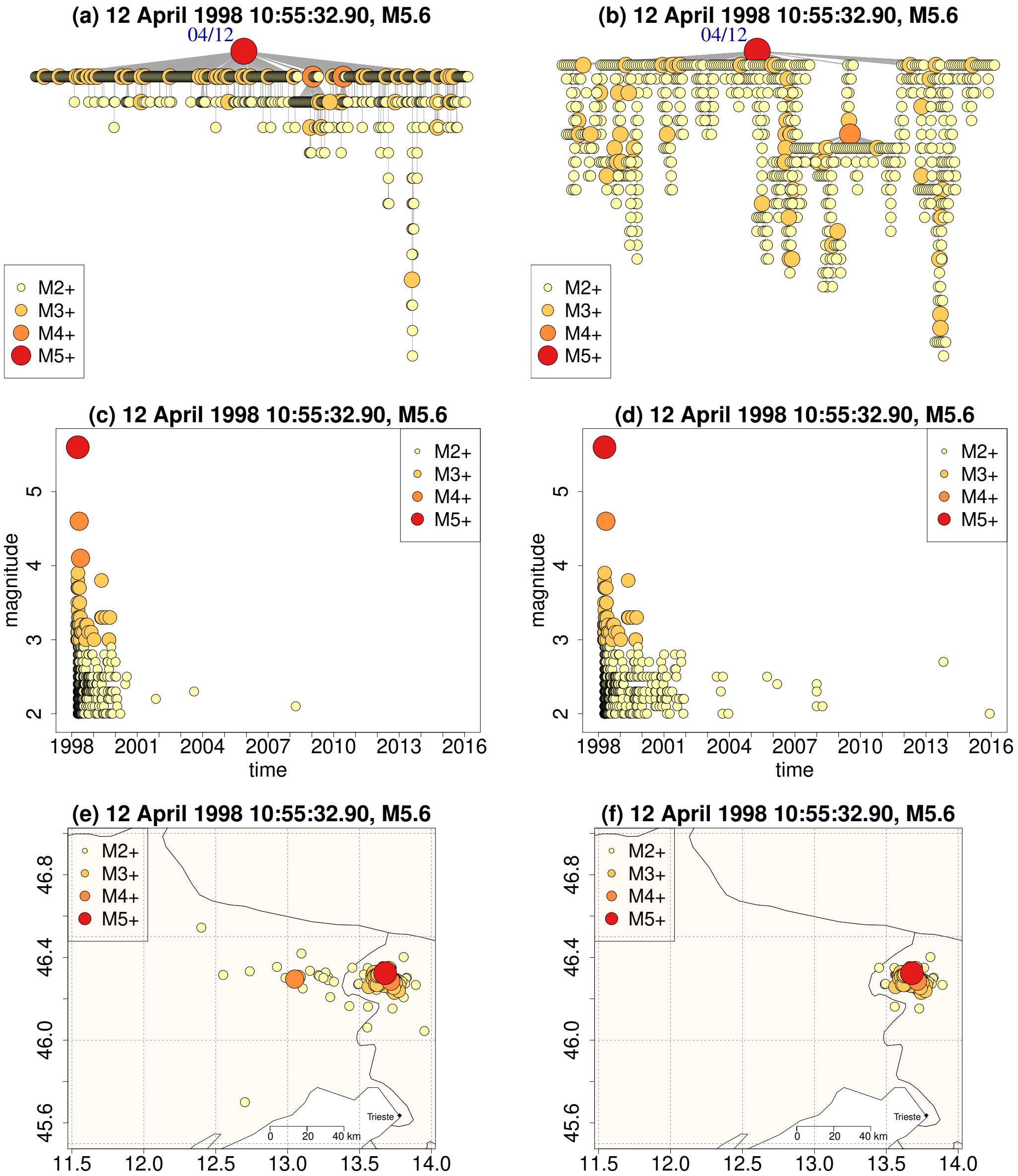

Figure 9. NN-cluster (left) and SD-cluster (right) of the seismic sequence occurred in 1998: (a-b) tree graph representation, (c-d) magnitude versus occurrence times, (e-f) map of the epicentres. Date and magnitude of the largest event is also reported.

The following formulas define the centralization based on outdegree centrality and closeness centrality:

$$C_\delta(T) = \frac{\sum_v c_\delta(v^*|T) - c_\delta(v|T)}{\#T - 1} \qquad \textit{outdegree centralization,} \tag{8}$$

$$C_c(T) = \frac{\sum_v c_c(v^*|T) - c_c(v|T)}{\#T - 1} \qquad \textit{closeness centralization.} \tag{9}$$

Centralization also ranges in $[0, 1]$ and high centralization indicates the tendency of a single node (i.e. an earthquake) to be more central than other nodes in the network (i.e. in the cluster). Both centrality measures and centralization are normalized on $[0, 1]$ and thus independent on the cluster size; this makes the topological comparison among tree graphs easier, compared to the use of other indices (e.g., average node depth and average leaf depth proposed by Zaliapin and Ben-Zion (2013)), especially for clusters with very different numbers of nodes.

Tab. 3 lists the centralization values of matched clusters with large cluster size. Fig. 10 compares all the matched clusters that have at least 5 events, in terms of both C_δ and C_c. Fig. 11 shows the spatial distribution of the epicentres of the representative events for all the matched NN-clusters and the SD-clusters. Overall, it emerges that centralization values of the NN-clusters are comparable to or higher than those of the SD-clusters. Thus, both centralizations C_δ and C_c are proved to be effective indices for expressing what has been observed in Figs. 7-9: whenever a NN-cluster exhibits similar or even simpler structural complexity than its matched SD-cluster, its centralization value is similar to or greater than that of its matched SD-cluster.

We also verified that C_δ and C_c have a strong positive correlation to each other (0.87 for NN-clusters and 0.86 for SD-clusters). Their correlations to the magnitudes of the representative events are moderate (0.60 and 0.46 for NN-clusters, and 0.42 and 0.26 for SD-clusters, respectively) and also their correlations with clusters size are close to zero (between -0.2 and 0.2). This suggests that the complexity of clusters structure does not depend simply on magnitude and related clusters size.

The spatial distribution of centralization values obtained for NN- and SD-clusters (Fig. 11) highlights the basic difference between the two approaches, namely the comparatively higher complexity of SD-clusters structure, which reflects the multilevel triggering property of this approach; in the color scale dark colors correspond to low values of centralization, which are associated with swarm-like sequences, whereas light colors correspond to burst-like sequences. This is particularly evident for the largest earthquakes (events with $M \geq 5$ in Table 2), which are represented by stars in the maps. These events are associated to rather simple clusters by NN (i.e. high centrality values, close to 1), whereas they correspond to complex clusters in SD (i.e. low centrality values, close to 0); this effect is less evident for the 1977 earthquake, possibly because the event occurred at the beginning of the considered data set. In addition, while the spatial distribution of centralization values from NN-clusters does not contradict the spatial pattern identified by Peresan and Gentili (2018), in both maps from SD-clusters, the complex swarm-like sequences appear scattered all over the study area.

6 Conclusions

In this study, we compared the performances of the NN and SD algorithms in classifying events from an earthquake catalogue into clusters and background seismicity. Both methods provide data-driven identifications of earthquake clusters and permit to disclose possible complex features in their internal structure. The two declustering algorithms have been applied to the seismicity data of Northeastern Italy, whose completeness and scaling parameters were already analysed in some detail by Peresan and Gentili (2018).

Table 3. Centralization scores based on outdegree centrality (C_δ) and on closeness centrality (C_c) for the selection of matched large clusters listed in Table 2. The clusters including earthquakes with $M \geq 5$ are marked by numbers as in Fig. 10.

	C_δ		C_c	
largest event	*NN-cluster*	*SD-cluster*	*NN-cluster*	*SD-cluster*
(1) 12 April 1998 M5.6	0.6490	0.1868	0.6307	0.1601
(2) 12 July 2004 M5.1	0.8191	0.4576	0.7832	0.1431
13 April 1996 M4.3	0.3802	0.3602	0.2274	0.2335
(4) 16 September 1977 M5.2	0.8462	0.5836	0.8472	0.6763
1 February 1988 M4.1	0.3756	0.1627	0.2632	0.2754
18 April 1979 M4.8	0.4623	0.5041	0.4798	0.5013
20 April 1994 M3.7	0.5275	0.4808	0.1898	0.2162
14 February 2002 M4.9	0.8827	0.9172	0.4466	0.4476
5 October 1991 M3.8	0.5640	0.2377	0.3855	0.2052
12 February 2013 M3.8	0.1200	0.1074	0.1798	0.1063
25 February 2018 M3.9	0.4643	0.3878	0.3737	0.3351
29 August 2015 M4.3	1.0000	0.6686	1.0000	0.6985

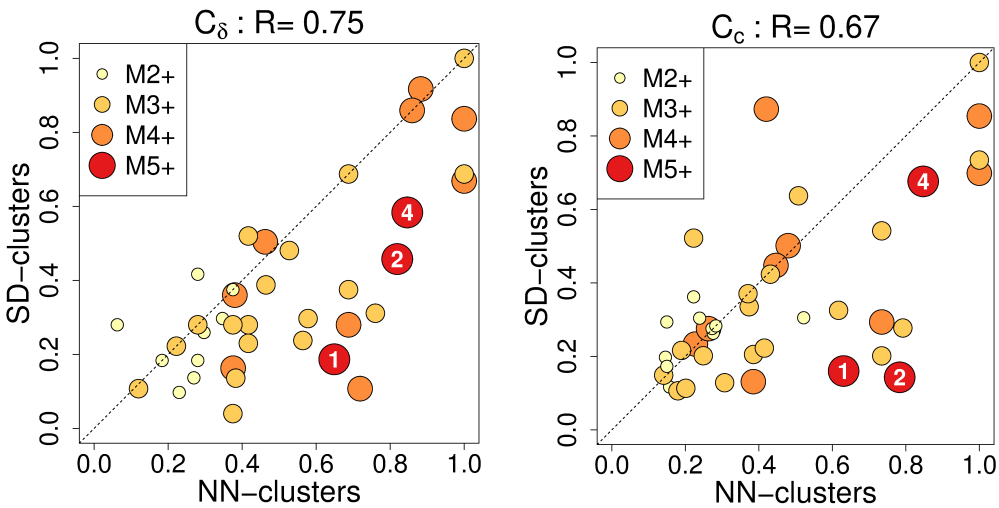

Figure 10. Comparison of the matched clusters that have at least 5 events, in terms of outdegree centralization (left) and closeness centralization (right); correlation values are also reported. The colors and sizes of the dots refer to the magnitude level of the largest event in the clusters. Numbered symbols refer to the events listed in Tab. 3.

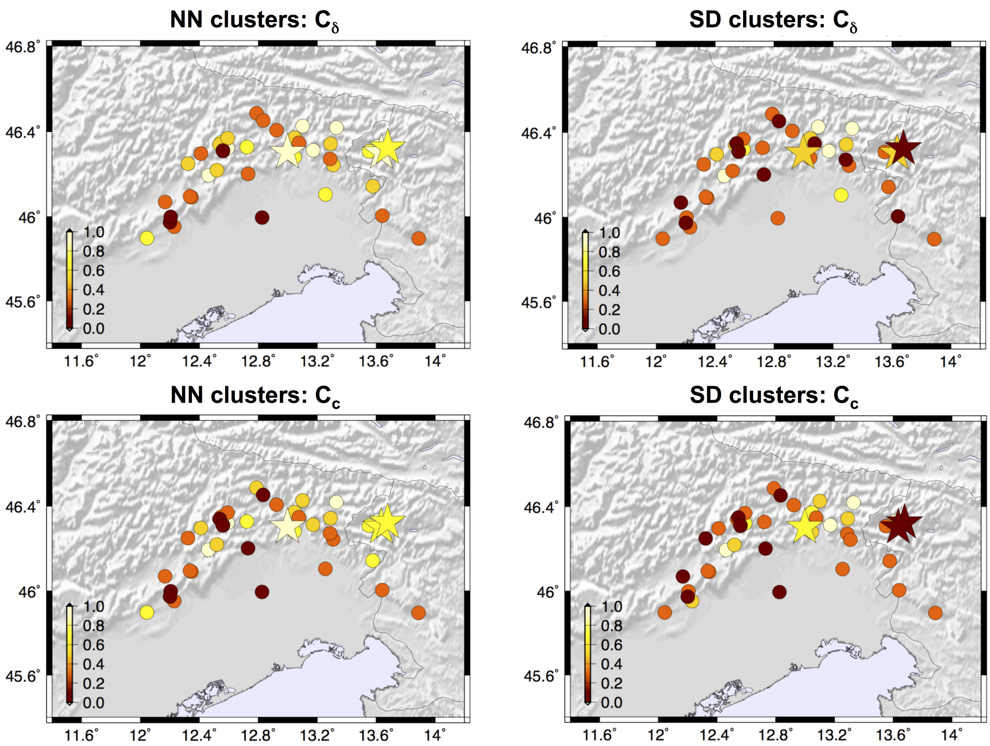

Figure 11. Spatial distribution of the epicentres of the representative events (the largest events within each of the NN-clusters (left) and the SD-clusters (right)) for the clusters that have at least 5 events. Each epicentre is associated with the outdegree centralization (top) or the closeness centralization (bottom) of its cluster. The matched clusters are denoted by circles and the events with $M \geq 5$ are highlighted by stars.

The global features of the resulting background seismicity and earthquake clusters turn out well consistent, though the partitions are slightly different. Specifically, the statistics of clusters, singles and fore/aftershocks are quite comparable (Tab. 1). Both NN and SD results consistently show that background seismicity is composed by a large proportion of single events (about 45-50%) and by a limited number of clustered events (8-9%). However the events forming the background may be different (especially in presence of foreshocks), due to the different definitions used by the two methods: NN assigns to background the largest earthquake from each cluster, whereas SD the first independent earthquake in the cluster.

Since the two methods also allow to outline the internal structure of clusters, an in-depth comparison was carried out both for selected clusters (Figs. 7, 8, 9) and for all matching clusters identified by NN and SD (Figs. 10, 11). The concepts of outdegree centrality and closeness centrality have been introduced from network theory to quantitatively compare the characteristics of the declustering outputs, by regarding earthquake clusters as tree graphs. The proposed centrality measures, C_δ and C_c, are especially advantageous when clusters with different and large sizes are compared; in these cases, the tree graph representation of the cluster might be very unclear due to the large number of nodes, while centralization indices are still able to capture some key properties of the hierarchical complexity of the cluster and to rank earthquakes within the cluster according to their importance/centrality. These quantitative measures are shown to be able to characterize the internal structure of the clusters in a robust and consistent way. Accordingly, we found that NN-clusters usually display simpler internal structures than SD-clusters and that the corresponding centralization values of NN-clusters are higher than those of SD-clusters.

Given the outcomes of this in-depth comparative analysis of NN and SD methods, there are still some open issues that need to be addressed and will be matter for future research. The main outcome of this study consists in the identification of the basic similarities and differences between the NN and SD methods, both in their theoretical formulation and operational results. From a methodological point of view, we believe the use of centrality measures and other tools borrowed from network theory may open new possibilities in the study of earthquake sequences and their evolution. Another issue is to verify generality of above conclusions, that is to assess to what extent they depend on the considered catalog and study area by performing the same analysis in different regions. Finally, there is the problem of investigating how these declustering algorithms influence the forecasting performance in short-term and long-term earthquake hazard assessment.

Acknowledgments

This work is supported by National grant MIUR, PRIN-2015 program, Prot. 20157PRZC4: Complex space-time modeling and functional analysis for probabilistic forecast of seismic events. J. Zhuang is also partially supported by the JSPS Grant-in-aid 19H04073. We also acknowledge financial support from Civil Defence of the Friuli Venezia Giulia Region. Graph visualization is realized by the R package *igraph* (Csardi & Nepusz, 2006). Earthquake data used in this study are published as yearly and monthly bulletins by the National Institute of Oceanography and Experimental Geophysics (OGS) and are publicly available via the OGS website (http://www.crs.inogs.it/bollettino/RSFVG/).

References

Baiesi, M., & Paczuski, M. (2004). Scale-free networks of earthquakes and aftershocks. *Physical Review E*, *69*, 066106. doi: 10.1103/PhysRevE.69.066106

Bonacich, P. (1987). Power and centrality: A family of measures. *American Journal of Sociology*, *92*(5), 1170-1182. doi: 10.1086/228631

Bonacich, P., & Lloyd, P. (2001). Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, *23*(3), 191 - 201. doi: 10.1016/S0378-8733(01)00038-7

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, *27*(1), 55 - 71. doi: 10.1016/j.socnet.2004.11.008

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, *Complex Systems*, 1695. (http://igraph.org)

Davoudi, N., Tavakoli, H. R., Zare, M., & Jalilian, A. (2018). Declustering of Iran earthquake catalog (1983-2017) using the epidemic-type aftershock sequence (ETAS) model. *Acta Geophysica*, *66*(6), 1359-1373. doi: 10.1007/s11600-018-0211-5

Freeman, L. C. (1978). Centrality in social networks. conceptual clarification. *Social Networks*, *1*(3), 215–239. doi: 10.1016/0378-8733(78)90021-7

Gardner, J. K., & Knopoff, L. (1974). Is the sequence of earthquakes in southern California, with aftershocks removed, Poissonian? *Bulletin of the Seismological Society of America*, *64*(5), 1363-1367.

Guo, Y., Zhuang, J., Hirata, N., & Zhou, S. (2017). Heterogeneity of direct aftershock productivity of the main shock rupture. *Journal of Geophysical Research: Solid Earth*, *122*(7), 5288–5305. doi: 10.1002/2017JB014064

Guo, Y., Zhuang, J., & Zhou, S. (2015). An improved space-time ETAS model for inverting the rupture geometry from seismicity triggering. *Journal of Geophysical Research: Solid Earth*, *120*(5), 3309–3323. doi: 10.1002/2015JB011979

Nandan, S., Ouillon, G., Sornette, D., & Wiemer, S. (2019). Forecasting the rates of future aftershocks of all generations is essential to develop better earthquake forecast models. *Journal of Geophysical Research: Solid Earth*, *124*(8), 8404-8425. doi: 10.1029/2018JB016668

Nekrasova, A., Kossobokov, V., Peresan, A., Aoudia, A., & Panza, G. F. (2011). A multiscale application of the unified scaling law for earthquakes in the central mediterranean area and alpine region. *Pure and Applied Geophysics*, *168*(1), 297–327. doi: 10.1007/s00024-010-0163-4

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, *50*(2), 379–402. doi: 10.1023/A:1003403601725

Peresan, A., & Gentili, S. (2018). Seismic clusters analysis in Northeastern Italy by the nearest-neighbor approach. *Physics of the Earth and Planetary Interiors*, *274*, 87 - 104. doi: 10.1016/j.pepi.2017.11.007

Reasenberg, P. (1985). Second-moment of central California seismicity, 1969–1982. *Journal of Geophysical Research*, *90*(B7), 5479–5495.

Rotondi, R. (2010). Bayesian nonparametric inference for earthquake recurrence time distributions in different tectonic regimes. *Journal of Geophysical Research: Solid Earth*, *115*(B1), 1–25. doi: 10.1029/2008JB006272

Rotondi, R., & Varini, E. (2019). Failure models driven by a self-correcting point process in earthquake occurrence modeling. *Stochastic Environmental Research and Risk Assessment*, *33*(3), 709–724. doi: 10.1007/s00477-019-01663-5

Rovida, A., Locati, M., Camassi, R., Lolli, B., & Gasperini, P. (2014). *Italian Parametric Earthquake Catalogue (CPTI15), version 2.0. Istituto Nazionale di Geofisica e Vulcanologia (INGV).* (https://doi.org/10.13127/CPTI/CPTI15.2)

Slejko, D., Neri, G., Orozova, I., Renner, G., & Wyss, M. (1999). Stress field in friuli (ne italy) from fault plane solutions of activity following the 1976 main shock. *Bulletin Seismological Society of America*, *89*(4), 1037–1052.

Talbi, A., Nanjo, K., Satake, K., Zhuang, J., & Hamdache, M. (2013). Comparison of seismicity declustering methods using a probabilistic measure of clustering. *Journal of Seismology*, *17*(3), 1041–1061. doi: 10.1007/s10950-013-9371-6

van Stiphout, T., Zhuang, J., & Marsan, D. (2012). Seismicity declustering. *Community Online Resource for Statistical Seismicity Analysis*, 1–25. (Available at http://www.corssa.org) doi: 10.5078/corssa-52382934

Vere-Jones, D. (1978). Earthquake prediction - A statistician's view. *Journal of Physics of the Earth*, *26*(2), 129-146. doi: 10.4294/jpe1952.26.129

Wang, Q., Jackson, D. D., & Zhuang, J. (2010). Are spontaneous earthquakes stationary in California? *Journal of Geophysical Research: Solid Earth*, *115*(B8), 1–12. doi: 10.1029/2009JB007031

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications.* Cambridge University Press. doi: 10.1017/CBO9780511815478

Zaliapin, I., & Ben-Zion, Y. (2013). Earthquake clusters in southern California II: Classification and relation to physical properties of the crust. *Journal of Geophysical Research: Solid Earth*, *118*(6), 2865-2877. doi: 10.1002/jgrb.50178

Zaliapin, I., & Ben-Zion, Y. (2016). A global classification and characterization of earthquake clusters. *Geophysical Journal International*, *207*(1), 608-634. doi: 10.1093/gji/ggw300

Zaliapin, I., Gabrielov, A., Keilis-Borok, V., & Wong, H. (2008, Jun). Clustering analysis of seismicity and aftershock identification. *Phys. Rev. Lett.*, *101*, 018501. doi: 10.1103/PhysRevLett.101.018501

Zhang, Q., & Shearer, P. M. (2016). A new method to identify earthquake swarms applied to seismicity near the San Jacinto Fault, California. *Geophysical Journal International*, *205*(2), 995-1005. doi: 10.1093/gji/ggw073

Zhuang, J. (2006). Second-order residual analysis of spatiotemporal point processes and applications in model evaluation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(4), 635–653. doi: 10.1111/j.1467-9868.2006.00559.x

Zhuang, J., Chang, C.-P., Ogata, Y., & Chen, Y.-I. (2005). A study on the background and clustering seismicity in the Taiwan region by using point process models. *Journal of Geophysical Research: Solid Earth*, *110*(B5), 1–12. doi: 10.1029/2004JB003157

Zhuang, J., Ogata, Y., & Vere-Jones, D. (2002). Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, *97*(3), 369–380.

Zhuang, J., Ogata, Y., & Vere-Jones, D. (2004). Analyzing earthquake clustering features by using stochastic reconstruction. *Journal of Geophysical Research: Solid Earth*, *109*(3), B05301. doi: 10.1029/2003JB002879