

Visualization of Earthquake Clusters over Multi-dimensional Space

David A. Yuen^a, Witold Dzwiniel^b, Yehuda Ben-Zion^c, Ben Kadlec^d

^aDept. of Geology and Geophysics, University of Minnesota, Minneapolis, MN 55415-1227, USA

^bDept. of Computer Science, AGH University of Sci. and Technol., 30-059, Kraków, Poland

^cDepartment of Earth Sciences, University of Southern California, Los Angeles, CA 90089, USA

^dDepartment of Computer Science, University of Colorado at Boulder, USA

Article outline

Glossary and Notation

I Definition of the subject and its importance.

II Introduction.

III Tools for Detection and Visualization of Clusters

- 1 Basic concepts of clustering.
- 2 Clustering Techniques.
- 3 Multidimensional Scaling.

IV Data description.

- 1 Earthquake datasets.
- 2 Physical model of earthquake dynamics.
- 3 Synthetic catalogs.

V Earthquake Clustering in Space and Time

VI Multidimensional feature space.

VII Visual Clustering of Earthquakes in feature space.

- 1 Short-time period.
- 2 Time period of 150 years.
- 3 Time period of 1500 years.

VIII Remote Problem Solving Environment (PSE) for Analyzing Earthquake Clusters.

- 1 Remote visualization and analysis of large data sets.
- 2 Grid Environment.
- 3 Example of Remote PSE.

IX Future Directions

X References

XI Supplementary material

Glossary and Notation

Grid: Virtual metacomputer, which uses a network of geographically distributed local networks, computers and services. **Grid computing** focuses on distributed computing technologies, which are not in the traditional sense of dedicated clusters. **Data grids** – represent controlled sharing and management of large amounts of distributed data.

Problem Solving Environment (PSE): A specialized computer software for solving one class of problems. They use the language of the respective field and often employ modern graphical user interfaces. The goal is to make the software easy to use for specialists in fields other than computer science. PSEs are available for generic problems like data visualization of large systems of partial differential equations and for specialized fields in science or engineering.

Global Seismographic Network (GSN): The goal of the GSN is to deploy permanent seismic recording stations uniformly over the earth's surface. The GSN stations continuously record seismic data from very broad band seismometers at 20 samples per second, and to provide for high-frequency (40 sps) and strong-motion (1 and 100 sps) sensors where scientifically warranted. It is also the goal of the GSN to provide for real-time access to its data via Internet or satellite. Over 75% of the 128 GSN stations meet this goal as of 2003.

WEB-IS: A software tool developed at the University of Minnesota in 2003 that allows remote, interactive visualization and analysis of large-scale 3-D earthquake clusters over the Internet through the interaction via a workflow procedure between client and server.

Scientific visualization is a branch of computer graphics and user interface design that are dealing with presenting clearly scientific data to users, by means of patterns and images. The goal of scientific visualization is to improve understanding of the large amounts of data being presented.

Interactive visualization is a branch of graphic visualization that studies how users interact in real time with computers to create rapidly graphic illustrations of information such as volume-rendering of time-dependent scenarios and how this process can be made more efficient.

Remote-visualization: The tools for interactive visualization of high-resolution images on remote client machine, rendered and preprocessed on the server. This allows for collaboration over long distances.

OpenGL: A standard specification defining a cross-language cross-platform application programming interface (API) for writing applications that produce 2D and 3D computer graphics.

Sumatra-Andaman Earthquake: An extremely large undersea earthquake at a subduction zone with a magnitude of 9.3 that occurred at 00:58:53 UTC (07:58:53 local time) December 26, 2004, with an epicentre off the west coast of Sumatra, Indonesia. The earthquake triggered a series of devastating tsunamis along the coasts of all of the countries bordering the Indian Ocean, killing around 300,000 people and inundating coastal communities across South and Southeast Asia, including parts of Indonesia, Sri Lanka, India, Pakistan and Thailand.

Earthquake catalog: Data set consisting of earthquake hypocenters, origin times, and magnitudes. Additional information may include phase and amplitude readings of seismograms, as well as first-motion mechanisms and moment tensors.

Pattern recognition: The methods, algorithms and tools to analyze data based on either statistical information or on a priori knowledge extracted from the patterns. The patterns for classification are groups of observations, measurements, objects, defining feature vectors in an appropriate multidimensional feature space.

Data mining: The algorithms, tools, methods and systems used in extraction of knowledge hidden in a large amount of data.

Features – denoted f_i or F_j (i, j – feature indices) - a set of variables which carry discriminating and characteristic information about the objects under consideration. The features can represent raw measurements (data) f_i or can be generated in a non-linear way from the data F_j (features).

Feature space: The multidimensional space in which the \mathbf{F}_k vectors are defined. Data and feature vectors represent vectors in their respective spaces.

Feature vector: A collection of features ordered in some meaningful way into multi-dimensional feature vectors \mathbf{F}_l (\mathbf{F}_l where l – feature vector index) that represents the signature of the object to be identified.

Feature extraction: A procedure of mapping source feature space into output feature space of lower dimensionality, by retaining the minimal value of the error cost function.

Multidimensional scaling: A nonlinear method of feature extraction, which minimizes the value of the “stress” being the function of differences of all the distances between feature vectors in the source space and corresponding distances in the resultant space of lower dimensionality.

Data space: The multi-dimensional space in which the data vectors \mathbf{f}_k exist.

Data vector: A collection of features ordered in some meaningful way into multi-dimensional vectors \mathbf{f}_k (\mathbf{f}_k , k – data vector index) and $\mathbf{f}_k = [m_k, z_k, \mathbf{x}_k, t_k]$ where m_k is the magnitude and \mathbf{x}_k, z_k, t_k – its epicentral coordinates, depth and the time of occurrence, respectively.

Cluster: Isolated set of feature (or data) vectors in data and feature spaces.

Clustering: A computational method for extracting clusters in multidimensional feature space.

Agglomerative (hierarchical) clustering algorithm: A clustering algorithm in which at the start the feature vectors represent separate clusters and the larger clusters are built-up in a hierarchical way. The procedure repeats the process of gluing-up the closest clusters up to the stage, when a desired number of clusters is achieved.

k-means clustering: Non-hierarchical clustering algorithm in which the randomly generated centers of clusters are improved iteratively.

Multi-resolutional clustering analysis: A hierarchy of clusters can be obtained due to clustering. The analysis of the results of clustering in various resolution levels allows for the extraction of knowledge hidden in both local (small clusters) and global (large clusters) similarity associated with the multi-dimensional feature vectors.

N-body solver: The numerical algorithm for exploiting the concept of temporal evolution of an ensemble of N mutually interacting particles.

Non-hierarchical clustering algorithm: A clustering algorithm in which the clusters are sought by using global optimization algorithms. The most representative algorithm of this type is coined the **k-means** procedure.

I Definition of the subject and its importance

Earthquakes have a direct societal relevance because of their tremendous impact on human community [90]. The genesis of earthquakes is an unsolved problem in the earth sciences, because of the still unknown underlying physical mechanisms. Unlike the weather, which can be predicted for several days in advance by numerically integrating several non-linear coupled partial differential equations on massively parallel systems, earthquake forecasting remains an elusive goal, because of the lack of direct observations and the fact that the governing equations and physics are still unknown. Instead one must employ statistical approaches (e.g. [58,71,81]) and data-assimilation techniques (e.g., [1,92,107]).

The nature of the spatio-temporal evolution of earthquakes has to be assessed from the observed seismicity and geodetic measurements. Problems of this nature can be analyzed by recognizing non-linear patterns - clusters – which are hidden in the vast amount of seemingly unrelated information. The clustering of seismic events may arise from multi-scale feed-back mechanisms between the nonlinear rock rheology and earthquake failures [13,14,65,67]. With the proliferation of large-scale computations, data mining [112] - which is a time-honored and well-understood process - has come into its own for extracting meaningful patterns from large incoherent data sets found in diverse fields, such as astronomy, medical imaging, combinatorial chemistry, bio-informatics, seismology, remote sensing and stock markets [75]. Examples of clusters abound in nature and include stars in galaxies, hubs in airline routes and centers of various human relationships [6].

Indeed, clustering analysis - i.e., the procedure of extracting clusters - has become a widely successful approach for revealing features and patterns in the data-mining business. The concept of clustering has been used for many years in pattern recognition [91, 103, 105,111]. Clustering can use more (e.g. [48]) or less mathematically rigorous principles (e.g. [34]). Nowadays clustering is recognized as an important tool for revealing coherent features in the earth sciences [33,64,65,109], bioinformatics [28,106] and in data mining [98,101,102,108].

We describe the method of using clustering as a tool for analyzing seismic data sets together with visualization techniques necessary for interpreting the results. Combination of clustering with feature extraction methods, which enables 3-D visualization of multidimensional spaces, gives us a new emergent dimension in looking at seismic data. It allows for direct, visual exploration of multivariate and mutiresolutional feature space by the specialist in the field and for deducing an optimal strategy for knowledge extraction. Armed with this visual clustering concept, one can extract multiresolutional structure of seismic clusters both in spatio-temporal space and feature space.

Recent advances in information technology (IT), high performance computing, and satellite imagery have given rise to extremely large data sets, exceeding terabytes at each turn, that are coming regularly to physical scientists who need to analyze them quickly. These data sets are non-trivial to analyze without the use of algorithms that find solutions with a minimal computing complexity. The arrival of petascale computing will also spur the development of new visualization techniques, which are sorely needed to understand the vast amounts of data compressed in many different kinds of spaces, with spatial, temporal and other types of dimensions [111].

The Grid infrastructure, enabling instantaneous access to huge widely distributed computing and storage resources and other types of IT services, will serve as ideal platform for online visualization and for remote visual exploration of vast amount of seismic data from the data bases scattered over the world.

II Introduction

There are a few statistical laws which represent the basis for earthquake models development. The frequency-size statistics of regular tectonic earthquakes (excluding swarms and deep focus earthquakes) are found to follow the Gutenberg-Richter relationship [38,78,83]:

$$\log N(M) = a - bM \quad (6)$$

where N is the number of events with magnitude larger than M and a, b are constants giving, respectively, the overall seismicity rate and relative rates of events in different magnitude ranges. Observed b -values of regional seismicity typically fall in the range 0.7-1.3.

Aftershock decay rates can usually be described by the Omori-Utsu law [70,113]:

$$\Delta N / \Delta t = K(t + c)^{-p} \quad (7)$$

where N is the cumulative number of events, t is the time after the mainshock, and $K, c,$ and p are empirical constants. The epidemic-type aftershock-sequences (ETAS) model combines the Omori-Utsu law with the Gutenberg-Richter frequency-magnitude relationship for a history dependent occurrence rate associated with a point process in the form (e.g., [58])

$$\lambda(t | H_t) = \mu + \sum_{t_i < t} \frac{K_0 \exp[\alpha(M_i - M_c)]}{(t - t_i + c)^p} \quad (8)$$

where α is a constant background rate, M_i is the magnitude of earthquake at time t_i , M_c is a lower magnitude cut-off, H_t denotes the history, and the productivity factor $K_0 \exp[\alpha(M_i - M_c)]$ gives the number of events triggered by a parent earthquake with magnitude M_i . The ETAS model is used widely in analysis of seismic data, owing to its built-in clustering of events in time associated with the incorporation of the Gutenberg-Richter and Omori-Utsu laws. Examples of their recent applications can be found in [41,59,68].

These results can be used to derive additional properties, such as the average recurrence times (e.g., [5,20,21,22,68,88]). It is usually defined as the number of years between occurrences of an earthquake of a given magnitude in a particular area. For example, the probability of a devastating earthquake striking the greater San Francisco Bay Region over the following 25 years (2007-2031) is estimated to be 0.62 [68]. Corral [20,21,22] proposed the existence of a universal scaling law for the probability density function $H(\tau)$ of recurrence times (or interevent times) τ between earthquakes in a given region:

$$H(\tau) \cong \lambda \times f(\lambda\tau). \quad (9)$$

The function $f(x)$ appears to be similar for many different seismic regions, which suggests some universal properties. The average rate λ represent the region specific constant, whose reciprocal is the only relevant characteristic timescale for the recurrence times. Molchan [54] showed that under general conditions, the only universal distribution of inter-event times in a stationary point process is exponential. Hainzl et al. [39] and Saichev and Sornette [68] discussed the relationships between statistics of interevent times, the ETAS model of triggered seismicity, and the Corral [20,21] distribution of equation 9.

These empirical findings with a broad magnitude range suggest that the largest seismic events are surrounded by a bunch of smaller events, according to certain statistical time arrangements. Earthquakes usually occur in sequences consisting of relatively small number of

foreshocks, a mainshock, and large number of aftershocks with rate following Eq.(7). The mainshock is assumed to have about a magnitude unit larger than the largest aftershock. However, this is not always satisfied. Occasionally, for so called earthquake doublets, a large earthquake is followed by an event of similar size. It may occur either on an adjacent portion of the fault or on a neighboring faults. Earthquake doublets present particularly difficult challenges for seismic hazard assessment after a large event, but also provide insights into earthquake clustering, triggering and stress cycling [3].

In the context of earthquake prediction it is important to analyze earthquake cycles with repeating sequences of events, such as foreshocks, mainshocks and aftershocks (e.g., [12,78,110]). Apart from tendencies reflected by statistical laws, earthquakes exhibit various types of more subtle spatio-temporal clustering, i.e., grouping of events of the same type both in time and in spatial coordinates. Recognition of these patterns followed by the analysis of the reasons for their appearance may lead to the development of improved prediction algorithms.

Seismic events produce non-random clusters [14,65,68]. One of the goals of earthquake clustering studies is to find spatial and/or temporal patterns of seismic activities associated with earthquake dynamics [26,27]. These patterns can occur both in geographical (spatial clustering of faults) and configuration-feature spaces (temporal clustering of activity). Spatial clustering of faults is a consequence of fractal structure of faults network and interactions among them [65]. In [62] a new pattern recognition method was proposed for imaging the 3D structure of the active part of a fault network, using the spatial location of earthquakes.

Information in seismicity data sets (either synthesized or real) can be both relevant and irrelevant from the point of view of deterministic earthquake dynamics. It can be also "entangled" and impossible to be interpreted with normal human perception. Meaningful patterns are usually surrounded by artifacts, noise and outliers. The role of data mining is to produce a mathematically rigorous algorithm for extracting relevant information from this deluge of data, and make it understandable. Clustering techniques allows us to produce specially crafted data models that can be employed for predicting the nature of future events. In more complex cases, these special data models can work in concert with formal mathematical and physical paradigms to give us a deeper physical insight.

Depending on the data structures and goals of classification, different clustering schemes must be applied [36,48,49]. Because clustering never gives consistent and firm outcome, 3-D visualization of mostly multidimensional clusters gives direct insight into the structure of the feature space to estimate the plausibility of the clustering scheme selected. Visualization of high-dimensional space allows also for finding simple dependences between clusters (e.g., their mutual position) and eliminates noise, outliers and artifacts.

In this chapter we emphasize the strategy for both visualizing and interpreting the computed results from clustering rather than just the role of clustering in the understanding of earthquake dynamics. All the seismic events occurring over a certain region during a given time period are viewed as a single cluster of correlated events. The strength of mutual correlations between events, such as correlations in spatial and time positions along with magnitude, cause this single cluster to have very complex internal structure. The correlations - the measures of similarity between events - divide the global cluster into a variety of small clusters of multi-scale nature, i.e., small clusters may consist of a cascade of smaller ones. Coming down to smaller scales, we record clusters with more and more tightly correlated events. Exploring the nature of events belonging to a single cluster, we can extract the underlying common features. Having more information about events belonging to the same cluster, we can uncover hidden relationships between them. Moreover, we can anticipate the type of an unknown event belonging to a certain cluster from the character of the other events belonging to this cluster.

In the following sections we describe the idea of visual clustering and the idea of higher dimensions associated with data sets. We show that the visualization of clusters is an essential tool for understanding data arrangement. It is well known that the quality of the best clustering

algorithms is very data dependent. Moreover, due to the difficulty of properly understanding the character and causes of earthquake clustering one needs to explore many hypotheses. For this purpose, instead of usually simply clustering techniques, we propose tying together visualization and clustering (visual clustering). Visual clustering enables an experienced specialist with specific knowledge to confront confidently the data structure.

In sum, visual clustering concentrates on the tight cooperation between visualization algorithms (e.g. MDS – multidimensional scaling) and clustering schemes. In our investigations we used many clustering techniques (some of which are mentioned in Section III) and several stress functions describing the character and goals of MDS transformation. The best combination can be selected on the basis of multiple visual inspections of the multidimensional structure associated with the dataset.

Application of higher-dimensional data for analyzing earthquakes was introduced by Keilis-Borok and Kossobokov in [50,51]. In their M8 algorithm the transient seismic activity in the region is described by a vector of seven parameters, which are used for selecting times of increased probability (TIPs) of strong earthquakes. We use another set of parameters devised by Eneva and Ben-Zion [30,31]. In the M8 algorithm six parameters are based on seismicity rates and one parameter combines a function of the event sizes with a function of the seismicity rate. These parameters are highly correlated by default and they provide minimal information on the seismic response in space. The seven seismicity parameters used by Eneva and Ben-Zion [30,31] attempted to provide a more comprehensive sampling of the seismic response in the space, time and size domains. Eneva and Ben-Zion [30,31] also proposed improved techniques for extracting information from the parameter time-series, which included not only high values but also on low values, the distinction between local extrema and trends, the difference between precursory- and post-seismic effects, and better normalization procedures.

We demonstrate the results of our clustering analysis using both synthetic and real data. Long synthetic data were derived by using a model for a segmented strike-slip fault zone in a 3D elastic half-space [10]. The real data represent short time (5 years interval) seismic activities of the Changbaishan volcano (the northeast frontier of the North China craton) and the Japanese Archipelago. Lastly, we also emphasized the potential role played by remote visualization environment specially devised for visualizing earthquake clusters.

III Tools for Detection and Visualization of Clusters

The basic goal of knowledge extraction is to discover possible dependences between data and/or between model parameters. They can be unveiled as patterns (clusters) in time, spatial and feature (parameter) space domains. In order to explore the structures involving mostly high-dimensional spaces, we should have a reliable procedure to pin-point the clusters. However, reliable cluster detection remains an elusive goal. A universal clustering scheme that is able to answer unambiguously two basic questions, i.e.,

1. Does the data structure have clusters?
2. How many clusters it contains?

simply does not exist. Moreover, the clusters can have finer structure (multi-resolution) or different densities. The existence of outliers and noisy background additionally exacerbates the search of a viable solution for this problem. There are many clustering methods which are specialized to find clusters of a specific shape (e.g. compact, the same density) or in specific conditions (lack of noise and outliers, and low dimensionality). As we will never know before hand what the structure of the feature space looks like, we have to check many clustering

methods, never being sure whether what we have obtained are really the clusters of interest or just accidental clumps. Fortunately, there exist feature extraction methods that can help in exploring visually multidimensional space. They allow for decreasing the high space dimensionality to 3-D by minimizing a cost function. Even though the resulting structures may be distorted, we can minimize this effect by controlling the various parameters associated with the feature extraction algorithms. Visual clustering consists in employing both clustering schemes and feature extraction algorithms in the process of knowledge extraction. The results of clustering in high dimensional spaces can be readily estimated by visualizing their 3-D representations instead of relying on simple graphs. The noise and outliers can be easily removed. Afterwards, the filtered data can be clustered again or another clustering scheme can be inspected to select the most plausible hypothesis. Single clusters can be investigated separately by searching for clusters with a finer resolution. Apart from fundamental results of clustering, we are able to estimate visually the shapes of clusters and their mutual positions. This information can be also valuable in constructing hypotheses. In this section we present a closer look at the visual clustering concept as a viable tool for knowledge extraction.

III.1 Basic concepts of clustering

Clustering analysis is a mathematical concept whose main useful role is to extract the most similar (or dissimilar) separated sets of objects according to a given similarity (or dissimilarity) measure [91, 103]. Clustering is one of the most fundamental processes generated by nature. For example, people gathering in groups, tribes, demonstrations, parties, cities, produce clusters. Similarly, towns and cities are clusters of buildings while galaxies are clusters of stars. Local computer networks and bacterial colonies are also clusters. The objects forming clusters can be the clusters of smaller objects, which in turn, are clusters of even smaller and smaller building bricks. The complexity of cluster structure reflects the complexity of the real world. The clusters of various shapes, densities and sizes, with additional attributes as colors, transparency etc. built up patterns, which are the underlying scaffold of all multi-scale processes and phenomena. Clusters are the primitives of the patterns.

Our main challenge is to devise a clustering scheme which can divide the M feature vectors \mathbf{x}_i , $i=1,2,\dots,M$ into k separate groups (clusters). More formally, by assuming that $X = \{\mathbf{x}_i\}_{i=1,\dots,M}$ and $\mathbf{x} \in \mathbf{R}^N$, $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{iN}]$ we define as an k -clustering of X the partition of X into k clusters C_1, \dots, C_k provided three conditions are satisfied:

- $C_i \neq \emptyset$, $i=1, \dots, k$ – the clusters are non empty sets,
- $\cup_{i=1, \dots, k} C_i = X$ – the sum of elements inside clusters is equal to the total number of feature vectors,
- $C_i \cap C_j = \emptyset$, $i \neq j$, $j=1, \dots, k$ – each feature vector belongs to only one cluster.

The computational problem with clustering is that the number of possible clustering of M vectors into k groups is given by the Stirling numbers (very large numbers) of the second kind:

$$S(M, k) = \frac{1}{k!} \sum_{i=1}^k (-1)^{k-i} \binom{k}{i} \cdot i^M \quad (10)$$

Some values of $S(N, k)$ are: $S(15, 3) \approx 2 \times 10^6$, $S(20, 4) \approx 45 \times 10^9$, $S(25, 8) \approx 7 \times 10^{17}$, $S(100, 5) \approx 2 \times 10^{68}$. Knowing that the value of N in typical clustering problems can be 10^2 to 10^9 or more, we see that the clustering problem is intrinsically hard and requires an exhaustive search. The special

clustering schemes based on the proximity measures between feature vectors must be exploited. We should follow these basic steps in order to develop a workflow for clustering:

1. Feature selection – Features must be properly selected to encode as much information as possible. Parsimony of the features and minimum redundancy among different features are major goals.
2. Proximity measure – This is the measure of the degree of similarity or dissimilarity between two feature vectors.
3. Clustering criterion, which depends on the interpretation of the term “sensible”, depending on the type of clusters expected in the data set, which are characterized as being oblate, elongated, “bridged”, circular etc.
4. Clustering algorithms. The method of choosing a particular algorithmic scheme that unravels the cluster structure of the data set.
5. Validation and interpretation of results are the final procedures of clustering.

There are two principal types of clustering algorithms [95,105,111].

1. Partitional clustering – which finds a one-level partitioning of the data into k disjoint groups.
2. Hierarchical clustering – which finds a hierarchy of nested clusters; may proceed by
 - a. bottom-up – agglomerative schemes,
 - b. top-down – divisive schemes.

Other techniques employing evolutionary algorithms, neural networks and other naturally inspired algorithms have also been in use [90,104].

III.2 Clustering techniques

Non-hierarchical (partitional) clustering algorithms are employed primarily for extracting compact clusters by using global knowledge concerning the data structure. Well-known schemes based on k -means concept [111], consist in finding the global minimum of the following goal function:

$$J(w, z) = \sum_j \sum_{i \in C_j} |x_i - z_j|^2 \quad (11)$$

where: z_j is the position of the center of mass of the cluster j , while x_i are the feature vectors closest to z_j . To find a global minimum of function J , one repeats many times the clustering procedures for different initial conditions [44]. Each new initial configuration is constructed in a special way from the previous results by using the methods from [44,86]. The cluster structure with the lowest $J(w, z)$ minimum is selected. Clones of this method, whose aim is to find the global minimum of (11), are referenced in journals in pattern recognition and books (e.g. c -means, k -medians, ISODATA [72,91,103,104,111]).

Agglomerative clustering schemes consist in the subsequent merging of smaller clusters into larger clusters, which are based on proximity and clustering criteria. Depending on the definition of these criteria, there exist many agglomerative schemes such as: average link, complete link, minimum variance [91, 103, 104,111] and many nearest neighbor algorithms. The hierarchical schemes are very fast for extracting localized clusters with non-spherical shapes. The proper choice of proximity and clustering criteria depend on many aspects, such as the dimensionality of data, noise sensitivity etc. For example, a smart clustering criterion based on linked-list scheme for finding neighbors used for molecular clustering is completely worthless for

clustering N -dimensional data for which it has extremely high computational complexity. All of agglomerative algorithms suffer from the problem of having no properly defined control parameters, which can be matched to the data of interest and hence can be regarded as invariants for other similar data structures.

The majority of the classical clustering algorithms requires some knowledge of the number of clusters. However this number is usually not known *a priori*. Furthermore, these methods do not perform well in the presence of heavy noise or outliers. Recently, new methods have been proposed that deal with noisy data, discover non-spherical clusters and allow for automatic assessment of the number of clusters. Some important examples are the Chameleon [49], DBSCAN [69] and CURE [37] algorithms. Unfortunately, these methods are suited only for low dimensional data and are rather inefficient, thereby limiting their use for data mining of large-scale sets. Other approaches are in great demand for clustering of large data sets of multidimensional data. In a recent innovative work by Frey and Dueck [34] the authors use the concept of “affinity propagation,” which takes as input measures of similarity between pairs of data points. Real-valued messages are exchanged between data points until a high-quality set of exemplars and corresponding clusters gradually emerges. Affinity propagation promises to find clusters with a much lower error than other methods, and it can do this in less than one-hundredth the amount of time.

Clustering schemes do not produce unequivocal or unambiguous results. For low dimensional space (2 or 3D) human eye can decide whether the clustering result is optimal or not. However, it becomes hopeless for higher dimensions. There exist many techniques for visualizing multidimensional clusters. One of them is the multi-dimensional scaling (MDS) (see the overview on mapping techniques in [73,111]) - the powerful non-linear mapping technique. This method is extremely useful for visualization of the multidimensional data in 2-D or 3-D and for interactive extraction of clusters.

III.3 Multi-dimensional scaling

Because human perception is limited only to three spatial dimensions, special techniques are needed to help us understand and explore multi-dimensional data sets. In terms of pattern recognition these techniques are called feature extraction. Mathematically, feature extraction represents transformation of N -dimensional data from the original feature space onto the n -dimensional ($n \ll N$) target space. For a particular problem of data visualization n is 3. This mapping optimizes some measure quantifying the desired properties of the representation. For example, the technique of principal component analysis (PCA) [111] is theoretically the optimum transform for a given dataset in terms of least squares. It maximizes the total variance in the target space. In general, feature extraction methods consist of linear and non-linear algorithms.

The principal component analysis (PCA), independent component analysis (ICA), linear discriminant analysis (LDA), entropy or divergence minimization [111] are mathematically defined as orthogonal linear transformations that map the data to a new coordinate system. However, linear transformations are too limited in exploring complicated cluster shapes with bridges, noise, different densities and/or positioned on manifolds embedded in higher dimensional spaces. This situation is very common, especially when there exist strong non-linear correlations between different features. Such correlations are expected for the set of features defined on the basis of narrower set of independent variables, e.g., as it is in the M8 algorithm [50,51]. Moreover, the histogram of distances between feature vectors in the original and the target spaces differs qualitatively. Thus, linear mapping cannot preserve adequately the complicated multi-resolutional structures from the original space in the target space of lower dimension [73].

Multi-dimensional scaling or MDS is a non-linear feature extraction method [25,73,111]. The MDS algorithm is based on the “stress function” criterion. The goal is to maintain all the

distances between points $\mathbf{R}_i \in \omega \subset \mathcal{R}^N$ in the Euclidean 3-D (or 2-D) space with a minimum error. The “stress function” can be written as follows:

$$E(\omega, \omega') = \sum_{j < i} D_{ij}^{w \cdot m} \cdot (D_{ij} - r_{ij})^m = \min \quad (12)$$

where: $r_{ij} = (\mathbf{r}_i - \mathbf{r}_j) \circ (\mathbf{r}_i - \mathbf{r}_j), i, j = 1, \dots, M,$

and D_{ij} — is a squared distance between points $\mathbf{R}_i, \mathbf{R}_j \in \omega \subset \mathcal{R}^N$ and $\mathbf{r}_i, \mathbf{r}_j \in \omega' \subset E^3$ — coordinates of the respective points in 3-D Euclidean space. The values of w and m are the parameters of transformation. In fact, Eq.12 describes the group of stress functions in which the choice of m and w parameters allow for extracting particular properties of the multidimensional feature space. For example, by assuming that w is positively definite we can enhance the global properties of the feature space. Otherwise, its local characteristics are better reconstructed.

There are many other non-linear techniques with more precisely defined cost function such as the curvilinear component analysis (CCA) or local MDS [52,77], curvilinear distance analysis (CDA) [52], locally linear embedding (LLE) [79,80], isomaps [77], which use distances based on combination of the Euclidean metrics and k-nearest neighbor proximity measures. However, the result of mapping depends on the quality of the minimum obtained for the cost function. Usually the dimensionality of the function domain is very high and is equal to $N \cdot M$, i.e., thousands, in the smallest and billions in large problems. For more than $M=10^3$ feature vectors, the high dimensionality of source space and data complexity may cause the resultant low dimensional patterns to be completely illegible [25]. Especially, for cost functions, which employs k-nearest neighbors based metrics, the application of standard numerical algorithms for finding global minimum is hopeless. Its computational complexity is simply too high.

Therefore, for visualization of $M > 10^3$ multidimensional data samples we recommend MDS techniques with the group of “stress functions” given by Eq.12. For MDS the fast and reliable minimization techniques extracting global minimum (or close to it) of the “stress function” exists. In [25] we proposed a N-body solver by ODE's as a heuristic means. The algorithm is as follows:

1. The initial configuration of M interacting “particles” is generated in E^3 ,
2. Every “particle” corresponds to the respective N -dimensional point from \mathcal{R}^N ,
3. The “particles” interact with each other with a two-body $\Phi_{i,j}$ particle-particle potential:

$$\Phi_{ij} = \frac{1}{4} \cdot k \cdot D_{ij}^{w \cdot m} \cdot (D_{ij} - r_{ij})^m \quad (13)$$

(k -is the stiffness factor) and the energy produced is dissipated by the friction force proportional to the velocity of the particles.

4. The entire system consisting of thousands to millions of particles evolves to an equilibrium configuration according to the Newtonian equations of motion.

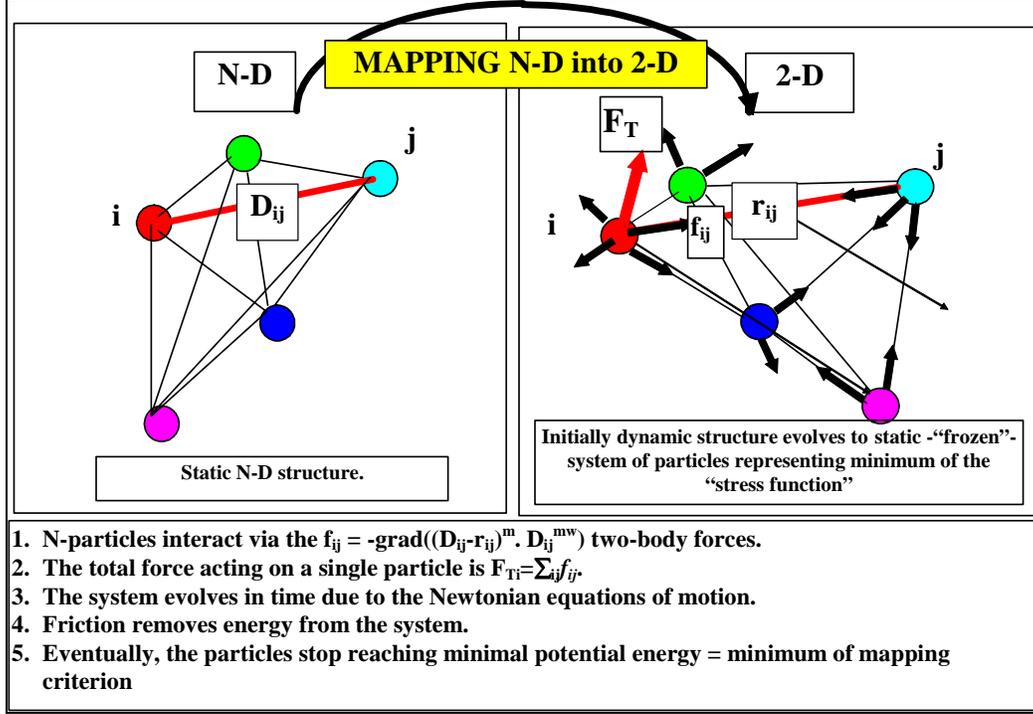


Fig.1 The conceptual diagram of MDS transformation by means of the molecular dynamics approach.

In this way the interactions between each pair of particles are described by various springs similar to intermolecular potentials, which depend on the separation distance between particles r_{ij} and the distance D_{ij} between respective multi-dimensional points in \mathfrak{R}^N . If the distance between particles i and j in the output 2(3)-D space is shorter than the distance between the respective i and j feature vectors in the source N -D space, these points then repel one another. Otherwise, i.e., the distance becomes longer, the particles attract one another. By using the *leap-frog* numerical scheme [100], the following formula for velocities and positions of “particles” can be derived from the Newtonian time-differenced equations:

$$\mathbf{v}_i^{n+1/2} = \frac{(1-\varphi)}{(1+\varphi)} \cdot \mathbf{v}_i^{n-1/2} + \frac{\alpha \Delta t}{(1+\varphi)} \cdot (D_{i,j})^{-wm} \cdot \left\{ \sum_{j=1}^K (r_{i,j}^n - D_{i,j})^{m-1} \mathbf{r}_{i,j}^n \right\} \quad (14)$$

$$\mathbf{r}_i^{n+1} = \mathbf{r}_i^n + \mathbf{v}_i^{n+1/2} \cdot \Delta t$$

$$\alpha = \frac{k}{m}, \quad \varphi = \frac{\lambda}{2m} \cdot \Delta t < 1$$

where \mathbf{v}_i^n , — the particle i , n — the time-step number, $m=1$ — particle mass.

As it is common in molecular dynamics [100], the system of “particles” evolves in time until the global (or close to the global) minimum of Eq.(12) (the total potential energy of the particle system) is attained. Two free parameters, λ and k , have to be fit for obtaining stable state, where the final positions of frozen “particles” reflect the result of the nonlinear mapping from N -D to 3-D. The conceptual scheme of multidimensional scaling by exploiting the N-body solver is shown in Fig.1. In Fig.11 we present the 7-dimensional feature space which has been transformed by using the MDS procedure and then mapped onto the 3-D space. Please take a look on the movies (Movie 1 and 2 in Supplementary Materials), which shows how rotation in the 3-D space can help in a better recognition of the cluster.

IV Data description

IV.1. Earthquake datasets

We analyze the observed and synthetic earthquake catalogs for three time intervals of 5, 150 and 1500 years respectively. The observed data (Fig.2) represents seismic activities of the Japanese islands collected by the Japan Meteorological Agency (JMA).

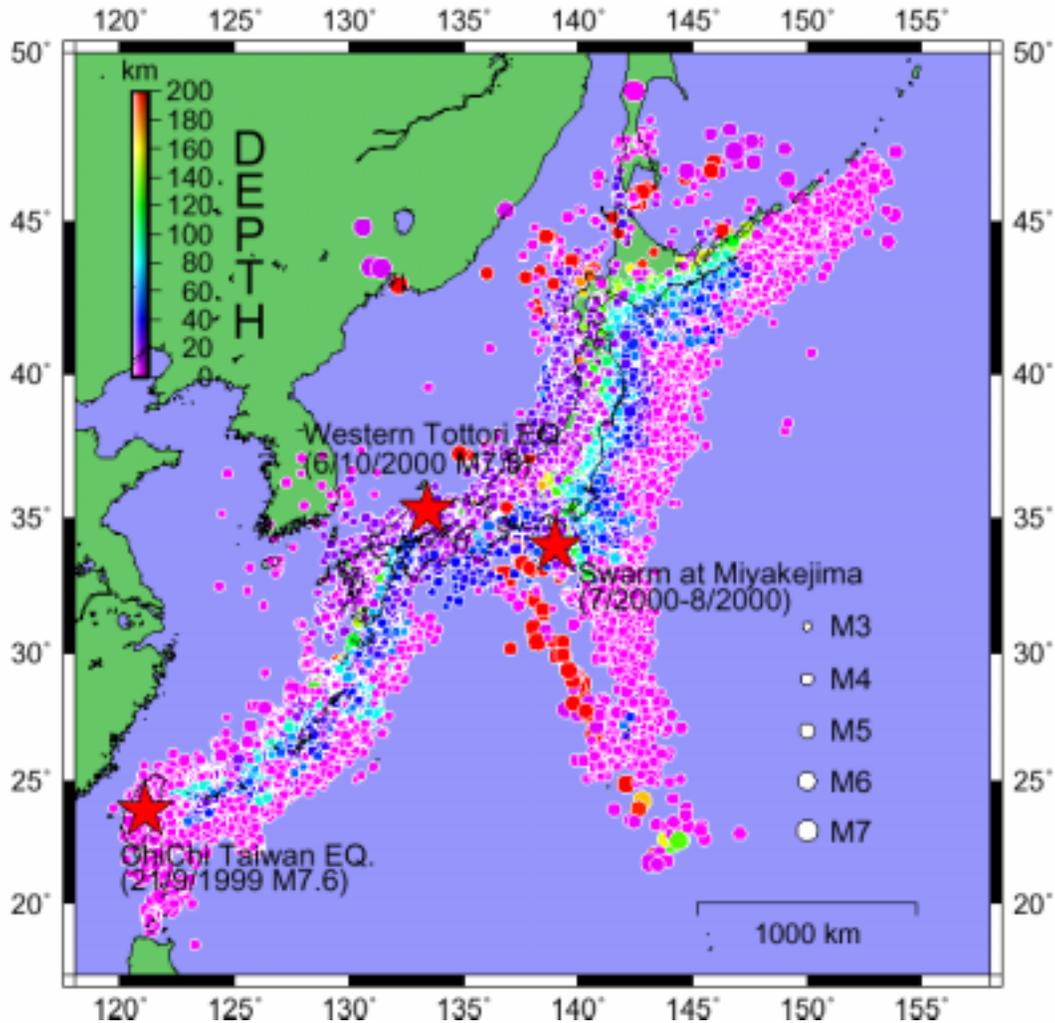


Fig.2 Seismic activities around the Japanese Archipelago within a time span of 5 years. We use the hypocentral data provided by the Japan Meteorological Agency (JMA). The magnitude of the earthquakes (JMA magnitude) and their depths are represented by differences of the radius of the circle and colors, respectively. The red stars symbolize large events such as: Chi-Chi Taiwan earthquake (21/9/1999 M7.6 latitude 23.8 longitude 121.1) Swarm at Miyakejima (7/2000-8/2000 latitude 34.0 longitude 139.0) Western Tottori earthquake (6/10/2000 M7.3 latitude 35.3 longitude 133.4) (from [27]).

The JMA catalogue consists of 915,829 events detected in Japan Islands between 1923 and January 31, 2003. The original catalogue includes also events with magnitudes less than 1.0. The lowest magnitudes were determined by using a detection level, estimated from the Gutenberg-

Richter frequency-magnitude earthquake distribution. We have assumed that the cutoff magnitude of earthquake is equal to 3 ($m > 3$). We do not use any cutoff depth of hypocenter events. The seismic events, shown in Fig.2, were recorded during the 5 year time interval from October 1, 1997 to January 31, 2003. The data set processed consists of $M=42,370$ seismic events with a magnitude m , position in space (latitude X , longitude Y , depth z) and occurrence time t .

To understand more deeply seismic activities over longer time periods, we use data from the synthetic catalogs generated by numerical earthquake models [10].

IV.2 Physical model of earthquake dynamics

The synthetic catalogs are generated by the model of Ben-Zion [10] for a segmented strike-slip fault zone in a 3D elastic half-space, based on earlier developments of Ben-Zion and Rice [8, 9]. This dynamical model attempts to account for the statistical properties of earthquake ruptures on long and narrow fault zones with bends, offsets, etc (Fig.3a), represented by a rectangular patch field structure on a 2D plane with discrete cells and spatial variations of frictional parameters (Fig.3b). The model contains a computational grid (region II of Fig.3b) where the evolving stress field and seismicity are generated in response to ongoing loading imposed as slip boundary conditions imposed on other fault regions. Regions III and V creep at constant plate velocity of 35 mm/yr, while regions I and IV follow staircase slip histories with recurrence times of 150 yr. The stress transfer due to the imposed boundary conditions and failing grid cells is calculated by using a discretized form of a boundary integral equation and employing the elastic solution for dislocations in a 3D elastic half-space [8, 60].

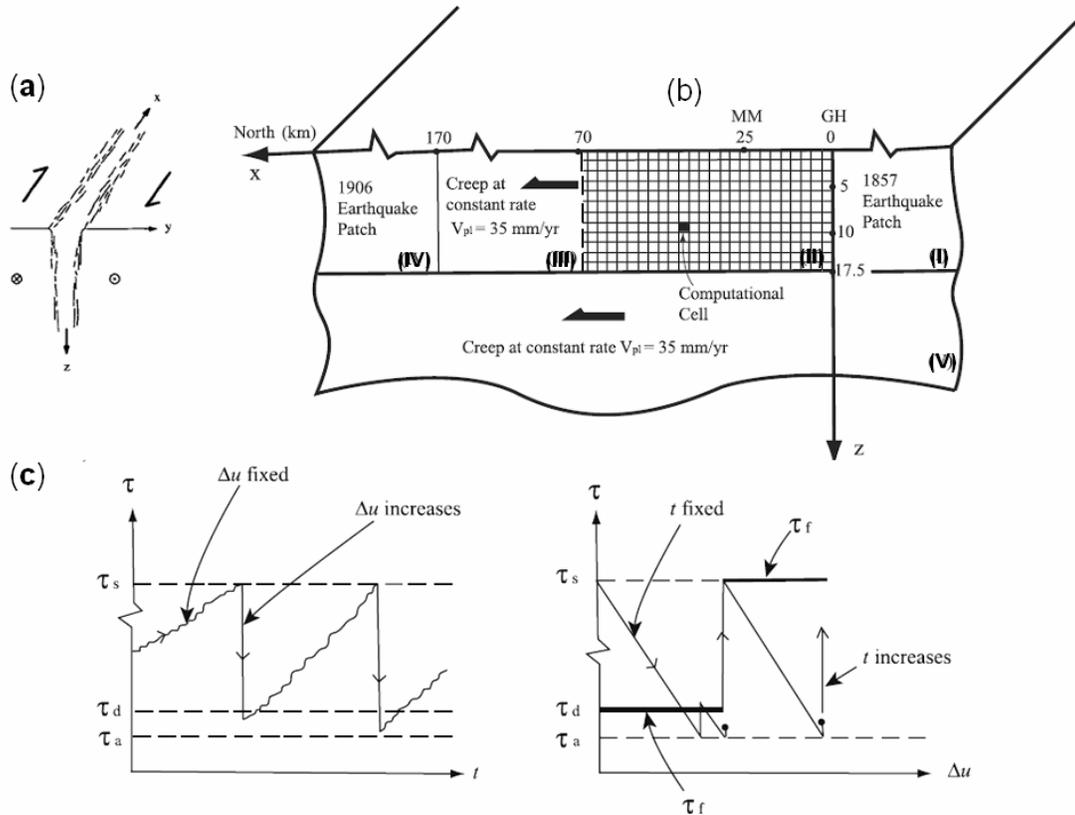


Fig.3 The schematics of the model of Ben-Zion [10,14] for a segmented strike-slip fault zone embedded in a 3D elastic half-space.

Deformation at each computational cell is the sum of slip contributions from both brittle and creep processes. The brittle process (Fig.3c) is governed by distributions of static friction τ_s , dynamic friction τ_d , and arrest stress τ_a . The static friction characterizes the brittle strength of a cell until its initial failure in a given model earthquake. When stress τ at a cell reaches the static friction, the strength drops to the dynamic friction for the remaining duration of the event. The stress at a failing cell drops to arrest level τ_a , which may be lower than τ_d in order to accommodate dynamic overshoot, producing local slip governed by elastic dislocation theory [19,60]. The static friction, dynamic friction, and arrest stress are connected via a dynamic overshoot coefficient $D=(\tau_s-\tau_a)/(\tau_s-\tau_d)$. If stress transfer from failing regions increases the stress at other cells to their static or dynamic strength thresholds, these cells fail and the event grows. When the stress at all cells is below the brittle failure thresholds, the model earthquake ends and the strength at all failing cells goes back to τ_s . The creep process is governed by a power-law dependence of the creep law on the local stress and space-dependent coefficients that increase exponentially with depth and with distance from the southern edge of the computational grid. The chosen parameters produce an overall "pine-tree" stress-depth profile with a "brittle-ductile" transition at a depth of about 12.5 km, and variable stress-along-strike profiles with a gradual "brittle-creep" transition near the boundary between regions II and III (see Ben-Zion [10] for additional details). The model generates many realistic features of seismicity, which are compatible with observations, including the frequency-size and temporal event statistics, hypocenter distribution with depth and along strike, intermittent criticality, accelerated seismic release, scaling of source time functions and more (e.g., [14,31,53,87]).

IV.3 Synthetic catalogs

Synthetic data generated by computational models comprise many events, which are able to cover large spatial areas and extremely long time spans. Moreover, the synthetic data retain the statistical reliability of the results. The data are free of measurement errors, which occur in estimating earthquake magnitudes and hypocentral locations, and do not suffer from incomplete recording of small events, which exist in natural catalogs. These are significant advantages for our study, which attempts to illustrate clearly the performance of both clustering analysis and visualization techniques.

In section 6 we will analyze synthetic catalogs generated by two model realizations (A and M) of Ben-Zion [10,11]. The catalogs contain the time, location and magnitude of earthquakes calculated by the model for 150 and 1500 years. Extensive numerical simulations with several different classes of models, summarized by Ben-Zion [10,11] and Zöller et al. [69,70], suggest that the degree of disorder in fault heterogeneities is a tuning parameter of the earthquake dynamics. Catalog A is generated by a model specifically tailored to the Parkfield section of the San Andreas fault. Catalog M is generated by a realization of a more-disordered system like the San Jacinto fault or the Eastern California Shear Zone in Southern California. In both data sets the time interval covers all events ($M \sim 1-3 \times 10^4$) that have occurred over the last 150 years of simulated fault activity. These simulations were repeated for a ten times longer time scale i.e. 1500 year interval (the number of events $M \sim 10^5$), covering hundreds of large earthquakes ($m > 6$) and a correspondingly wider time window.

The seismicity parameters were obtained by averaging the data using a sliding time window of constant width ΔT and shift dt . We employ $\Delta T = 10$ days and $dt = 2$ days for the Japanese data, $\Delta T = 10$ months and $dt = 2$ months for synthetic data covering 150-years and $\Delta T = 30$ months and $dt = 6$ months for the data covering time period of 1500 years. Each parameter in the clustering was normalized with respect to the standard deviation.

V Earthquakes clustering in space and time coordinates

The same notion of clustering also holds for geographical locations and other properties of earthquakes. In Fig.4a we present a spatial distribution of earthquake epicenters in the western hemisphere of the Earth (data from <http://quake.geo.berkeley.edu/cnss/maps/cnss-map.html>). One can see with the naked eye that their distribution is far from being uniform. We observe both elongated and oblate structures – the earthquake clusters - separated at this resolution by the large gaps of seismically quiescent areas.

Properties of the clusters result from dynamical characteristics of the generating processes. Visual representations of information on these processes are reflected by the shape and structure of clusters. Therefore, detection of clusters and their analysis should be the first step for knowledge extraction from this information. For example, as shown in Fig.4a, the earthquake clusters on Earth are located in geologically active regions, mainly, on the edges of colliding tectonic plates. The distribution and shape of the earthquake clusters follow the borders between the plates. In Fig.4b we show the large earthquake cluster from Fig.4a located at the US western coast. One can distinguish here many smaller clusters of different density separated by geologically inactive area. A similar pattern (see Fig.4c) is observed by zooming-in on one of denser clusters from Fig.4b. This multi-resolutional and self-similar system is characteristic of many critical phenomena [4,12,18,23,110]. A fractal structure is revealed in the worldwide fault network [24,29,113]. Wavelet-based multi-fractal analysis [29] shows clearly several distinct scaling domains in earthquake catalogues, thereby revealing rich self-similar multi-scale structure. However, the spatial structure of earthquake clusters alone is inadequate to formulate plausible hypotheses about earthquake dynamics. More information is needed, such as the correlation relationships between earthquake events [66].

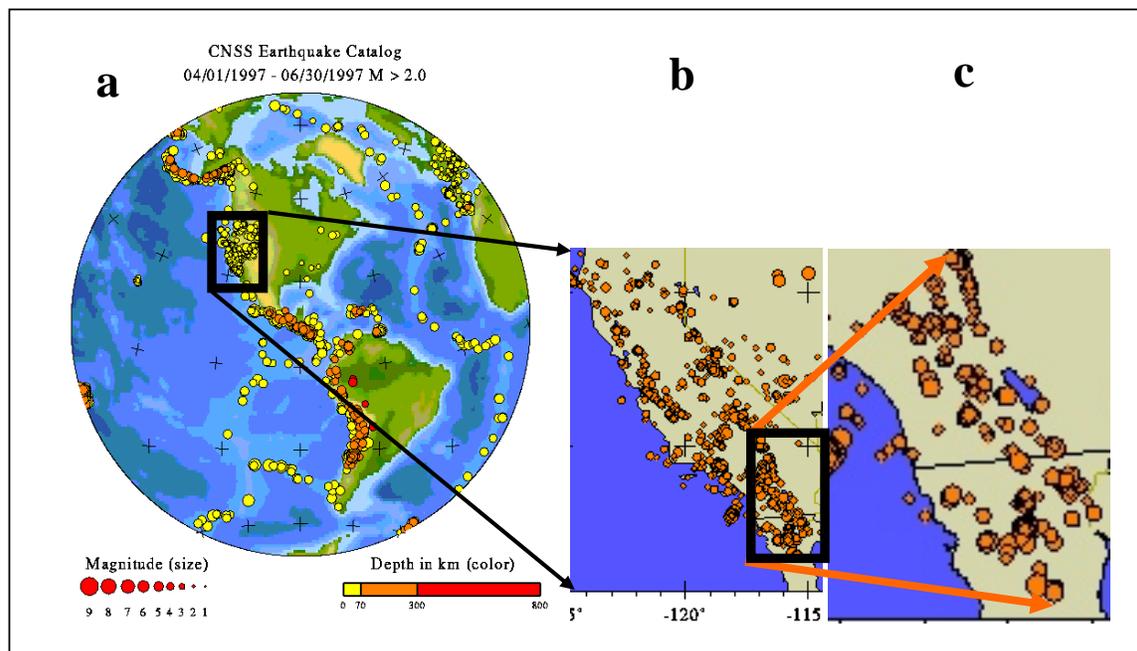


Fig.4 Multiscale character of earthquake clusters. The epicenters of earthquakes of various depth and magnitude are displayed. This dataset comes from the CNSS Earthquake Catalogue (<http://quake.geo.berkeley.edu/cnss/maps/cnss-map.html>). a) the western hemisphere, b) the US western coast c) California and Nevada.

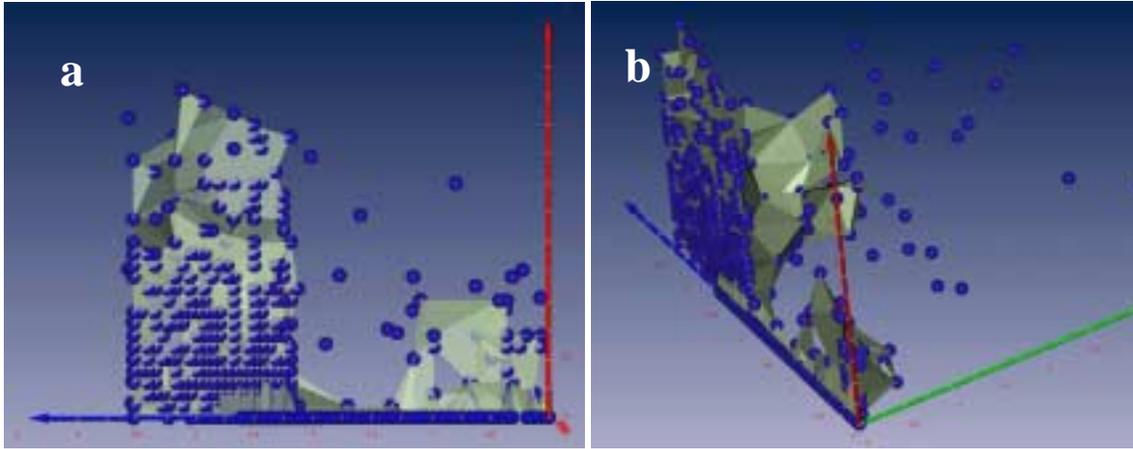


Fig.5 Seismic activity of the Changbaishan volcano during the 5-year time span from July 1999 to May 2004 (the north-east boundary of the North China craton) [43]. The plates represent the seismic events in 3-D feature space attributed by eruption time (blue axis), magnitude (red axis) and distance to the epicenter (green axis) coordinates. The surfaces approximate cluster shapes. Two different positions of coordinates are shown (a,b). A large cluster representing the earthquake swarm is preceded by a small precursory cluster of seismic activity and quiescent time period (see Fig5a). The Amira package (www.amiravis.com) was used for visualization.

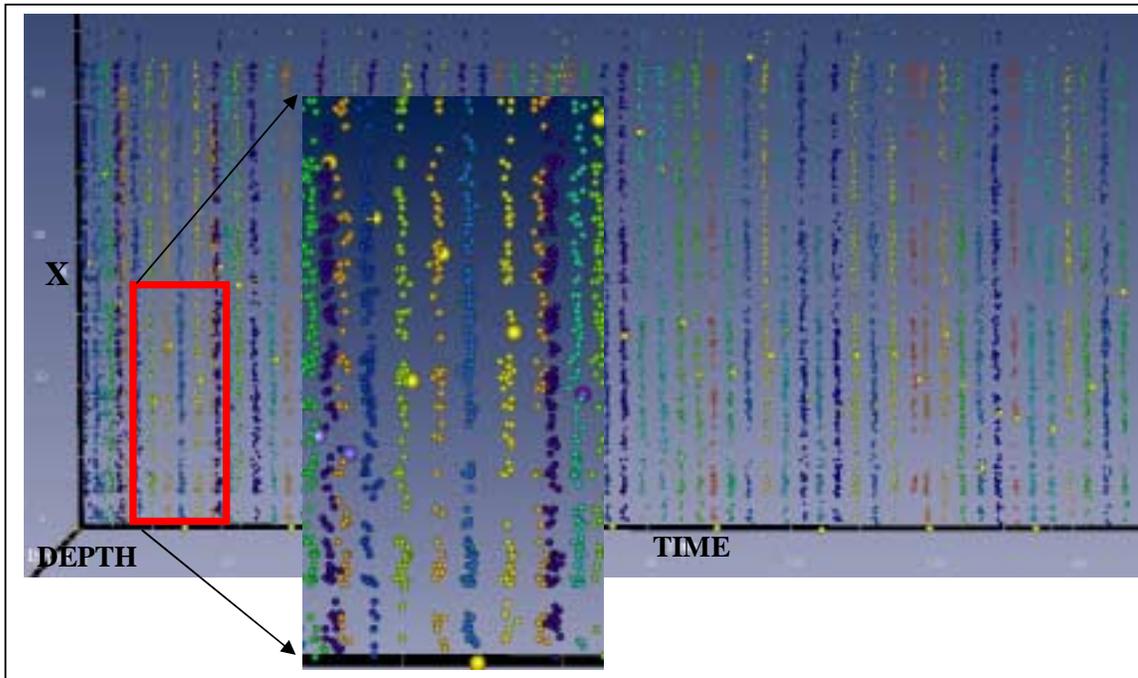


Fig.6 A plot reconstructing the seismic activity during a period of 150 years from the synthetic data [10] (horizontal distance - X , depth - z ; visualized by using the Amira visualization package [2]). Large events (with magnitude $m > 6$) are shown as distinctly larger dots (yellow) on the background of the lowest magnitude events ($m < 4$). One can observe the patches of low magnitude events preceding larger events [26]. The separate clusters are marked in colors. An agglomerative clustering scheme, employing the concept of the nearest mutual neighborhood [36], was used for extracting the clusters. We have employed the zoom-in option for looking at finer details.

As shown in Fig.4, besides the geographical location, earthquakes have additional features such as the time and depth of occurrence and the amount of energy released (proportional to $10^{\alpha m}$ with $\alpha \sim 1.5$ and m the magnitude). These attributes can be used as additional coordinates of feature space (e.g., [111]). In Fig.5 we display the earthquake clusters representing the seismic activity nearby the Changbaishan volcano in an abstract 3-D feature space. Apart from the geographical location - represented by the distance from the epicenter - other coordinates (features) are employed: the time of occurrence and the magnitude of the earthquake. As shown in Fig.5, the large cluster of seismic activity is preceded by the small precursory cluster and low activity region. The larger cluster is characterized by the seismic events from a broader interval of magnitudes and with outlying earthquakes more distant from the epicenter than in the preceding smaller cluster.

The dynamics of the volcanic earthquakes covers only a period of 5 years. The time is too short to draw any conclusions about the intermediate time-scale earthquake dynamics driven by magma injection. To obtain data covering much longer time period we used synthetic data generated by numerical simulations of seismicity on a heterogeneous fault governed by 3-D elastic dislocation theory, power-law creep and boundary conditions corresponding to the central San Andreas Fault [10,30,31]. In Fig.6 we represent seismic activity during 150 years. This period contains $M_f \sim 1-3 \times 10^4$ events (represented in Fig.6 by colored dots) in the magnitude interval [3.3-6.8]. Unlike in the Changbaishan case, the seismic events have one more feature - the earthquake depth. Thus the feature space has now four dimensions. In Fig.6 we display the data distribution in time-depth-position 3-D space. The fourth dimension - the magnitude - is displayed in Fig.6 by the size of the dot. To make the situation clearer only the large earthquakes with magnitudes $m > 6$ (large dots) and the smallest ones $m < 4$ (small dots) are distinguished in Fig.6. As shown in Fig.6 and in [26], the synthetic seismic events with magnitudes $m < 4$ produce stripe-like clusters in the data space. They precede the large earthquakes ($m > 6$) and are separated in time by the regions of mixed type of events (i.e., with $4 < m < 5$). A similar space-time seismicity pattern for the real data is shown in [3]. This represents the 2006-2007 Kuril islands earthquake sequence relative to the 15 November 2006 event.

Another system of earthquake clusters is displayed in Fig.7 The synthetic data ($M_f \sim 10^5$ events) corresponding to the seismic activity during 1500 years were generated by the same model [10] for similar geological and boundary conditions. Only medium-size events with $4.5 < m < 6$ were taken for clustering. In addition to the local striped clusters of smaller events ($m < 4$) detected for 150-years data, one can observe in Fig.7 a distinct spatio-temporal patchwork structure of clusters consisting of medium sized events ($4.5 < m < 6$). These clusters follow spatio-temporal changes in strength-stress physical properties of the fault in the region being modelled.

In summary, we can highlight that very fundamental properties of earthquakes, multi-resolutional clusters are built up by the earthquake epicenters. Clustering shows that earthquakes represent a dynamical process involving many spatio-temporal scales. The dynamical nature of earthquake clusters in a very long time horizon is obvious because everyone can expect that tectonic plates will change dynamically the geo-mechanical properties of the Earth crust. In a longer time period covering thousands years, the patterns from Fig.4 will evolve by following the changes in the fault network. More mysterious is the character of earthquake dynamics in spatio-temporal scales, which allow one to make realistic predictions. We show that in the medium-time period lasting more than a hundred of years the seismic events may produce periodic system of clusters in approximately equal time intervals with both increasing and decreasing seismic activity. The large earthquakes, preceded by the quiescent time periods, then appear. Moreover, the short-time dynamics reveal that the earthquake swarms are signaled by the smaller precursory cluster of seismic activity.

We showed that additional earthquake attributes such as the magnitude and the epicenter depth, allow for a better interpretation of spatio-temporal clusters and exploration of hypotheses

space. For studying various aspects of earthquake dynamics, including their prediction, we have to analyze the cluster structures in a multi-dimensional feature space. The features can be more abstract than just the location, time, magnitude and depth. They should encode important dependences and correlations between events, which cannot be perceived explicitly from the raw data, thus ensuring that none of important information will be lost.

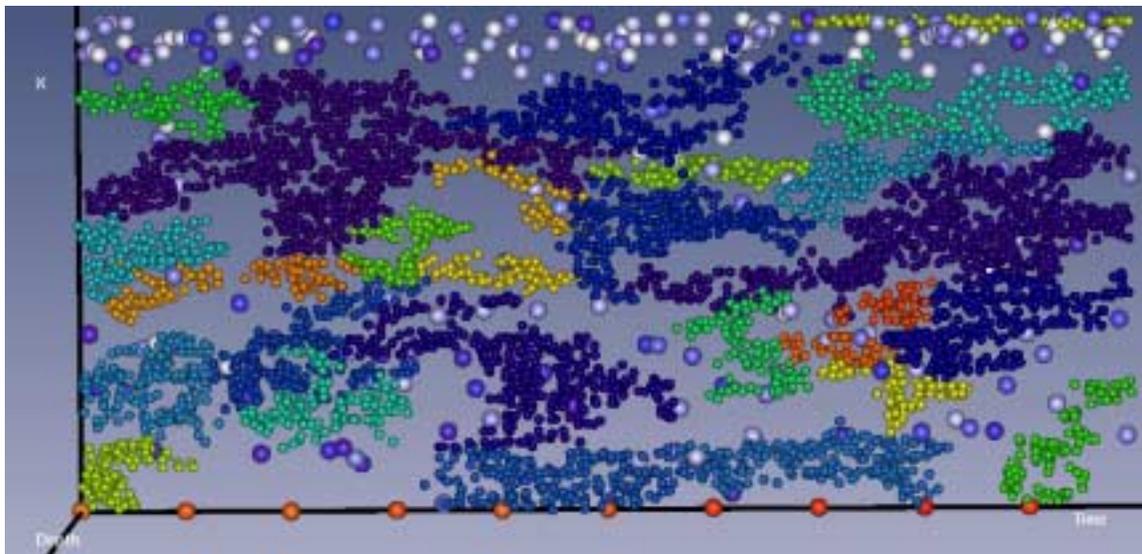


Fig.7 A plot reconstructing seismic activity during 1500 years from synthetic data [10]. The largest clusters obtained for events with magnitudes $4.5 < m < 6$. Large events ($m > 6$) are shown as distinctly larger plates. The separate clusters are marked in colors. The agglomerative clustering scheme SNN [KUMAR] based on the concept of the nearest mutual neighborhood [36] was employed for extracting clusters.

VI Multi-dimensional feature space

In Figs.2,4 every point i representing one of the M_f earthquakes has two dimensions - the geographical coordinates $\mathbf{x}_i = [x_1, x_2]$ of the epicenter. The point can be treated as 2-D vector \mathbf{f}_i in the feature space where $\mathbf{f}_i = \mathbf{x}_i$. Assuming additional coordinates, at the highest level of resolution, we can represent a single seismic event i as a five-dimensional data vector $\mathbf{f}_i = [m_i, z_i, \mathbf{x}_i, t_i]$ where m_i is the magnitude and \mathbf{x}_i, z_i, t_i - its epicentral coordinates, depth and the time of occurrence, respectively. The spatio-temporal clusters can be extracted by 3-D visualization similar to those of Figs.6,7 by distinguishing the extra dimension by the size of dots and colors. Only clusters in the three spatially visualized dimensions can be extracted, while the other attributes associated with the earthquake characteristics are used for interrogating the different types of clusters.

As shown in Fig.6 and 7 and in [26], at the lowest resolution level we can analyze the data locally by looking for clusters with similar events. However, from a general point of view, considering a single event on a given area as a feature vector [111] cannot be regarded as a good approach. The number of events is usually large. There are many noisy background events, which destroy the relevant clusters or produce artificial ones. Moreover, clustering of raw data neglects the important statistical information, which covers the entire inspected area. An alternative approach exists in which the entire seismic area can be described as a multi-dimensional feature vector evolving in time. In the following these features will represent descriptors a_k (seismicity parameters) corresponding to different statistical properties of all the events measured in a given time interval. The number of descriptors N defines the dimensionality of the feature vector

$F_i=[a_1, a_2, \dots, a_N]$, $i=1, 2, \dots, M$. The vector represents not a single seismic event but it corresponds to seismic situation on the whole controlled area in the subsequent time interval indexed by i . The number of feature vectors M is equal to the number of time intervals in which the descriptors are computed. The index i is a discrete equivalent of time. We expect that the features vectors representing different moments of time have also the tendency to produce clusters in the abstract N -dimensional feature space. Monitoring the changes of these time-series in abstract N -dimensional space may be used as a proxy for the evolution of stress and the development of a large earthquake cycle over a heterogeneous fault [14].

To explain this approach better, let us assume that we have to analyze the customer behavioral pattern in a hypermarket. We can watch every customer separately by assuming that it can be defined as a feature vector consisting of only two coordinates: the time he entered the shop, the money spent. Then we can try to find clusters emerging with time during a shopping day. This cannot be easy due to both a large number of feature vectors (clients) producing statistical noise and lack of correlations between them. Another approach consists in treating as a feature vector not a single client but every subsequent time interval $t_i=i \times \Delta T$ ($i=0, 1, 2, \dots, M_F$; $t_i < t_e$; $M_F=(t_e-t_b)/\Delta T$ and t_b -beginning of the working day and t_e – closing time). Let the coordinates of the subsequent feature vector define the following descriptors averaged in Δt : the number of people inside the shop (crowding), the flow, items bought, money spent per person. We note that now the number of feature vectors will be substantially smaller than in the previous approach but the dimensionality of feature space is larger. Let us assume that as a result of clustering we extract two distinct clusters. The first one consists of feature vectors (time intervals) from between 10:00-11:00 hr and 13:00-14:00 hr. The cluster is characterized by very small values of the first three descriptors (crowding, flow, numbers of items sold) and relatively large expenses. The second cluster consists of time intervals from between 8:00-9:00 hr and 16:00-18:00 hr with all descriptors large. We can conclude that the first cluster consists of time intervals from shopping hours that are the favorite for wealthy retired people from the rich village in the neighborhood, while the second cluster is associated with a frenetic shopping just before the beginning and after the end of the working hours.

In the same way, clusters of the feature vectors (time intervals) consisting of seismicity parameters should reflect the similarity between seismic activities in various time intervals. As we show before, the large seismic events are preceded by precursory events, reflected by an abnormal seismic activity in the whole area. We suppose that these instants of time are similar within the context of the set of seismicity parameters selected. For example, we could expect that feature vectors corresponding to precursory events should belong to the same cluster. We can also detect whether the current feature vector is the member of this cluster in earthquake prediction.

A similar concept was used in M8 algorithm of Keilis-Borok and Kossobokov [50,51], who employed multi-dimensional feature space to detect regions or time cycles of increased earthquake risk. The M8 algorithm is based on seven descriptors from smaller earthquakes within a specified “region of investigation” that are used as the feature space coordinates. The “Time of Increased Probability” (TIP) for larger events ($M>8$) in that region is searched in the following 6 month period. The M8 algorithm has played an important and often controversial role. Its strength is that it produces relatively well predictions. Their shortcomings are as follows [30,40]:

1. The algorithm is purely empirical without any physical interpretation.
2. Spatial information about seismicity is incorporated in a minimal way.
3. It depends on the interactions of large number of terms, making the result difficult to analyze and interpret.
4. Rigid specifications of the region are used (magnitudes and time intervals).

This makes its performance very sensitive to locations of critical events. Small measurement error can result in bad predictions. Moreover, no visualization was carried out on the results.

It is not known what the optimal parameters are for extracting informative signals from seismic catalogs, because the underlying physics is not yet understood. In these circumstances, a reasonable approach is to choose a set of parameters that sample the response of a fault system using observables in all possible domains, namely space, time, and size. The parameters sampling the different domains should consist of a relatively small set (for tractability), should have relatively small inter-correlation (to increase the information content), and should be physically motivated (to increase the likelihood of relevance to physical faulting). Here we use a set of seven parameters chosen by Eneva and Ben-Zion [31] based on the above guidelines. These parameters are not necessarily the best, but they have been shown by Eneva and Ben-Zion [31] and Ben-Zion et al. [14] as capable of extracting relatively large amount of relevant information on complex evolution of large earthquake cycles. Similar to the M8 descriptors, the results are still difficult to analyze and understand. This is the major reason for using more sophisticated pattern recognition methods by employing on-line visual exploration of multidimensional structures in the feature space.

The seismicity parameters are computed as time and space averages for a given time and space intervals within a sliding time window with a length ΔT and time step dt . The values of a_k represents one of the following seismicity parameters: **NS**, **NL**, **CD**, **SR**, **AZ**, **TI**, **MR**. The value of dt was assumed to be equal to the average time difference between two recorded consecutive events while ΔT is equal to about 1/10 of the average time distance between two successive large events ($m > 6$ or $m > 5$). By increasing the values of dt and ΔT , one can obtain smoother time series due to better statistics. On the other hand, we can then expect poorer prediction characteristics. We can define the seismicity parameters as follows [30,31]:

Table 1 Definition of seismicity parameters

NS	Degree of spatial non-randomness at short distances. The differences between distributions of event distances and distances between randomly distributed points.
NL	Degree of spatial non-randomness at short distances.
CD	Spatial correlation dimension calculated on the basis of correlation integrals and on interevent distances.
SR	Degree of spatial repetition represents the tendency of events with similar magnitudes to have nearly the same locations of hypocenters.
AZ	Average depth of the earthquake occurrence.
TI	Inverse of seismicity rate – time interval in which a given (constant) number of events occurs.
MR	Ratio of the numbers of events falling into two different magnitude ranges $=M_f(m \geq M_0) / M_f(m < M_0)$.

The seismicity parameters produces seven time series and create the abstract 7-dimensional feature space of time events $F_i = (NS_i, NL_i, CD_i, SR_i, AZ_i, TI_i, MR_i)$ where i are discretized values at time $t = t_b + i\Delta T$. In Fig.8 we display an example set of seismicity parameters (with an average magnitude **M**) for synthetic data [27]. The precise location of the clusters and the visualization of the clustering results are significant challenges in clustering over a multi-dimensional space. In the following section we present briefly the basics of clustering and algorithms needed in this venture.

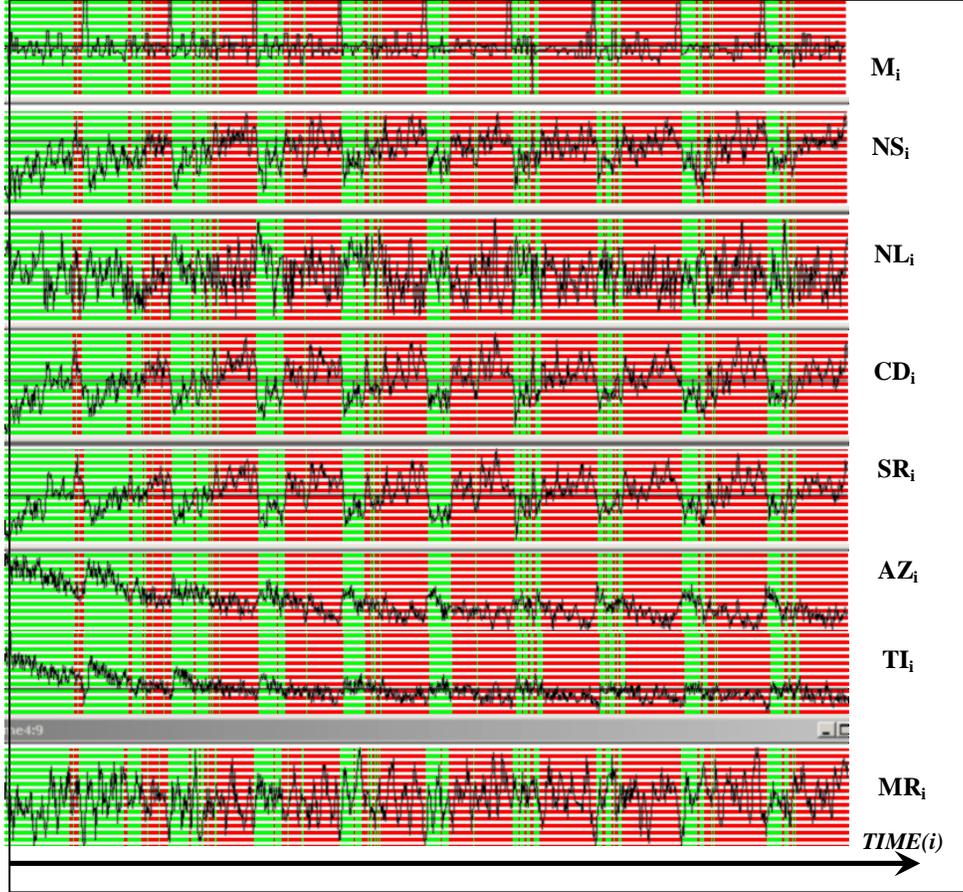


Fig.8 An exemplary set of seismicity parameters $\{\mathbf{M}, NS, NL, CD, SR, AZ, TI, MR\}$ evolving in time (i - subsequent number of the feature vector) for A file from the 1500-years synthetic data catalog (from [27]). The green and red strips show the time moments belonging to the two different clusters. The green cluster corresponds to the time intervals of lower, while the red cluster represents higher seismic activities. The above time series represent about $M_F=10^3$ feature vectors F_i .

VII Visual Clustering of Earthquakes in Feature Space

VII.1 Short-time period

Results of clustering of the observed Japanese seismic catalogs (see Fig.2) both in raw data and in feature spaces are shown in Figs.9. At the data resolution level a single seismic event i can be represented as a multi-dimensional data vector $f_i = [m_i, z_i, X_i, Y_i, t_i]$ where: m_i is the magnitude, X_i - the latitude, Y_i - the longitude, z_i and t_i - the depth and the time of occurrence, respectively. The seismic events are visualized with the Amira package in Fig.9A-B as irregular clouds of colored dots with (z, x, t) coordinates.

In accordance with the Gutenberg-Richter earthquake relationship, we find that the number of events from various ranges of magnitudes differs considerably, and divide the entire set of data onto three subsets. The first one comprises the small, the second medium and the last one represents the largest earthquakes displayed in Figs.9A-B as big dots. The deepest earthquakes for depths greater than 150 km are not displayed in the Fig.9. The various shades represent the magnitudes of earthquakes from $m=6$ (green) to $m=7$ (red). In Figs.9 we present the

clustering results in both the data f_i and the feature F_i spaces. We look for clusters of similar seismic events (data space) and time events (feature space). The dots (data vectors), belonging to the same clusters, have the same color. Fig.9A-B is very rich in cluster-like forms, some of them hard to interpret. Correspondence of the cluster structure of data f_j ($j=1, \dots, M_f$) with the clusters of averaged events F_i ($i=1, \dots, M_f$) in the feature space can reveal valuable information. As one can see from the panel C, only three clusters are obtained in the feature space consisting of small data events ($3 < m < 4$). The green cluster corresponds to two relatively large time intervals of small events preceding Miyakejima earthquake and many smaller post shock periods.

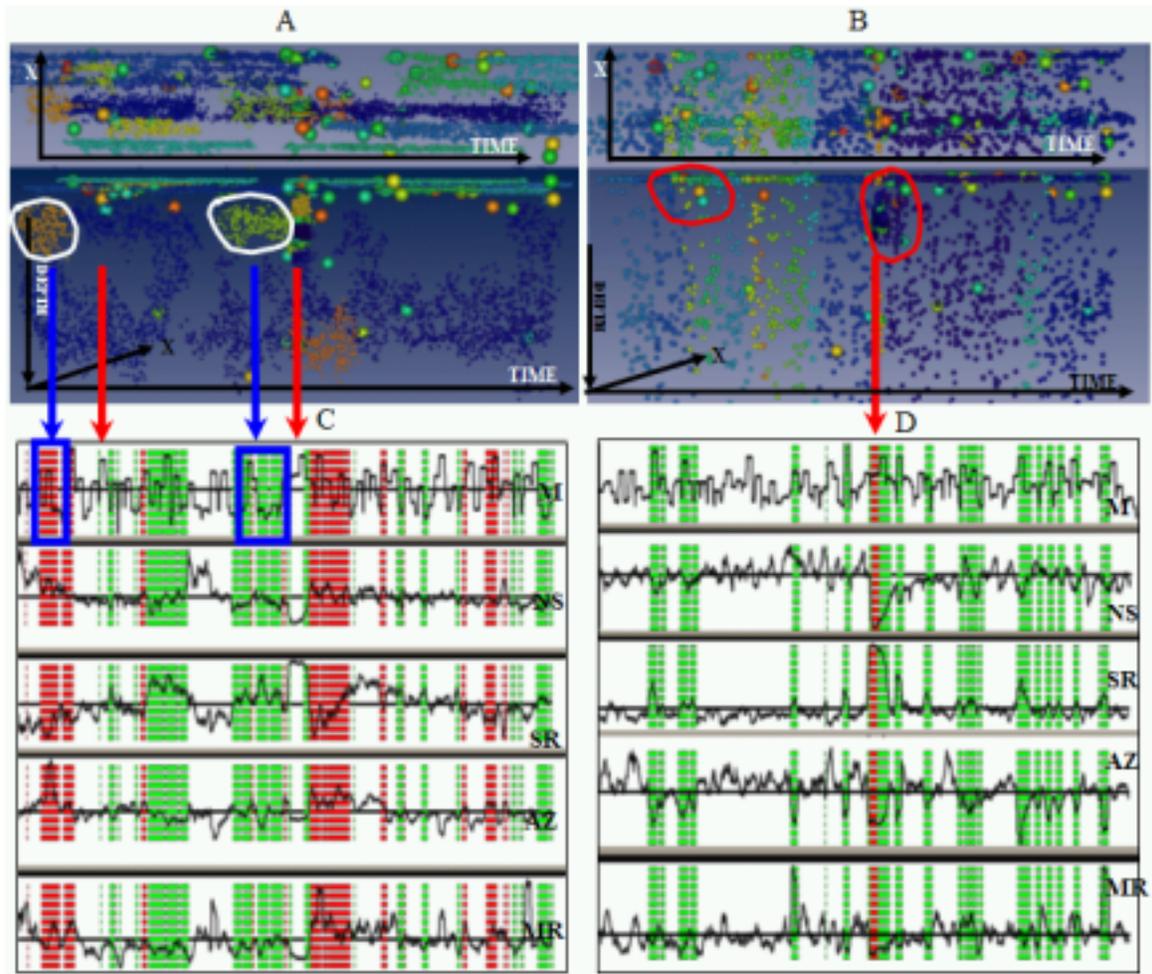


Fig.9 Actual seismic data from [45] analyzed by using clustering in both the data (A,B) and the feature (C-D) spaces. In panels A and B one can see the results of clustering in the data space (from two different perspectives, X-Time and Depth-Time) for small ($3 < m < 4$) and medium magnitude ($4 < m < 6$) events, respectively, represented by small dots. The different colors of the dots denote different clusters. Large events are visualized by the larger spheres. Their colors show the difference in magnitudes m (red – the largest, green - the smallest). The clusters in panels A-B encircled in red display the places with the largest seismic activity, while those in white represent clusters of small precursory events. The red, white and green stripes in panel C and D representing 4 (out of 7) seismic parameters and maximum magnitude M show the clusters of similar time events for situations corresponding to panels A and B, respectively. The Amira visualization package was used [2].

The time events F_i from this cluster represent averaged data events f_j , mainly shallow (**AZ**) with a high degree of spatial repetitiveness **SR** and small variations in magnitudes (**MR**). The red cluster consists of deeper events with a smaller rate of repetition, and more scattered in magnitudes. A longer time interval of this type of behavior is recognized just after the Miyakejima shock. The white cluster is not interesting in this scale of small events and includes all other events including the earthquake swarm.

In panel D we display the seismicity parameters, which form three clusters of time events obtained for seismic events of larger magnitude $4 < m < 6$. Clusters of these events have different structure than in the previous case. They are parallel to X-depth plane. The borders between clusters roughly correspond to the borders of successive showers of the earthquakes. The red cluster comprises only the earthquakes corresponding to the Miyakejima swarm encircled in red in Fig.9B. As we can see by the MDS visualization displayed in Fig. 10A, this cluster is made up from a needle of time events sprouting away from the two remaining oval clusters. The green cluster in Fig.9D represents the deep events, which have a wide variety of magnitudes with high repetition rate and with a rather high degree of spatial non-randomness at short distances (**NS**). As shown in Fig.9, these time events represent mainly the post-swarm series of shocks. As before, the white cluster, includes all the other events.

VII.2 Time period of 150 years

In Fig.11 we display the time series of seismicity parameters computed for the complete synthetic data catalog A. These time series follow the situation from Fig.6, where the dots represent separate data events.

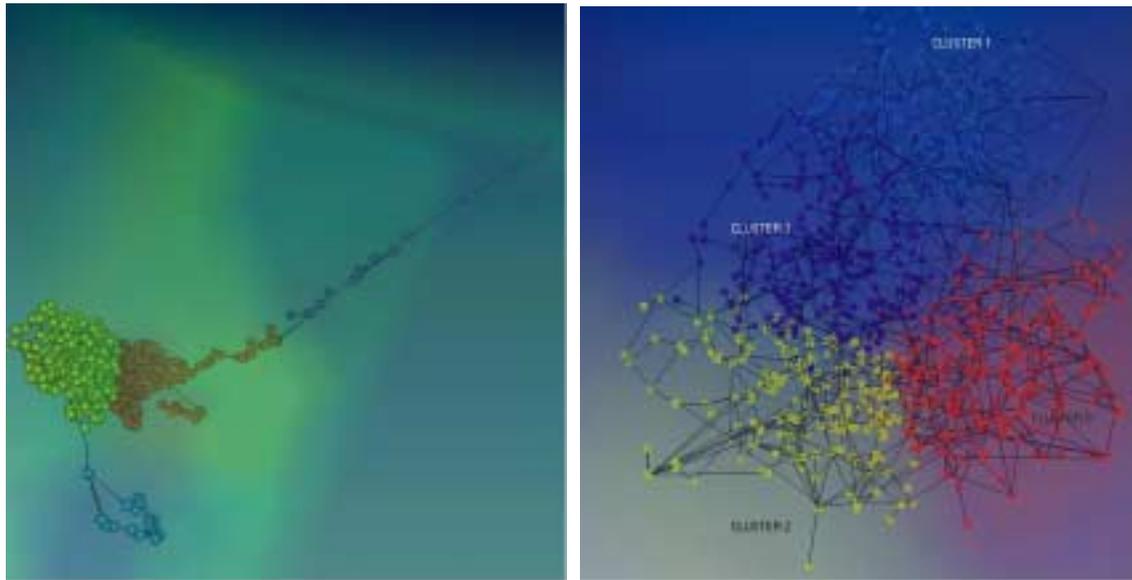


Fig.10 Clusters from feature space mapped into 3D space for A) realistic short-time interval seismic data. The small blue cluster at the bottom represents the events at the end of the time interval, which are averaged within a shrinking time window. B) The synthetic seismic data catalog A covering 150 years.

The green, red and white strips in Fig.11 show 3 clusters of similar time events represented by 7-dimensional feature vectors. In Fig.10B we visualize these clusters due to the MDS transformation of 7-dimensional feature space into 3-D space. In Fig.10B each dot represents a 7-dimensional feature vector mapped into 3-D space by MDS transformation. From the top panel of Fig.11 displaying the largest events \mathbf{M} over a the sliding time window, we may conclude that the

white (blue in Fig.10B) and red clusters from Figs.10B,11 comprise time events, which correspond to the aftershock effects. The white cluster represents the net aftershock events, while the red one includes the earthquake effects averaged over a sliding time window. Conversely, the green cluster (yellow in Fig.10B) contains the time events preceding the earthquakes. It should be clear that the visual clustering of patterns in the feature space is much more reliable and easier to interpret than analyzing arbitrary selected clusters or just thresholding the features, as had been proposed by Keilis-Borok [50,51].

The selectivity in time of the seismicity parameters depends on the width ΔT and shift dt of the sliding time window. Due to space and time averaging, it is impossible to correlate precisely the appearance of an earthquake with the rest of the seismicity parameters when two earthquakes are too close to each other. Therefore, the sequence of green-red-white cluster events can be broken (Fig.11) into time domains with many large earthquakes. As shown in Fig.11, the occurrence of the largest events correlates well with the minima of **NS**, **CD**, **SR**, **TI**, and maxima of **AZ**, **MR** parameters. This means that the occurrence of large earthquakes is preceded by an increasing spatial diffusion of events and greater seismicity rate. Moreover, the results confirm the findings from the actual seismic data over a shorter time-scale:

1. The events preceding large earthquakes are shallow and have small magnitudes. They have also higher degree of spatial repetition than events from different clusters.
2. The earthquakes accompanying and following the mainshock are deep and large in magnitude, have high seismicity rates and low spatial correlation dimension, which drops off rapidly at the onset of large events.

The analysis of synthetic data shows clearly that the clusters in the feature space reflect well the periodicity of both increasing and decreasing seismic activity in a given area. However, at this scale the fine details of precursory and after-shock effects become fuzzy.

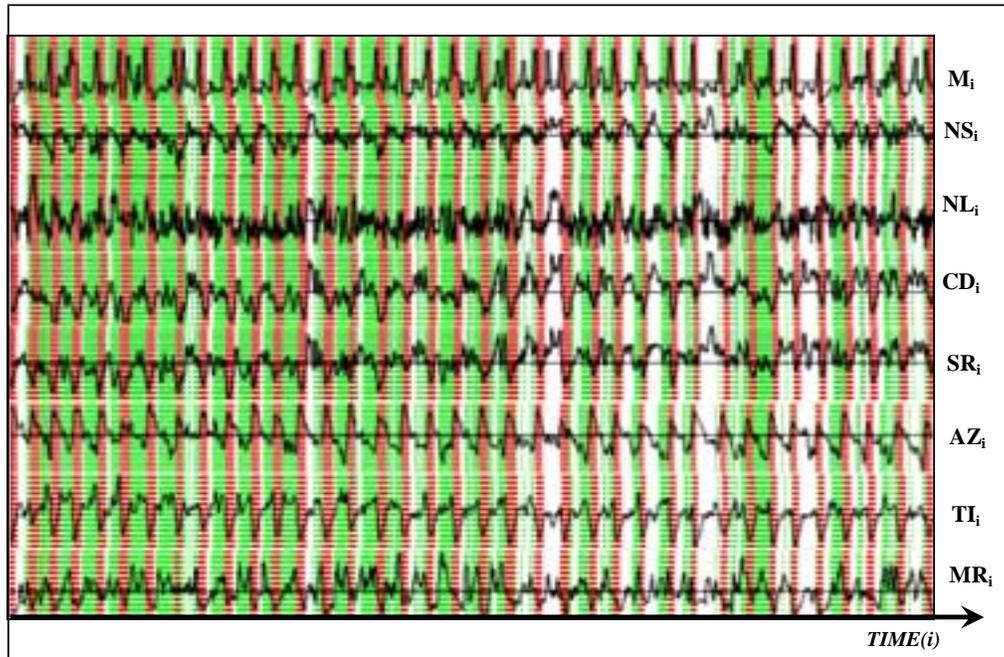


Fig.11 Time history of seismicity parameters $\{M,NS,NL,CD,SR,AZ,TI,MR\}$ for synthetic data catalog A (from [27]).

VII.3 Time period of 1500 years

In Fig.7,8 and Figs.12,13 we visualize the feature vectors for data covering 1500 year period for two models: the A model with a Parkfield-type asperity and the M model with multi-size-heterogeneities. In Fig.8 and Fig.12 one can recognize two types of clusters with different sizes. The larger cluster comprises feature vectors forming approximately 150-year long periodic intervals, which are represented by red strips in Fig.8 and by green strips in Fig.12. The second cluster consists of feature vectors from periodic gaps colored in green in Fig.8 and in white in Fig.12. This anomalous cluster corresponds to the periodic changes in the character of seismic activities. The third cluster (see Fig.13), marked in red for M type of data in Fig.12, consists of regularly periodic and short time intervals representing rapid bursts of seismic activity within every 150-year interval.

In both A and M models the gaps between 150-year long intervals are correlated with decrease of: the correlation dimension (**CD**), degree of spatial repetition (**SR**) and seismicity rate. These gaps are preceded by large earthquakes. The simulations used for generating the datasets incorporate the imposed large earthquakes in regions (I) and (IV) of Fig.3b that bound the computational grid (region II), as staircase boundary conditions with a step at every 150 years. This analysis can detect the effects of these boundary conditions on the seismicity that is calculated in region II.

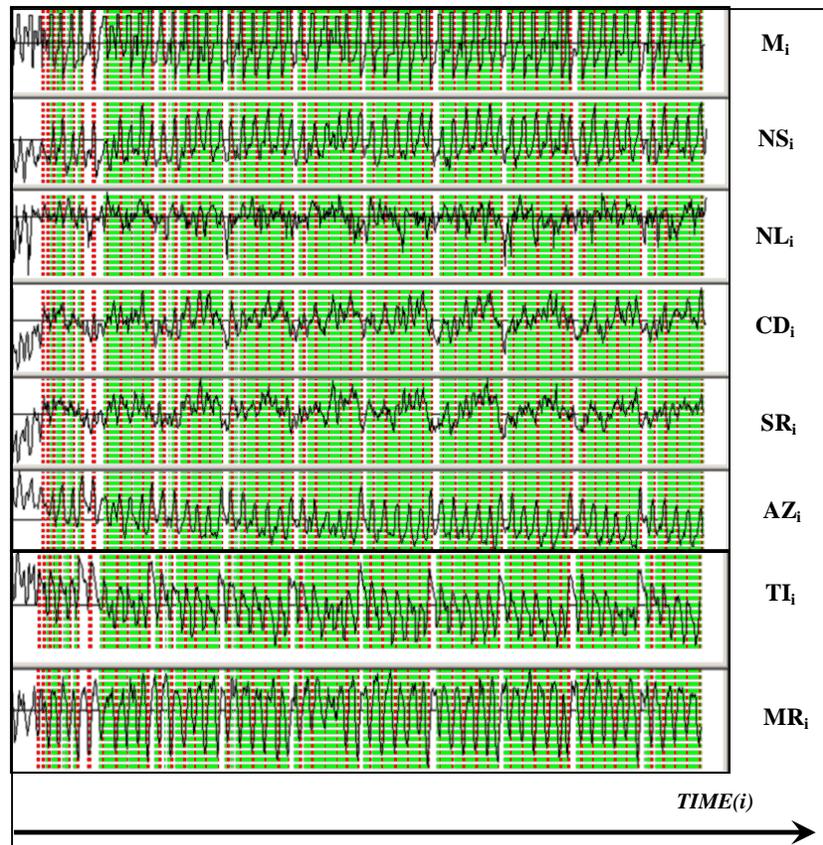


Fig.12 Evolution of the seismicity parameters with time for synthetic (catalog M) for seismic data representing time interval of 1500 years. The red and green strips depict the events belonging to the red and green clusters from Fig.13, respectively.

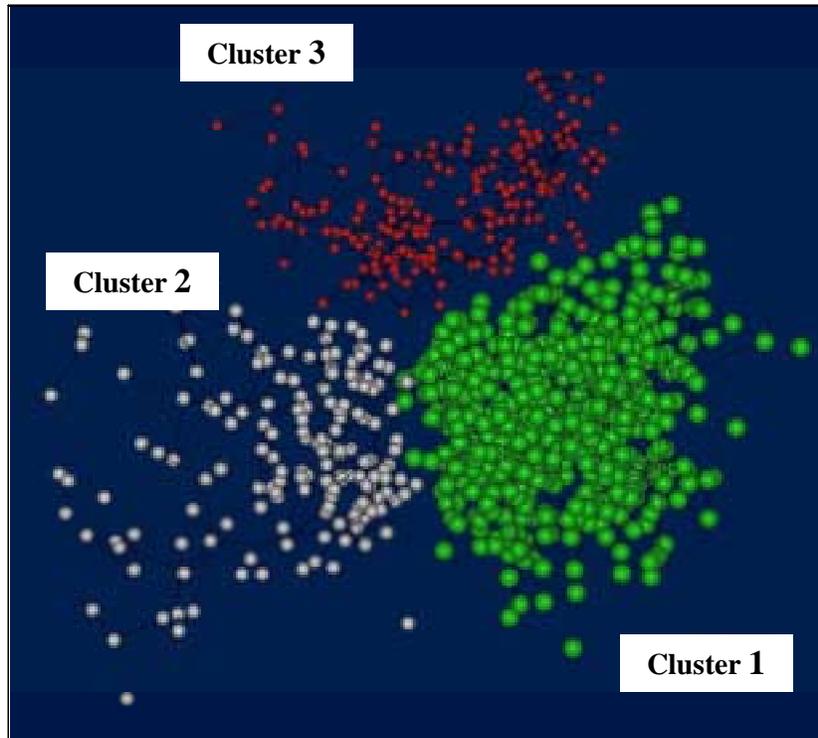


Fig.13 Clusters from Fig.12 mapped by using multidimensional scaling from a 7-dimensional feature space into 3D space for the synthetic seismic data catalog M covering 1500 years.

There are also evident differences between the A and M data in the time intervals belonging to the second cluster. For the A environment the gaps between 150-year intervals are greater and the secondary periodicity within them is less clear. Moreover, within the intervals from the second cluster, the degree of spatial non-randomness decreases at long distances (**NL**) for M model, while for the A data it decreases at short distances (**NS**). In addition, the average depth of earthquakes (**AZ**) is then clearly larger for A model, while for M data it remains at the average level.

We have employed two ways of earthquake visualization. First, the clusters are extracted from the original data from seismic catalogues. Then the results of clustering are mapped back onto the seismic events positions in space and time to visualize the clusters (see Figs.6,7). The second method consists also in two steps. However, now the clusters are detected and visualized in the feature space (Figs.10,13) and then they are mapped back on the set of time series representing the features (Figs.8,9,11,12). By means of analyzing and visualizing earthquake clusters in feature space over a long time-scale, we can investigate important characteristics of seismic activity such as:

1. The occurrence of hierarchical time-periodicity in seismic activity caused by increase of short-time correlations and their decay, respectively. The correlations can be broken due to both short-wave and long-wave resonances of the Poincarè type (e.g. during largest earthquakes) [68].
2. The dependence of seismic activities on the ambient rheological and geological properties of the tectonic environment, which strongly modify the cluster structure of the feature vectors.

VIII Remote Problem Solving Environment (PSE) for Analyzing Earthquake clusters

VIII.1 Remote visualization and analysis of large data sets

We need fast access to large databases in order to forecast earthquakes by observation of similarities between thousands and hundred thousands of seismic events in the course of visualizing earthquake clusters. The largest earthquake catalogs contain TBytes of data. Taking into account also the data from tsunami earthquakes and micro-earthquakes in mines, the total amount of data collected by seismic centers spread all over the world is humongous. Moreover, knowledge extraction of earthquake precursors may demand exploration of cross-correlation relationships involving many variables among many different catalogs. Therefore, both fast communication between data centers and large centralized storage facilities are sorely needed.

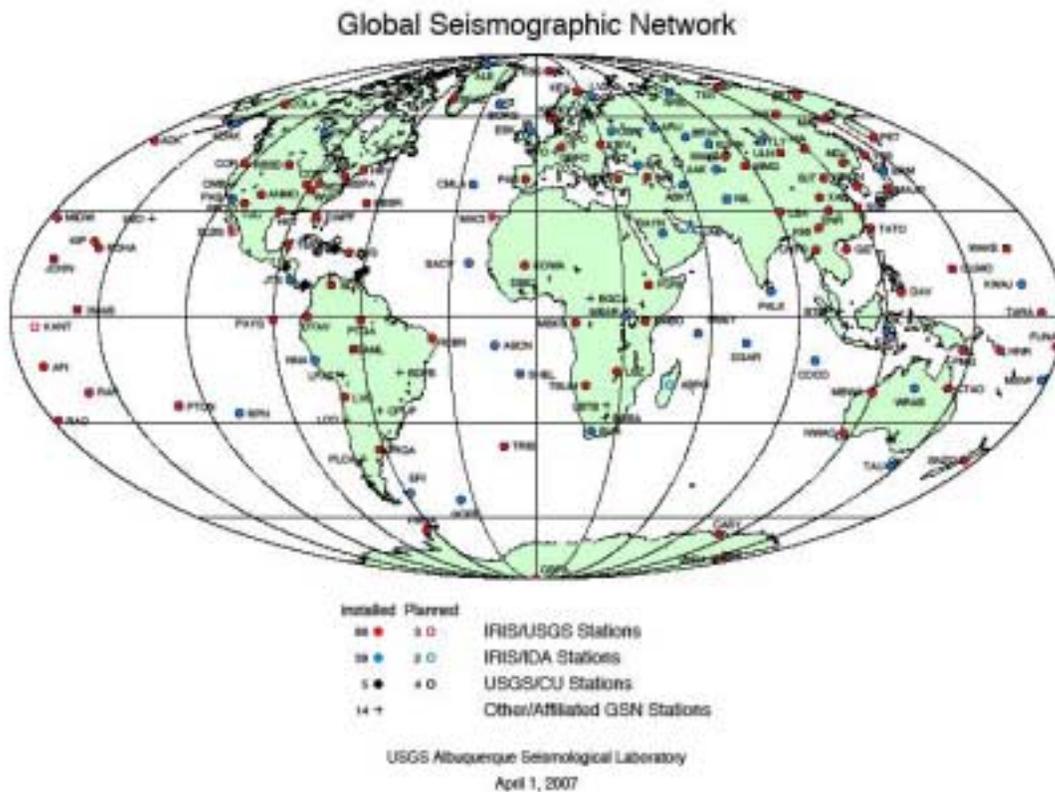


Fig. 14 Worldwide distribution of earthquake seismographic stations (©USGS)

As shown in Fig.14, earthquake seismograph stations, which collect earthquake data from regions with high seismic activity, are distributed worldwide. Therefore, the unprocessed data needs to be stored and then transferred to a dedicated remote server for data processing. After processing, the results must be returned to data acquisition centers and/or other clients. Broadband access to remote facilities dedicated specifically to pattern recognition and visualization allows for scrutinizing local data catalogs by using peer-to-peer connections of data acquisition centers to data preprocessing servers. Clients in the network can automatically compare various types of

earthquake catalogs, data measured in distant geological regions, and the results from various theoretical and computational models. By comparing data accessible in the network we have a chance to eliminate the environmental factors and to extract the resultant earthquake precursory effects.

Integration of a variety of hardware, operating systems, and their proper configuration results in many communication problems between data centers. Efficient, reliable, and secure integration of distributed data and software resources, such as pattern recognition and visualization packages, is possible only within the GRID paradigm of computing [32,93,96]. The GRID mode of computing has flourished rapidly in recent years and has facilitated collaboration and accessibility to many types of resources, such as large data sets, visualization servers and compute engines. Scientific teams have developed easy-to-use, flexible, generic and modular middleware, enabling today's applications to make innovative use of global computing resources. Remote access tools were also produced to visualize huge datasets and monitor performance and data analysis properties, effectively steering the data processing procedures interactively [74]. The TeraGrid project, serving over 10,000 customers in 2007 (<http://www.teragrid.org>) is a highly successful high-performance implementation of such a GRID infrastructure and is being used as an integrated, persistent computational resource at universities and laboratories across the USA. The Teragrid development impacts also the earthquake science [7,17]. The National Science Foundation has awarded the Southern California Earthquake Center 15 million hours of computer time on the Teragrid [*Grid Today*, August 2007]. These computational resources will be used for simulating thousands of possible earthquakes scenarios in Southern California, including the largest breaks on the San Andreas fault (www.scec.org/cybershake). SCEC will be able to simulate the most disastrous earthquakes ($M > 7$), such as events that could produce Sumatran disasters.

We discuss the idea of an integrated problem-solving environment (PSE) intended for the analysis of earthquake clusters for the forecasting earthquakes. A simplified scheme for data acquisition and visualization of earthquake clusters is displayed in Fig.15. This system promotes portability, dynamic results on-demand, and collaboration among researchers separated by long distances by using a client-server paradigm. This is provided through a lightweight front-end interface for users to run locally, while the a remote server takes care of intensive processing tasks on large databases, off-screen rendering, and data visualization.

VIII.2 Grid Environment

Today large datasets and high-performance computing resources are distributed across the world. When collaboration and sharing of resources are required, a computational GRID infrastructure needs to be in place to connect up these servers (see e.g., [17,32,93,96]). There must exist protocols available to allow clients to tap into these resources and harness their power. The computational grid can be seen as a distributed system of "clients", which consists of either "users" or "resources" and proxies. A GRID can be implemented using an event brokering system designed to run on a large network of brokering nodes. Individually, these brokering nodes are competent servers, but when connected to the brokering system, they are able to share the weight of client requests in a powerful and efficient manner. Examples of this include GRID Resource Brokering [95], NaradaBrokering, GLOBUS [95], and VLab developed at the University of Minnesota (see <http://vlab.msi.umn.edu>) [47,74].

These GRID architectures are well suited to function as a PSE for earthquake cluster analysis and as an integrated computational environment for data exchange. The seismic data centers from the networking point of view represent a complex hierarchical cluster structure. They are located geographically in the regions of high seismic activity within heavily populated areas of economic importance. Therefore, the seismic data centers create distant superclusters of

various "density" of computational resources corresponding to the size and importance of the regions. These superclusters are sparse in the sense of computational resources devoted for earthquake detection and data acquisition. However, these same structures contain important computational, scientific and visualization facilities with strong interest in the analysis of earthquake data and earthquake modeling. The efficient interconnection of these sites is of principal interest. Due to the "small world network" structure [6,57] of GRID architectures, it is possible to select the most efficient routing schemes, considerably shortening the average communication path length between brokers. GRID architectures are appropriate to link the clients, both users and resources, together. Construction of efficient and user friendly Problem Solving Environments (PSE) requires integration of data analysis and visualization software within a GRID environment, in such a way that it can be easily accessed over the Internet. We have created an integrated data interrogation toolkit to act as a PSE for visualization and clustering of earthquake data, which we called WEB-IS [46].

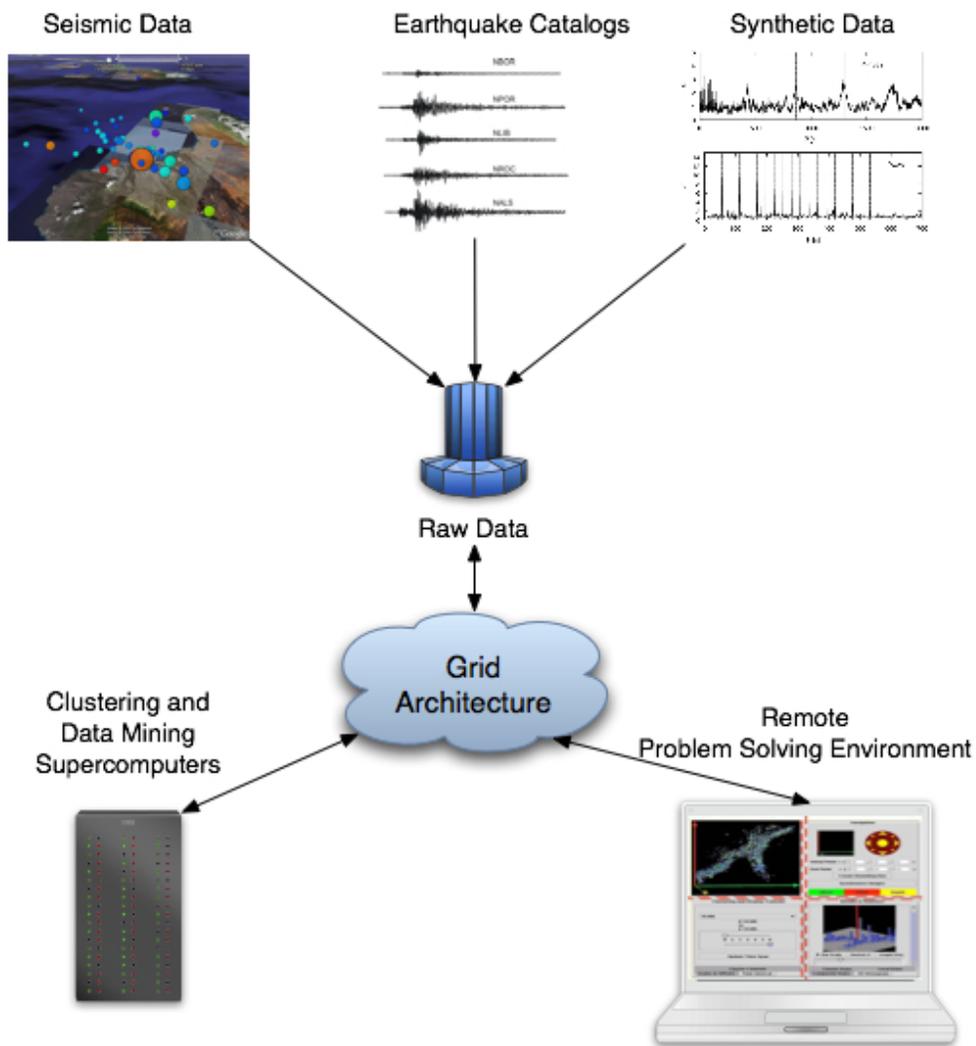


Fig.15 Data acquisition, storage, processing, and remote problem solving environments in earthquake clustering.

VIII.3 Example of Remote PSE

WEB-IS is a recently developed software tool that allows remote, interactive visualization and analysis of large-scale 3-D earthquake clusters over the Internet [84,85] through the interaction between client and server. WEB-IS acts as a PSE through a web portal used to solve problems by visualizing and analyzing geophysical datasets, without requiring a full understanding of the underlying details in software, hardware and communication [35,46]. As shown in Fig.16, the primary goal of WEB-IS in the geosciences is to provide middleware that sits between the modeling, data analysis tools and the display systems that local or remote users access. In the case of large and physically distributed datasets, it is necessary to perform some preprocessing and then transmit a subset of the data to one or more processes or visualization servers to display. The details of where and how the data migrate should be transparent to the user. WEB-IS makes available to the end users the capability of interactively exploring their data, even though they may not have the necessary resources such as sufficient software, hardware or datasets at their local sites. This method of visualization offers to the user a workflow to navigate through their rendered 3-D data [46,47] and analyze for statistics or apply earthquake cluster analysis. To the client, the process of accessing and manipulating the data appears simple and robust, while the middleware takes care of the network communication, computer security and data preparation.

Complete realization of an earthquake clustering PSE consists of:

1. Data analysis tools to implement earthquake clustering techniques;
2. High performance visualization techniques using OpenGL, scripting language such as Tcl, or Amira package;
3. Coupling the workflow to a Grid environment;
4. Integration of the visualization toolkit, such as WEB-IS.

These exist and can work both independently and coupled in a single special purpose system. This system can be developed creating the backbone of the sophisticated computational data acquisition environment, which can be devised specifically for earthquake clustering or for general needs of the geophysical community. Equipped with only hand-held computers or laptops, and working on location in unreachable desert terrains with remote data acquisition centers or perhaps just analyzing data in one of the many computation facilities located around the globe, geophysicists will have readily accessible gate to data resources spread all over the world.

We see the principal goal of our work in contributing to the construction of a global warning system, which can be used for prediction of catastrophes such as various types of earthquakes along the circum Pacific belt, where there is a great concentration of people along the coastal region. For example, similar methodology can be used for tsunami earthquake alerting. Theoretical models of faulting and seismic wave propagation used for the computation of radiated seismic energy from broad-band records at teleseismic distances [15] can be adapted to the real-time situation when neither the depth nor the focal geometry of the source is known accurately. The distance-dependent approximation was used in [56]. By analyzing important geophysical parameters (such as the energy-to moment ratio H [56] for normal earthquakes, the results obtained from the theoretical models agree well with values computed from available source parameters (e.g., as published by the National Earthquake Information Center). It appears however that the so called "tsunami earthquakes" - characterized by the significant deficiency of moment release at high frequencies—yield the values of H considerably different than regular earthquakes. Thus H value can be used as a suitable criterion for discriminating various types of earthquakes in a short duration of time, like an hour. However, this hypothesis holds only for a few cases. For, the so-called, "tsunamigenic earthquakes", this difference is not so clear.

Moreover, the value of the moment computed on the base of long-period seismic waves can be underestimated. For example, analysis of the longest period normal modes of the Earth, ${}_0S_2$ and ${}_0S_3$, excited by the December 26, 2004 Sumatra earthquake [76], yields a seismic moment of 1.3×10^{30} dyn-cm, approximately three times larger than the 4×10^{29} dyn-cm measured from long-period surface waves. Therefore, instead of a single-value discrimination we recommend using more parameters (dimensions) for detecting tsunami earthquakes. As shown in [61] and [82], one could employ other T-phase characteristics such as its duration, seismic moment, and spectral strength or even similar features associated with the S-phase. Progress in prediction efforts is hindered, somewhat by lack of communication between researchers from different fields and fast access to the various types of data. Therefore, we hope that globalization of computation, data acquisition and visualization resources, together with fast access through a scale-free network, will provide an important solution to this critical problem.

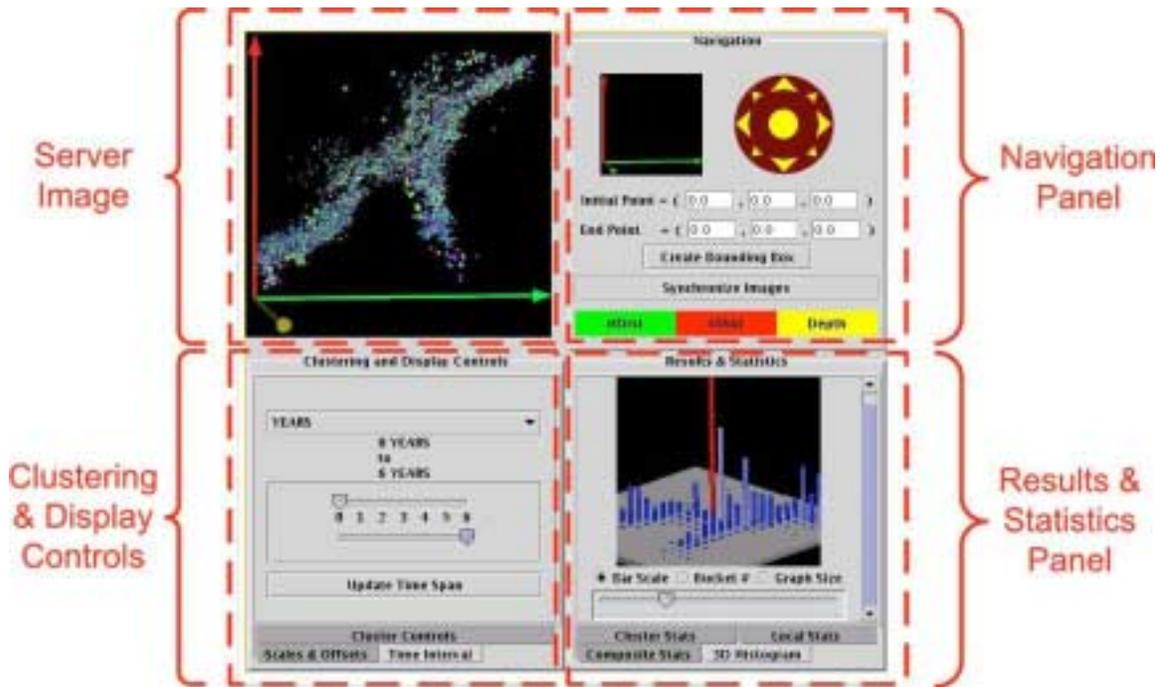


Fig.16 WEB-IS is an example of a PSE devoted to remote earthquake clustering.

IX Future Directions

In this chapter we endeavor to bring across the basic concept of clustering and its role in visualization of earthquake statistics. Indeed we find that the clustering of seismic activities reflects both the similarity associated with them and their correlation properties. As discussed in, e.g., Ben-Zion et al. [14], Saichev and Sornette [68] and Zöller et al. [89], there exists an evolutionary process or memory between successive earthquakes, which impact the distribution of the inter-event times. We believe that by means of earthquake clustering we can capture the essence of this predictive information [29]. Therefore, in order to carry out real-time earthquake forecasting for short-time scales, it is necessary to derive a thorough understanding of all families of earthquake clusters produced over an earthquake-prone region.

We stress that in obtaining this type of information one must first be able to detect the precise location of the significant clusters by filtering out simultaneously the noise and the

outliers, and finally to visualize these objects seamlessly. While the existence of spatial-temporal clusters is important, they do not reveal the subtle information hidden behind the relationships among the data events, such as: spatial-temporal correlation dimensions, correspondence between the numbers of small and large magnitude events, degree of spatial randomness, repetitiveness at different distances and other factors. The features - "descriptors" or seismicity parameters - constructed from the empirical knowledge of the researcher should be largely independent and should represent aptly distinctive features, which are useful for the purpose of pattern recognition. Unlike single events described only by spatio-temporal features (and magnitude), the N -dimensional feature vectors can portray better the dynamics of the seismically active area in different moments of time. By following the basic rules of learning theory [111], we may be able to arrive at the number N and quality of features, which can assure the generalization power of the data and allow us to construct reliable data-models or classifiers.

We have shown that clustering, as a well-honed tool in data mining and pattern recognition, represents the classifier without the teacher, which means that the nature of the clustering is unknown and its exact background must be guessed at from expert knowledge and analysis of the cluster properties. Clustering is a process based on *a priori* knowledge extraction for constructing the hypothesis space needed for reliable classifiers that can be taught and used for forecasting [27]. However, the quality of these data models depends strongly on the quality of hypothesis space constructed. Consequently, it depends on the quality of cluster extraction. The major problem comes from the lack of a universal clustering scheme, thus making the clustering process somewhat subjective. In this case we must visualize the multidimensional feature space. Visual confirmation lends great confidence on the validity of the clustering process. We can then adjust for the optimal clustering procedures by removing the noise and outliers. Among the major goals of earthquake clustering analysis, we can list the following salient points:

- classification of the branching properties of seismicity patterns [97], for example to recognize the three main groups of shocks: foreshocks, mainshocks and aftershocks or to remove the temporary clustering for estimating the background seismicity;
- understanding the correlations between observed properties of earthquakes in different domains (e.g., space, time, number, size);
- understanding the relations between various physical parameters of the models and properties of the generated earthquakes;
- investigating the multi-scale nature of the cluster structure and reconstructing any hidden information.

Classification of type of shocks remains an outstanding problem because there are no observable differences between foreshocks, main shocks and aftershocks [68]. Each earthquake is able of triggering other earthquakes according to the basic laws from Section 2.1 [42,67]. In spite of this difficulty, as shown in [14, 27, 29, 31, 50], it is possible to construct some sort of stochastic classifiers based on theoretical footing. The method proposed here is closely related to the epidemic-type aftershock sequence (ETAS) model [58]. It is important that the principal characteristics of ETAS-based models correspond to experimental verifications, i.e., they treat all earthquakes on the same footing and there is no marked distinction among foreshocks, main shocks and aftershocks. The key points of the method are the probabilities of one event being triggered by a previous event (e.g., [81]). Making use of these probabilities, we can reconstruct the functions associated with the characteristics of earthquake clusters in order to test a number of plausible hypotheses about earthquake clustering phenomena.

As shown above by our results on seismicity clustering for the three different time epochs, clustering can be truly regarded as a coarse-graining procedure. We can see the lucid details only after the smaller scales have been erased, thereby exposing the general trends associated with the long correlation length. For large data bases covering long time intervals we can unveil the

shorter timescale characteristics by removing the background events, using successive clustering. Eventually, we can build up the strong classifiers. In the case where the long-time data catalogs are missing, we can employ the stochastic classifiers advocated Ben-Zion et al. [14] for prior thresholding of the background data or what is sometimes called "fuzzification" [114]. By following this procedure we can construct the hypothesis space for data models by clustering (or fuzzy clustering) procedures.

In this chapter we contribute to the development of improved software infrastructure and information technology for visual analysis of seismicity. A combined clustering analysis of observed and synthetic data, aided by state-of-the-art visualization of multi-dimensional clusters, will lead to improved earthquake forecasting algorithms with shorter time windows of increased probability of large seismic events. We emphasize the vital role played by information technology constructed around a GRID infrastructure in visually investigating earthquakes. Grid services, due to their role in integration of computational resources, enable on-line access to distributed seismic data, both on-field measurements and seismic catalogs. Simultaneously, they allow for instant access to the modeling and data mining tools, computational power and visualization facilities.

Acknowledgments

This research was supported by NSF ITR and Math-Geo grants. WD acknowledges support from the Polish Ministry of Science and Higher Education Grant No.3T11C05926. YBZ acknowledges support from the NSF, USGS and SCEC. We thank Dr. David Vere-Jones for an elucidating review and Dr. Willie K. Lee for constant encouragement. We thank Hanjing Hong and Roy Liu for discussion of the Changbaishan seismicity.

X References

Primary Literature

- [1] van Aalsburg, J., Grant, L.B., Yakovlev, G., Rundle, P.B., Rundle, J.B., Turcotte, D.L. and A. Donnellan, A feasibility study of data assimilation in numerical simulations of earthquake fault systems, *Phys. Earth Planet Inter.*, 163, 149-162, 2007.
- [2] Amira visualization package, <http://www.amiravis.com>
- [3] Ammon, C., J., Kanamori, H., Lay, T., A great earthquake doublet and seismic stress transfer cycle in the central Kuril islands, *Nature*, 451/31, 561-566, 2008.
- [4] Bak, P., Tang, C., Earthquakes as a self-organized critical phenomena, *J. Geoph. Res.*, 94/B11, 15,635–15,637, 1989.
- [5] Bak, P., K. Christensen, L. Danon, and T. Scanlon (2002), Unified scaling law for earthquakes, *Phys. Rev. Lett.*, 88, 178501, 2002.
- [6] Barabasi AL, Jeong H, Neda Z, Ravasz E, Schubert A, Vicsek T, Evolution of the social network of scientific collaborations, *Physica A*, 311 (3-4), 590-614, 2002.

- [7] Beckman, P.,H., Building the TeraGrid, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 363/1833, 2005.
- [8] Ben-Zion, Y. and J. R. Rice, Earthquake failure sequences along a cellular fault zone in a three-dimensional elastic solid containing asperity and nonasperity regions, *J. Geophys. Res.*, 98, 14109-14131, 1993.
- [9] Ben-Zion, Y. and J. R. Rice, Slip patterns and earthquake populations along different classes of faults in elastic solids, *J. Geophys. Res.*, 100, 12959-12983, 1995.
- [10] Ben-Zion, Y., Stress, slip and earthquakes in models of complex single-fault systems incorporating brittle and creep deformations, *J. Geophys. Res.*, 101, 5677-5706, 1996.
- [11] Ben-Zion, Y., Dynamic Rupture in Recent Models of Earthquake Faults, *J. Mech. Phys. Solids*, 49, 2209-2244, 2001.
- [12] Ben-Zion, Y., Appendix 2, Key Formulas in Earthquake Seismology, in *International Handbook of Earthquake and Engineering Seismology*, eds. W. HK Lee, H. Kanamori, P. C. Jennings, and C. Kisslinger, *Part B*, 1857-1875, Academic Press, 2003.
- [13] Ben-Zion, Y., Sammis, C.,G., Characterization of Fault Zones , *Pure Appl. Geophys.*, 160, 677-715, 2003.
- [14] Ben-Zion Y., Eneva, M., Liu, Y., Large earthquake cycles and intermittent criticality on heterogeneous faults due to evolving stress and seismicity, *J. Geoph. Res.*, 108/B6, 2307-27, 2003.
- [15] Boatwright J, Choy G L (1986) Teleseismic estimates of the energy radiated by shallow earthquakes, *J Geophys Res*, 91:2,095–2,112
- [16] Bollig, E.F., Jensen, P. A., Lyness, M. D., Nacar, M. A., da Silveira, P. R., Erlebacher, G., Pierce, M., Yuen, D. A., VLAB: Web Services, Portlets, and Workflows for Enabling Cyber infrastructure in Computational Mineral Physics 2007, *Journal of Physics of the Earth and Planetary Interiors (PEPI)*, 163,333-346, 2007.
- [17] Catlett, C., E., TeraGrid: A Foundation for US Cyberinfrastructure, *Lecture Notes in Computer Science*, 3779/2005, 2005.
- [18] Chen, C-C., Rundle, J.B., Li, H.-C., Holliday, J.R., Turcotte, D.L., and K. F. Tiampo,, Critical point theory of earthquakes: Observations of correlated and cooperative behavior on earthquake fault systems, *Geophys. Res. Lett.*, L18302, 2006.
- [19] Chinnery, M., The stress changes that accompany strike-slip faulting, *Bull. Seismol. Soc. Am.*, 53, 921–932, 1963.
- [20] Corral, A., Mixing of rescaled data and Bayesian inference for earthquake recurrence times, *Nonlinear Processes Geophys.*, 12, 89– 100, 2005.
- [21] Corral, A., Renormalization-group transformations and correlations of seismicity, *Phys. Rev. Lett.*, 95, 028501, 2005.

- [22] Corral, A., and K. Christensen, Comment on ‘Earthquakes descaled: On waiting time distributions and scaling laws, *Phys. Rev. Lett.*, 96, 109801, 2006.
- [23] Dahmen, K. and Y. Ben-Zion, The physics of jerky motion in slowly driven magnetic and earthquake fault systems, submitted to *Encyclopedia of Complexity and System Science*, 2008.
- [24] Davy, P., A. Sornette, and D. Sornette, Some consequences of a proposed fractal nature of continental faulting, *Nature*, 348, 56–58, 1990.
- [25] Dzwiniel W, Blasiak J, Method of particles in visual clustering of multi-dimensional and large data sets, *Future Generation Computers Systems*, 15, 365-379, 1999.
- [26] Dzwiniel W, Yuen D.A., Kaneko Y, Boryczko K, Ben-Zion Y, Multi-Resolution Clustering Analysis and 3-D Visualization of Multitudinous Synthetic Earthquakes, *Visual Geosciences*, 8, 12-25, 2003
- [27] Dzwiniel, W., Yuen, D.A., Boryczko, K., Ben-Zion, Y., Yoshioka, S., Ito, T., Nonlinear multidimensional scaling and visualization of earthquake clusters over space, time and feature space, *Nonlinear Processes in Geophysics*, 12, 117-128, 2005.
- [28] Eisen M.B., Spellman P.T., Brown P.O., & Botstein D, Cluster analysis and display of genome-wide expression patterns. *Proc Natl. Acad Sci U S A* 95, 14863–14868, 1998.
- [29] Enescu, B., Ito, K. and Struzik, Z.R., Wavelet-based multiscale analysis of real and simulated time-series of earthquakes, *Geophys. J. Int.* 164, 63-74, 2006.
- [30] Eneva, M., Ben-Zion, Y., Techniques and parameters to analyze seismicity patterns associated with large earthquakes, *Journal of Geophysical Research*, 102/B8, 785-795, 1997a
- [31] Eneva, M. and Ben-Zion Y., Application of pattern recognition techniques to earthquake catalogs generated by models of segmented fault systems in three-dimensional elastic solids, *J. Geophys. Res.*, 102, 24513-24528, 1997b.
- [32] Foster, I., The Grid: A new infrastructure for 21st century science, *Physics Today*, 55/2, 42-47, 2002.
- [33] Freed, A.M. and Lin, J., Delayed triggering of the 1999 Hector Mine earthquake by viscoelastic stress transfer, *Nature*, 411, 180-183, 2001.
- [34] Frey, B.J., Dueck, D., Clustering by Passing Messages Between Data Points, *Science*, 315, 5814, 972 – 976, 2007.
- [35] Garbow Z.A., Erlebacher G., Yuen D.A., Sevre E.O., Nagle A.R., Kaneko Y., Web-Based Interrogation of Large-Scale Geophysical Datasets and Clustering Analysis of Many Earthquake Events from Desktop and Handheld Devices, *American Geophysical Union Fall Meeting*, Abstract , 2002
- [36] Gowda C.K., Krishna G., Agglomerative clustering using the concept of nearest neighborhood, *Pattern Recogn.*, 10, 105, 1978.

- [37] Guha, S., Rastogi, R., and Shim, K., *CURE: An efficient algorithm for large databases*, in Proceedings of SIGMOD '98, Seattle, June 1998, pp. 73-84.
- [38] Gutenberg, B., Earthquake magnitude, intensity, energy and acceleration. *Bull. Seism. Soc. Am.*, 32, 163-191, 1942.
- [39] Hainzl, S.F. Scherbaum and C. Beauval, Estimating Background Activity Based on Interevent-Time Distribution, *Bulletin of the Seismological Society of America*, 96, 313–320, 2006.
- [40] Harte, D., Li D-F., Vreede, M., Vere-Jones, D., Quantifying the M8 prediction algorithm: reduction to a single critical variable and stability results, *New Zealand Journal of Geology and Geophysics*, 46, 141–152, 2003.
- [41] Helmstetter, A., Sornette, D., and J.-R. Grasso, Mainshocks are aftershocks of conditional foreshocks: How do foreshock statistical properties emerge from aftershock laws, *J. Geophys. Res.*, 108, 2046, 2003.
- [42] Helmstetter A., Kagan Y.,Y., Jackson D.,D., Importance of small earthquakes for stress transfers and earthquake triggering, *J. Geophys. Res.*, 110, B05S08, 2005.
- [43] Hong H., Kadlec, D.J., Yuen, D.A., Zheng, Y., Zhang, H., Liu, G., Dzwinel, W., Fast timescale phenomena at Changbaishan volcano as inferred from recent seismic activity, *Eos Trans. AGU*, Fall Meet., 85(47), 2004.
- [44] Ismail, M.A, Kamel M.S., Multi-dimensional data clustering utilizing hybrid search strategies, *Pattern Recognition*, 22/1, 77-89, 1989.
- [45] Ito T., Yoshioka, S., A dike intrusion model in and around Miyakejima, Niijima and Kozushima. *Tectonophysics* 359: 171–187, 2002.
- [46] Kadlec BJ, Yang XL, Wang Y, Bollig EF, Garbow ZA, Yuen DA, Erlebacher G WEB-IS (Integrated System): *An Overall View*, *Eos. Trans. AGU*, 84(46), Fall Meet. Suppl., Abstract NG11A-0163, 2003.
- [47] Kadlec, B.J., Yuen D.,A., and Erlebacher, G., Visualization and Analysis of Multi-terabyte Geophysical Datasets in an Interactive Setting With Remote Webcam Capabilities, *Pure and Applied Geophysics*, 163/(11-12), 2455-2465, 2006.
- [48] Karypis G., Han, E., Kumar V., Chameleon: A hierarchical clustering algorithms using dynamic modeling, *IEEE Computer*, 32(8), 68-75, August 1999.
- [49] Karypis, G. Aggarwal, R. Kumar, V. Shekhar, S., Multilevel hypergraph partitioning: applications in VLSI domain, *IEEE Transactions on Very Large Scale Systems Integration (VLSI)*, 7/1, 69-79, 1999.
- [50] Keilis-Borok, V.I., Kossobokov, V.B., Premonitory activation of earthquake flow, *Physics of the Earth and Planetary Interiors*, 61, 73-83, 1990.

- [51] Keilis-Borok, V.I., Earthquake prediction: State-of-the-Art and Emerging Possibilities, *Annual Rev. of Earth and Planetary Sci.*, 30, 1-33, 2002.
- [52] Lee, J.,A., Lendasse A., Verleysen M., Nonlinear projection with curvilinear distances: Isomap versus curvilinear distance analysis. *Neurocomputing*, 57, 49-76, 2004.
- [53] Mehta, A.P., K. A. Dahmen and Y. Ben-Zion, Universal mean moment rate profiles of earthquake ruptures, *Phys. Rev. E.*, **73**, 056104, 2006.
- [54] Molchan, G.M., Interevent time distribution of seismicity: A theoretical approach, *Pure Appl. Geophys.*, 162, 1135– 1150, 2005.
- [55] Nacar, M., Aktas, M., Pierce, M., Lu, Z., Erlebacher, G., Kigelman, D., Bollig, E.F., de Silva, C., Sowell, B. and Yuen, D.A., VLab: Collaborative Grid Services and Portals to Support Computational Material Science, *Concurrency and Computation: Practice and Experience* 19/12, 1717-1728, 2007.
- [56] Newman A.V., Okal E.A., Teleseismic estimates of radiated seismic energy: the S/M0 discriminant for tsunami earthquakes, *J Geophys Res* 103/B11:23,885–23,898, 1998.
- [57] Newman M.E.J., The structure and function of complex networks, *SIAM Review*, 45/2, 167-256, 2003.
- [58] Ogata, Y., Seismicity analysis through point-process modeling: A review, *Pure Appl. Geophys.*, 155, 471-507, 1999.
- [59] Ogata, Y., and Zhuang, J., Space-time ETAS models and an improved extension, *Tectonophysics*, 413, 13–23, 2006.
- [60] Okada Y., Internal deformation due to shear and tensile faults in a half space. *Bull.Seism. Soc. Am.* 82, 1018-1040, 1992.
- [61] Okal E.A., Alasset P-J, Hyvernaud O., Schindele' F., The deficient T waves of tsunami earthquakes, *Geophys J Int*, 152:416–432, 2003.
- [62] Ouillon, G., Ducorbier, C., and Sornette, D., Automatic Reconstruction of Fault Networks from Seismicity Catalogs: 3D Optimal Anisotropic Dynamic Clustering, in press in *J. Geophys. Res.* 2007 (<http://arxiv.org/abs/physics/0703084>)
- [63] Roweis, S.T., Saul, L., K., Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290, 2323-2326, 2000.
- [64] Rundle J.B., Gross S, Klein W, Ferguson C, Turcotte DL., The statistical mechanics of earthquakes, *Tectonophysics*, 277, 147- 164, 1997.
- [65] Rundle, J.B., Klein, W., Gross, S., Physical Basis for Statistical Patterns in Complex Earthquake Populations: Models, Predictions and Tests, *Pure and Applied Geophysics*, 155/2-4, 575-607, 1999.
- [66] Rundle J.B., Klein W, Tiampo K, Gross S, Linear pattern dynamics in nonlinear threshold systems, *Phys. Rev. E*, 61/3, 2418-2143, 2000a.

- [67] Saichev, A.,A., Helmstetter, A., and Sornette, D., Power Law Distributions of Offspring and Generation Numbers in Branching Models of Earthquake Triggering, *Pure. Appl. Geophys.*, 162, 1113-1134, 2005.
- [68] Saichev, A., Sornette, D., Theory of earthquake recurrence times, *J. Geoph. Res.*, 112/B4, 1-26, 2007.
- [69] Sander, J., Ester M., Krieger H, Density based clustering in spatial databases, The algorithm DBSCAN and its applications, *Data Mining and Knowledge Discovery*, 2(2), 169-194, 1998.
- [70] Shcherbakov, R. and Turcotte, D.L., A modified form of Bath's law, *Bull. Seismol., Soc. America*, 94, 1968-1975, 2004.
- [71] Shcherbakov, R., Turcotte D.L., Rundle J.,B., Aftershock statistics, *Pure and Applied geophysics*, 162 (6-7), 1051-1076, 2005.
- [72] Shokri S.,Z., Ismail, M.,A., K-means-type algorithms: A generalized convergence theorem and characterization of local optimality, *IEEE Trans. Pattern Anal. Mach. Intell.* 6, 81-87, 1984.
- [73] Siedlecki W., Siedlecka K., Sklanski, J, An overview of mapping for exploratory pattern analysis, *Pattern Recogn.*, 21/5: 411-430, 1988.
- [74] da Silva, C.R.S., da Silveira, P.R.C., Karki, B., Wentzcovitch, R.M., Jensen, P.A., Bollig, E.F., Pierce, M., Erlebacher, G., and D.A. Yuen, Virtual laboratory for planetary materials: System service architecture overview, *Phys. Earth Planet.Inter.*, 163, 323-332, 2007.
- [75] Sornette, D., Johansen, A., and Bauchaud, J.-P., Stock market crashes, precursors and replicas, *J. Phys. I Finance*, 5, 167-175, 1996.
- [76] Stein S, Okal E Ultra-long period seismic moment of the great December 26, 2004 Sumatra earthquake and implications for the slip process, <http://www.earth.nwu.edu/people/emile/research/Sumatra.pdf>, 2004.
- [77] Tenenbaum, J.B., de Silva, V., Langford, J.C.A., Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290, 2319-2323, 2000.
- [78] Utsu, T., Statistical Features of Seismicity, in *International Handbook of Earthquake and Engineering Seismology*, eds. W. HK Lee, H. Kanamori, P. C. Jennings, and C. Kisslinger, pp. 719-732, Part A, 2002.
- [79] Varini C., Degenhard A., Nattkemper, T., W., ISOLLE: LLE with geodesic distance, *Neurocomputing*, 69, 1768-1771, 2006.
- [80] Venna, J., Kaski S., Local multidimensional scaling, *Neural Networks*, 19, 889-899, 2006.
- [81] Vere-Jones, D., A branching model for crack propagation: *Pure and Appl. Geophys.*, 114, no.4, p. 711-726, 1976.

- [82] Walker D.A., Mc Creery C.S., Hiyoshi Y., T-phase spectra, seismic moment and tsunamigenesis, *Bull Seism Soc Am* 82:1,275–1, 305, 1992.
- [83] Wesnousky, S. G., The Gutenberg-Richter or characteristic earthquake distribution, which is it?, *Bull. Seismol. Soc. Amer.*, 84, 1940-1959, 1994.
- [84] Yuen DA, Garbow ZA, Erlebacher G., Remote data analysis, Visualization and Problem Solving Environment (PSE) Based on Wavelet Analysis in the Geosciences, *Visual Geosciences* 8:83–92, 2004.
- [85] Yuen, D.A., Kadlec, B.J., Bollig, E.F., Dzwinel, W., Garbow, Z.A., da Silva, C., Clustering and Visualization of Earthquake Data in a Grid Environment, *Visual Geosciences*, 10/1, 2005.
- [86] Zhang Q. and Boyle R., A new clustering algorithm with multiple runs of iterative procedures, *Pattern Recognition*, 24/9, 835-848, 1991.
- [87] Zöller, G., S. Hainzl, Y. Ben-Zion, and M. Holschneider, Earthquake activity related to seismic cycles in a model for a heterogeneous strike-slip fault, *Tectonophysics*, 423, 137-145, 2006.
- [88] Zöller, G., Y. Ben-Zion and M. Holschneider, Estimating recurrence times and seismic hazard of large earthquakes on an individual fault, *Geophys. J. Int.*, 170, 1300–1310, 2007.
- [89] Zöller, G., S. Hainzl, Y. Ben-Zion and M. Holschneider, Critical states of seismicity: From models to practical seismic hazard estimates, submitted to *Encyclopedia of Complexity and System Science*, 2008.

Books, Reports and Reviews

- [90] Living on an active Earth, Perspectives on Earthquake Sciences, National Research Council, *The National Academies Press*, Washington D.C., 2003.
- [91] Andenberg M.R, *Clusters Analysis for Applications*, New York, Academic Press, 1973.
- [92] Bennett, A.F., *Inverse Methods in Physical Oceanography*, 346 pp., Cambridge University Press, Cambridge, 1992.
- [93] Berman, F., Fox, G.C., Hey, A.J.G., *Grid Computing - Making the Global Infrastructure a Reality*, Wiley Series in Communications Networking and Distributed Systems, pp. 1007, 2003.
- [94] Ertoz L., Steinbach M., Kumar V., Finding Clusters of Different Size, Shapes and Densities in Noisy, High-Dimensional Data, Army High Performance Center, technical report, April 2003.
- [95] Ferreira L., (2002) *Introduction to Grid Computing with Globus*, IBM Redbook series, IBM Corporation, <http://ibm.-com/redbooks>

- [96] Foster I., Kesselman C (eds), Building a computational grid: state-of-the art and future directions in high-performance distributed computing. Morgan-Kaufmann, San Francisco, 1998.
- [97] Goltz, C., Fractal and Chaotic Properties of Earthquakes, Editor: Christian Goltz, *Lecture Notes in Earth Sciences*, 77, p.3-164, SpringerVerlag, Berlin, 1997.
- [98] Grossman R.L., Karnath, Ch, Kegelmeyer, P., Kumar, V., Namburu, R.,R., *Data Mining for Scientific and Engineering Applications*, Kluwer Academic Publisher, 2001.
- [99] Gutenberg, B. and C.F. Richter, Seismicity of the Earth and Associated Phenomena, Princeton Univ. Press, Princeton, N.J., 1954.
- [100] Haile P.M., *Molecular Dynamics Simulation*, Wiley: New York, 1992.
- [101] Hand, D., Mannila, H. and Smyth P., *Principles of Data Mining*, M.I.T. Press, Cambridge, Ma, 2001.
- [102] Hastie, T., Tibshirani, R. and Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 533 pp., Springer Verlag, New York, 2001.
- [103] Hartigan, J.A., *Clustering Algorithms*, John Wiley & Sons, Inc. New York, NY, USA, 351pp, 1975.
- [104] Jain, A., K., Dubes A., C., *Algorithms for Clustering Data*, Prentice-Hall Advanced Reference Series, pp.320, 1988.
- [105] Jajuga, K., Sokolowski, A. and Hermann, H., Editors, *Classification, Clustering and Data Analysis*, 497 pp., Springer Verlag, 2002.
- [106] Jones, N.C. and Pevzner, P., *An Introduction to Bioinformatics Algorithms*, MIT press, 2004.
- [107] Kalnay, E., *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 341pp., 2003.
- [108] Mitra, S., and Acharya, T., *Data Mining: Multimedia, Soft Computing and Bioinformatics*. Wiley, New Jersey, September 2003
- [109] Rundle, J.B., Turcotte, D.L. and W. Klein (Edtiors), *GeoComplexity and the Physics of Earthquakes*, 284 pp., American Geophysical Union, Washington , D.C., 2000b.
- [110] Sornette, D., *Critical Phenomena in Natural Sciences*, Springer Series in Synergetics, Berlin-Heidelberg, , 528 pp, 2006.
- [111] Theodoris, S. and Koutroumbas, K., *Pattern Recognition*, Academic Press, San Diego, London, Boston, 1998.
- [112] Tan, P-N, Steinbach, M., Kumar, V., *Introduction to data mining*, Addison Wesley; Boston MA, USA, , 769 pp., 2005.

- [113] Turcotte, D.L., *Fractals and Chaos in Geology and Geophysics*, 2nd Edition, Cambridge University Press, 1997.
- [114] Zadeh, L.A., *Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers by Lotfi A. Zadeh*, World Scientific Series In Advances In Fuzzy Systems, World Scientific Publishing Co., Inc. River Edge, NJ, USA, pp 826, 1996.

XI Supplementary material

Movie 1 and 2: The 3-D visualization of multi-dimensional structures is a very powerful instrument in pattern recognition and data learning. The colored points represent 6-dimensional feature vectors transformed by MDS procedure onto 3-D image and displayed by Amira package. Four clusters can be easily detected. The rotation allows for better penetration of structure of the clusters, and check they separation abilities. The colors show the cluster membership. In case of overlapping clusters the better clustering scheme can be tried, or/and the noise and outliers removed. We show rotations around two different axes.