A novel tree-based algorithm to discover seismic patterns in earthquake catalogs

E. Florido [a], G. Asencio–Cortés [a], J.L. Aznarte [b], C. Rubio-Escudero [c], F. Martínez–Álvarez [a,*]

[a] Department of Computer Science, Pablo de Olavide University of Seville, ES-41013, Spain
[b] Department of Artificial Intelligence, Universidad Nacional de Educación a Distancia, Spain
[c] Department of Computer Science, University of Seville, Spain

Keywords:

Seismic time series

Earthquake prediction

Pattern discovery

Clustering

A B S T R A C T

A novel methodology is introduced in this research study to detect seismic precursors. Based on an existing approach, the new methodology searches for patterns in the historical data. Such patterns may contain statistical or soil dynamics information. It improves the original version in several aspects. First, new seismicity indicators have been used to characterize earthquakes. Second, a machine learning clustering algorithm has been applied in a very flexible way, thus allowing the discovery of new data groupings. Third, a novel search strategy is proposed in order to obtain non-overlapped patterns. And, fourth, arbitrary lengths of patterns are searched for, thus discovering long and short-term behaviors that may influence in the occurrence of medium-large earthquakes. The methodology has been applied to seven different datasets, from three different regions, namely the Iberian Peninsula, Chile and Japan. Reported results show a remarkable improvement with respect to the former version, in terms of all evaluated quality measures. In particular, the number of false positives has decreased and the positive predictive values increased, both of them in a very remarkable manner.

1. Introduction

The discovery of earthquake precursors is a task of utmost relevance in order to take precautionary measures and prevent human losses. According to Ishibashi (1998), such events can be classified into two categories: physical (irreversible rupture process) and tectonics (tectonic slide). Physical precursors are mainly considered for the short and intermediate term.

It is well-known that certain precursory events are correlated to large earthquakes. Actually, a vast majority of major earthquakes exhibit anomalous seismic activity just before they occur. The features include changes in regional activity rate and changes in the pattern of small earthquakes, including alignments on unmapped linear features near the (future) main shock. It has long been suggested that large earthquakes are preceded by observable variations in regional seismicity (Shanker et al., 2010).

The main objective of this work is to generalize the methodology introduced in (Morales-Esteban et al., 2010) and extended in (Florido et al., 2015). In it, authors applied unsupervised learning to discover significant precursory anomalies. Although the results they obtained were relevant in terms of accuracy, the approach itself exhibited several limitations:

1. Only *b*-value and time occurrence were considered to discover meaningful anomalies. That is, only two features were considered to characterize seismicity. There exist many other features that may be used.
2. The search strategy was not exhaustive and some patterns were just sub-patterns or shorter patterns of other patterns. Therefore, an improved search strategy must be developed in order to avoid overlapping in discovered patterns.
3. Only three labels were assigned when k-means were applied (the number of clusters was not thoroughly discussed and was set to 3). This number was particularly suitable for visualization, but certain physical behaviors may remain unrevealed.
4. The length of the patterns were limited to three elements, which involved the consideration of only 15 last events.
5. Results showed similar behavior for all the seven analyzed zones.

In view of the above mentioned, a novel methodology is introduced in

* Corresponding author.
E-mail addresses: eflonav@alu.upo.es (E. Florido), guaasecor@upo.es (G. Asencio–Cortés), jlaznarte@dia.uned.es (J.L. Aznarte), crubioescudero@us.es (C. Rubio-Escudero), fmaralv@upo.es (F. Martínez–Álvarez).

order to find non-overlapped patterns and to explore the entire search space. The heuristic is fed with a relevant set of seismicity parameters, already published in the literature. Moreover, a voting system is used to determine the optimal number of partitions to be used. And, finally, data for different zones in the world (Iberian Peninsula (Morales-Esteban et al., 2010), Chile (Reyes et al., 2013) and Japan (Asencio-Cortés et al., 2017)) are analyzed aiming at discovering common patterns.

The rest of the paper is structured as follows. Section 2 reviews recent and relevant works in this field of research. The followed methodology to discover earthquake precursors is described in Section 3. Results achieved from the application of the proposed methodology to several catalogs are summarized and discussed in Section 4. Finally, Section 5 presents the conclusions drawn in this work.

2. Related works

A very thorough survey of earthquake precursors was published in (Cicerone et al., 2009). Seismicity, surface deformations, temperature changes or magnetic fields can be encountered among all analyzed precursors. Three main conclusions were drawn from this study: largest anomalies occur before largest earthquakes, the number of anomalies increases when approaching to the earthquake and precursory anomalies usually take place closer to the epicenter. Another interesting survey can be found in (Florido et al., 2016), in which the performance of artificial neural networks is reviewed under a variety of earthquake prediction problems.

The *b*-value was pointed out to be an earthquake precursor in 1981, when Smith (1981) performed an analysis in New Zealand, showing that in areas surrounding eventual earthquake epicenters, *b*-value initially increases and, later, it decreases after earthquake occurrence.

In 1992, it was found that a pattern for large earthquakes, consisting in an intermediate-term increase followed by short-term decrease in the *b*-value (Sammonds et al., 1992). Additionally, they limited the influence of the *b*-value to a period no longer than seven years.

Similar conclusions were drawn by Nuannin et al. in 2005 for the region of Andaman-Sumatra (Nuannin et al., 2004). The authors introduced a thorough study on *b*-value variations by means of a time-sliding window. It was concluded that an acute decrease in the *b*-value is usually reported prior to major earthquakes.

The predictive ability of three seismic parameters (number of earthquakes, *b*-value and energy released) was studied in (Baskoutas and Popandopoulos, 2014). These parameters were included in the proposed FastBEE algorithm. Data from Greece were used to assess its performance, showing promising results.

Unglert et al. compared the performance of Self-Organizing Maps versus Principal Component Analysis when applied to synthetic data (Unglert et al., 2016). These data were built from retrieved information from two volcanic eruptions. They concluded that hierarchical cluster performs better.

Seismic signals have also been analyzed by means of pattern recognition techniques in Chile (Curilem et al., 2016). In particular, the authors proposed two strategies to combine information retrieved from different monitoring stations, in order to improve the precursor classification performance.

Last et al. proposed the use of new seismic parameters, along with other existing ones, to predict earthquakes in Israel (Last et al., 2016). Both foreshocks and aftershocks were removed from the catalog. Their new proposed features, based on the number of earthquakes and the maximum earthquake magnitude during the same year, exhibited remarkable predictive ability.

Fault deformations prior to the 2016 Qinghai Menyuan earthquake, 6.4 M_s, were studied in (Li et al., 2016). The authors found long anomalous tendencies near the epicenter the days before the earthquake took place. They even reported anomaly sites approximately a year earlier and kept increasing and migrating towards areas surrounding the epicenter several months prior to the event.

Swarm magnetic data from the 2015 Nepal earthquake, 7.8 M_s, were analyzed in (De Santis et al., 2017). The authors applied a statistical approach to detect temporal patterns and relevant anomalies were discovered. The authors claimed that this fact indicates an internal origin of the anomalies discovered.

A hydrogeochemical dataset associated with the 2016 Amatrice-Norcia, Italy, seismic sequence (Anzidei and Pondrelli, 2016) was analyzed in (Barberio et al., 2017). They reported variations of pH values, an increase of As, V, and Fe concentrations, whereas Cr concentrations increased immediately after the main shock. The authors interpreted the anomalies within the dataset as reliable seismic precursors for a dilational tectonic setting.

3. Methodology

The methodology proposed to perform earthquake predictions is described in this section. It is a methodology based on finding precursor patterns that involve further large magnitude earthquake events. Those patterns are based on seismic features derived from the Gutenberg-Ritcher's *b*-value (Gutenberg and Richter, 1944). The parameters on which the methodology depends are automatically optimized leaving the user free to tune them.

In Section 3.1 the entire procedure is summarized. Section 3.2 explains how seismic information was extracted from earthquake catalogs producing propositional datasets. Once those datasets are generated, the training procedure carried out to produce precursor models is explained in Section 3.3. Finally, the prediction procedure is described in Section 3.4.

3.1. General procedure

The methodology proposed is drawn in a schematic way in Figs. 1 and 2. The first figure shows the training procedure while the second one shows the prediction procedure. First, catalogs of earthquake events were taken from various works in the literature (Morales-Esteban et al., 2010; Reyes et al., 2013; Asencio-Cortés et al., 2017) in order to establish comparisons. Specifically, Section 4.1 describes the seven catalogs considered. Therefore, the proposed methodology was tested on seven different datasets.

The univariate time series corresponding to the magnitude of the earthquake events in the catalogs was considered in this work. The time series is defined by the sequence of magnitudes $\{m_i\}$ and it is indexed by the sequence of times $\{t_i\}$, where $1 \leq i \leq n$ and n is the time series length.

A propositional dataset is constructed for each earthquake catalog. Such dataset contains a set of seismic features $(a_1, a_2, \ldots, a_m)$ (see Section 3.2) and a target to predict (C). The target is defined as in the reference work. In all these works the target was the maximum earthquake magnitude in the next days.

Once the propositional dataset is generated, it was divided in two subsets named training and test. This division was performed exactly as in the reference work for each catalog. Reference works splitted the datasets using hold-out based on specific separated training and test subsets.

The training procedure is explained in Fig. 1. The core of the training process is based on the following sequence of tasks: clustering (Section 3.3.1), grouping (Section 3.3.2), construction of a precursor tree (Section 3.3.3), pattern extraction (Section 3.3.4) and pattern selection (Section 3.3.5).

Because the clustering and grouping tasks depend on parameters K and A, respectively, an exhaustive search was introduced wrapping the core. Note that K is the number of clusters to be created and A is the length of the pattern sequence to be found within historical data. This search iterates over all combinations in the given K, A grids and produces the best parameter values according to a measure of performance (explained in Section 4.2). The final precursor model contains several objects: the training clusters, the best value obtained for the parameter A

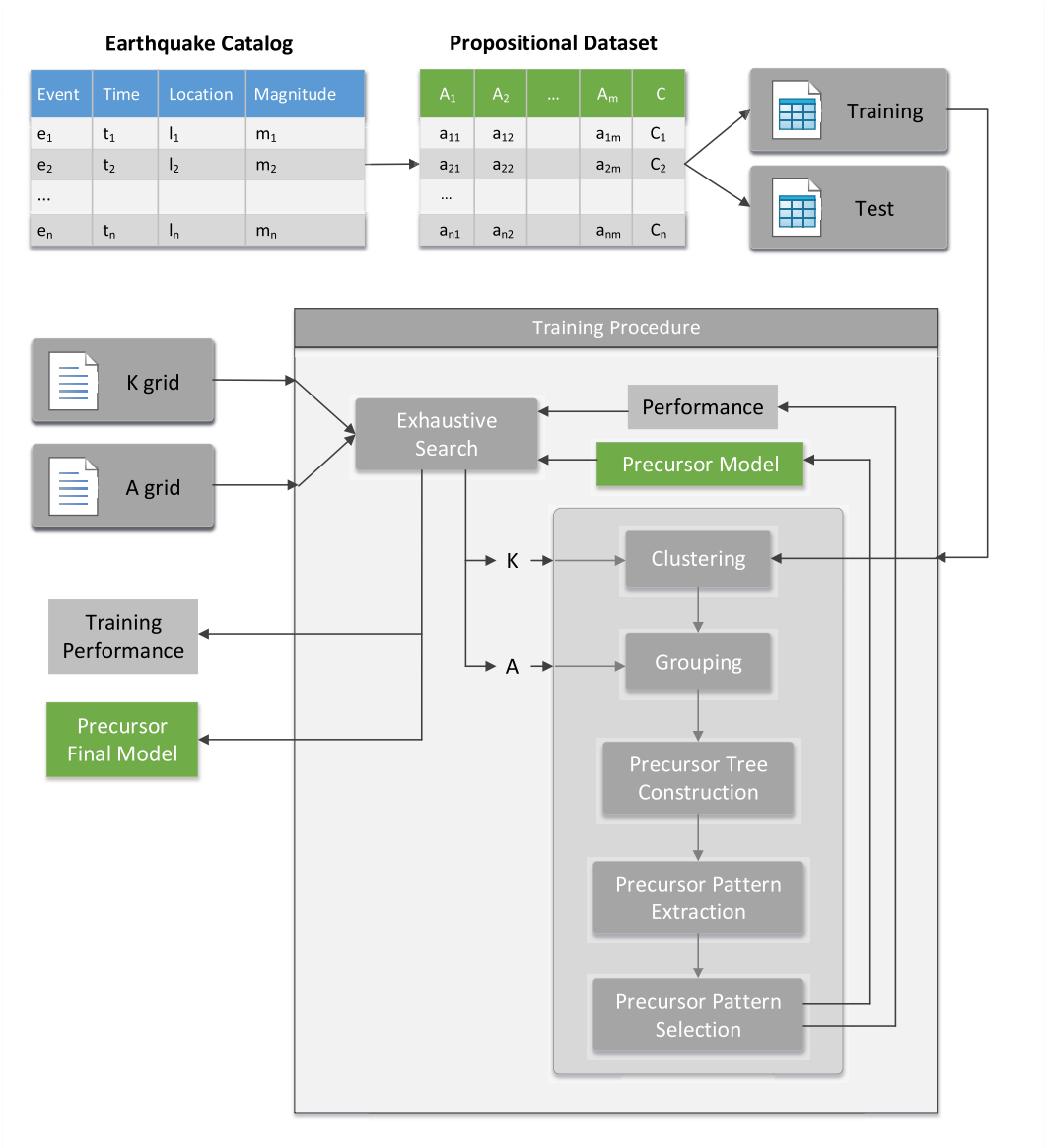

Fig. 1. Methodology proposed for training and producing precursor models from earthquake catalogs. It involves the creation of propositional datasets and an exhaustive search over parameters K and A. The precursor model is built through a sequence of tasks including clustering, grouping, construction of a precursor tree and pattern extraction and selection. The training procedure finishes returning the precursor model and its performance.

and the selected precursors. Precursors are formally defined in Section 3.3.3.

The test procedure is shown in Fig. 2. In first place, the propositional instances corresponding to the test events of the catalog are generated. From the previously generated precursor final model, a sequence of tasks were performed: cluster assignment, grouping instances, find precursor patterns and, finally, the prediction assignment. Once predictions were made, their performance is analyzed in Section 4.

3.2. Propositional dataset generation

A propositional dataset is constructed for each earthquake catalog. Such dataset contains tabular data structured on a set of attributes $a_1, a_2, \ldots, a_m$ and a class C. Such data structure has the standard meaning in machine learning area: attributes could contain relevant information to infer the class, which is the objective to estimate.

The class C is defined as the maximum magnitude of events in the next days. Depending on the specific work, those days can vary. The class considered in the proposed methodology is the specific for each reference work to perform correct comparisons.

Attributes $a_1, a_2, \ldots, a_m$ are seismic features mostly derived from the Gutenberg-Ritcher's b-value (Gutenberg and Richter, 1944). The b-value is the size distribution factor. It reflects the tectonics of the underlying zone and it is related to the geophysical properties of the zone. Specifically, Table 1 shows the set of seismic features used in the reference works (Morales-Esteban et al., 2010; Reyes et al., 2013; Asencio-Cortés et al., 2017).

The features $x_1, x_2, x_3, x_4, x_5, x_6$ and x_7 were proposed in (Reyes et al., 2013) and $b, a, \eta, \Delta M, T, \mu, c, dE^{1/2}$ and M_{mean} were introduced in (Panakkat and Adeli, 2007). To assess the features for a given event, the previous n events must be calculated. In this work n was set to the same specific value of the reference works ($n = 50$). The seismic features considered in this work were the specific for each reference work, in order to provide the same input information to the proposed methodology and to compare properly its prediction results.

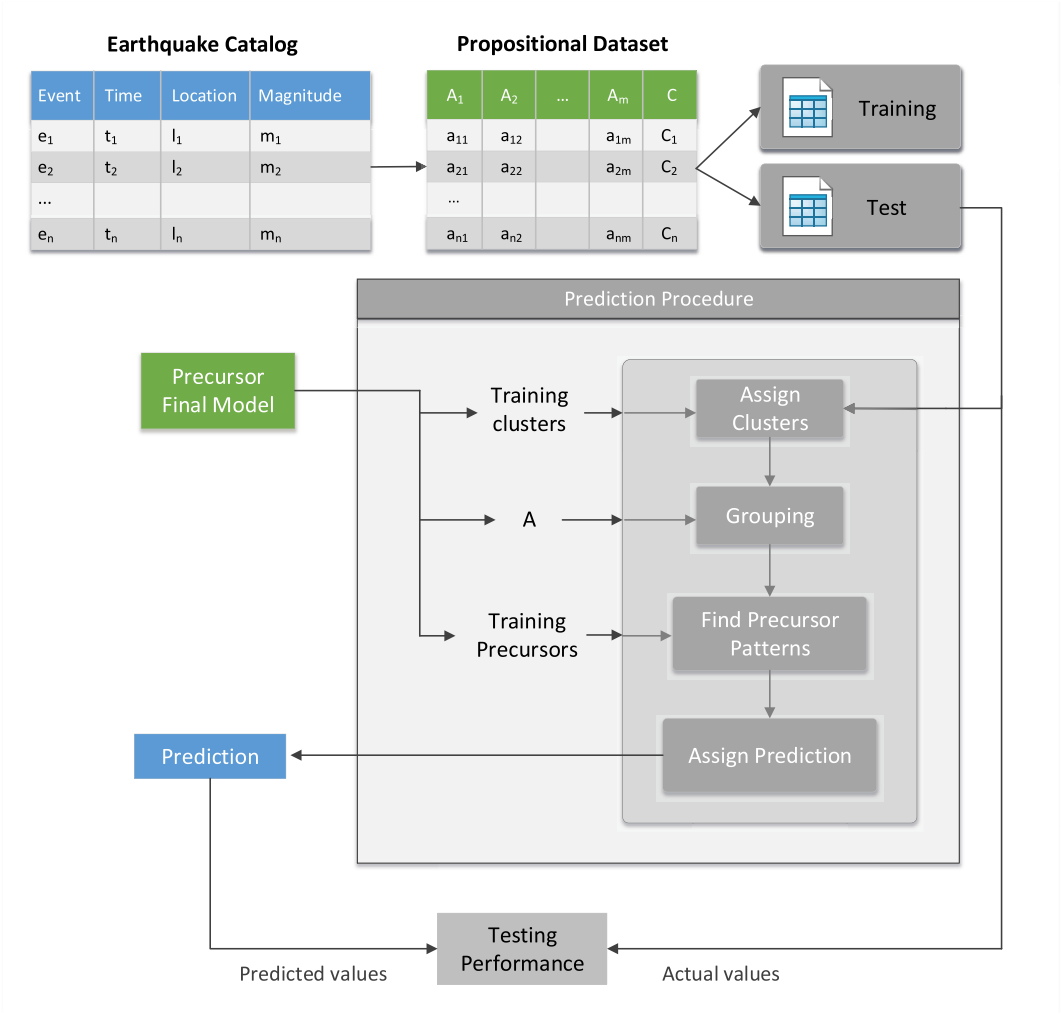

Fig. 2. The prediction procedure of the proposed methodology. First, propositional instances from test events are generated. Then, from such instances and a precursor final model, a sequence of tasks are performed, including cluster assignment, grouping instances, find precursor patterns and prediction assignment. Finally, performance of predictions is assessed and analyzed.

Table 1
Seismic feature set used in the reference works.

Feature	Description
b	Gutenberg-Richter law's b-value
x_1	Increment of b between the events i and $i-4$
x_2	Increment of b between the events $i-4$ and $i-8$
x_3	Increment of b between the events $i-8$ and $i-12$
x_4	Increment of b between the events $i-12$ and $i-16$
x_5	Increment of b between the events $i-16$ and $i-20$
x_6	Maximum magnitude from the events recorded during the last week (OU's law)
x_7	Probability of recording an event with magnitude larger or equal to 6.0 using a probability density function
a	Gutenberg-Richter law's a-value
η	Mean square deviation
ΔM	Magnitude deficit
T	Elapsed time
μ	Mean time
c	Coefficient of variation
$dE^{1/2}$	Rate of square root of seismic energy
M_{mean}	Mean magnitude

3.3. Training procedure

The procedure to generate the precursor model from the propositional training subset is shown in Algorithm 1. Note that it is based, in high level terms, on an exhaustive search over parameters K and A. Parameter K is the number of clusters used in the clustering phase. Parameter A is the group size used in the grouping phase.

The algorithm receives a grid of values for both K and A to perform the search for their optimum values. In this work, the following grids were used: $K_{grid} = 2, 3, 4, 5, 6$ and $A_{grid} = 2, 3, 4, 5, 6$. The algorithm also receives the training attributes ($a_{i,j} \in \mathbb{R}, \forall i, j : 1 \leq i \leq n, 1 \leq j \leq m$) and its classes ($C_i \in \{0, 1\}, \forall i : 1 \leq i \leq n$).

The algorithm returns the set of precursors obtained M^*, the best parameter values K^* and A^*, the cluster assignment to training instances L^* and the performance ϕ of the obtained precursors. This performance measurement is defined in Section 4.2. All these objects compose the knowledge model produced by the training procedure of the proposed methodology. This model will be used to perform earthquake predictions, as it is explained in Section 3.4.

Algorithm 1: Training procedure

Input :

$a \in \mathbb{R}^{n \times m}$ (training attributes)
$C \in \{0,1\}^n$ (training classes)
$K_{grid} \in \mathbb{N}^{K_l}$ (K grid values)
$A_{grid} \in \mathbb{N}^{A_l}$ (A grid values)

Output:

$M^* \in < S_j^i, H, G >^{M_l}$ (precursors)
$K^* \in K_{grid}$ (best value for parameter K)
$A^* \in A_{grid}$ (best value for parameter A)
$CT^* \in \mathbb{R}^{n \times (m+1)}$ (clustered training)
$\phi \in [0,1] \in \mathbb{R}$ (performance)

```
1  φ = 0;
2  foreach K ∈ K_grid do
3      foreach A ∈ A_grid do
4          CT := clustering(a, K);
5          G := grouping(CT, A, C);
6          T := precursortree(G, K, A);
7          M := extraction(T);
8          <M, φ_0> := selection(M, G);
9          if φ_0 > φ then
10             <M*, K*, A*, CT*, φ> := <M, K, A, CT, φ_0>;
11         end
12     end
13 end
```

3.3.1. Clustering instances

Clustering process is aimed to group seismic similarities among the training instances. For this purpose, all the attributes $a_1, a_2, \ldots, a_m$ were normalized between 0 and 1. Then, the k-means clustering algorithm (Lloyd, 1982) was applied to the training instances (excluding their classes C). The number of clusters is determined by the parameter K managed by the training procedure.

After clustering process is completed, the result $CT \in \mathbb{R}^{n \times (m+1)} = a|L$ is composed of the training instances plus a column vector L with their cluster assignments. These assignments must be within the number of clusters K: $L_i \in \{1, \ldots, K\}$, $\forall i: 1 \leq i \leq n$.

3.3.2. Grouping instances

The aim of the grouping process is to form sequences of consecutive cluster assignments in training instances (precursors) and their consequence in possible next large earthquakes. For this purpose, instances are condensed in groups of A elements. The process is shown in Fig. 3 for $A = 3$ (an example).

The grouping process results in a matrix named grouped training $G \in \mathbb{N}^{(n-A+1) \times (A+1)}$. The first A columns of the matrix $G_1, G_2, \ldots, G_A$ contain the cluster assignments of the grouped instances. The last column $C^* = C_i \in \{0,1\}$, $\forall i: A \leq i \leq n$, contains the class value of the last grouped instance.

Elements of the matrix G are defined in two parts: a) the submatrix with the columns G_i is defined as $g_{ij} \in \mathbb{N}$, $\forall i: 1 \leq i \leq (n - A + 1)$, $\forall j: 1 \leq j \leq A$; b) elements of the column vector C^* are defined as $c_i^* \in \{0,1\}, \forall i: 1 \leq i \leq (n - A + 1)$. Thus, the matrix $G = g|c^*$.

Thereby, the class of last grouped instances, C^*, means whether a large earthquake will occur in the next days and the columns $G_1, G_2, \ldots, G_A$ represent the pattern of previous seismic features (precursors) which could be the cause of these further earthquakes.

3.3.3. Construction of the precursor tree

The main objective was to select the most characteristic and general

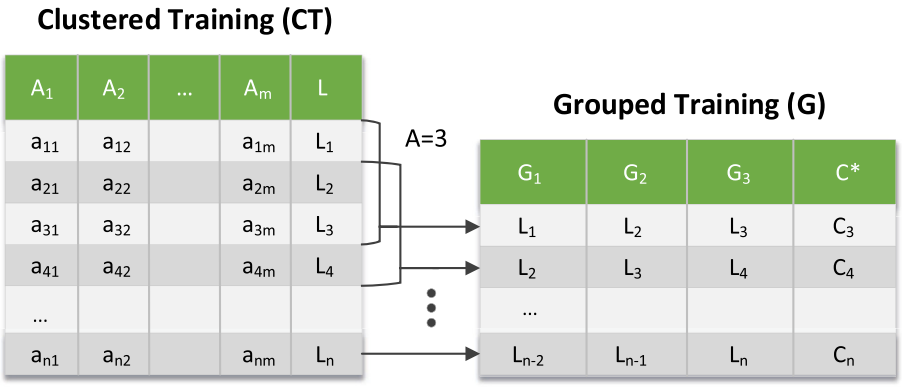

Fig. 3. Example of the process of grouping instances for $A = 3$.

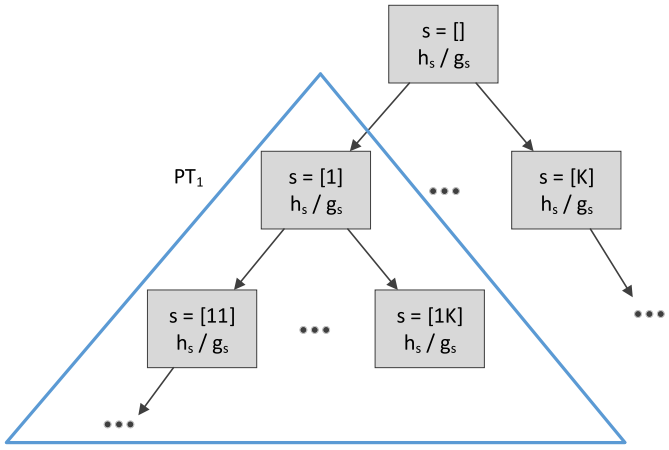

Fig. 4. Representation of a precursor tree with K clusters.

precursors that involves large further earthquakes. Once such precursors are obtained they can be used as patterns to predict earthquakes. To achieve this goal, the next stage of the proposed methodology, once the grouped training (G) was built, is to construct a tree-based model we named precursor tree. Fig. 4 shows a representation of a precursor tree defined with K clusters. In this section all fundamentals about precursor trees are defined.

In first place, let's introduce all the definitions about precursors. A precursor P_s is a tuple of three components, as defined in Eq. (1a). The component s is a sequence containing up to A elements with values between 1 and K, as it is shown in Eqs. (1b) and (1c). Specifically, s is a sequence of cluster assignments that corresponds, partial or totally, to a row in the submatrix g.

$$P_s =< s, h_s, g_s > \tag{1a}$$

$$s \in S_j : \ 0 \leq j \leq A \tag{1b}$$

$$S_j = \{s_1, \ldots, s_j\} : 1 \leq s_k \leq K, \ \ \forall k = 1..j; S_0 = \{\} \tag{1c}$$

$$0 \leq h_s \leq g_s \wedge 0 \leq g_s \leq n - A + 1 \tag{1d}$$

If s contains exactly A elements, then $\exists\, i\,,\ 1 \leq i \leq (n - A + 1) : \ s = g_i$. For sequences s with less than A elements, s corresponds to a row in the submatrix g' formed by the last columns of g. Specifically, $\exists i, 1 \leq i \leq (n - A + 1) : s = g_i^{'} \wedge g^{'} = G_{A-|s|+1} | \ldots | G_A \in \mathbb{N}^{(n-A+1)\times|s|}$, where $|s|$ is the length of s.

The component g_s of the precursor is the number of rows of G with which s matches (partial or totally); i.e. the number of occurrences of s in G. Since the matrix G is derived from earthquake catalogs, seismic features and cluster assignments, precursors P_s are linked to the original data.

Due to only precursors that involves large further earthquakes are object of interest for prediction, the component h_s of precursors counts the occurrences of s in G which their corresponding class C^* is 1. According to these definitions, h_s must be less than or equal to g_s, as it is shown in Eq. (1d).

The nodes of the precursor tree are all precursors P_s such that $g_s > 0$, in order to omit sequences of patterns that are not present in the original data. As it can be seen in Fig. 4, the root node contains the precursor with the void sequence ($s = [\ ]$). In this precursor, $g_s = n - A + 1$ and $h_s = \sum C^*$ (i.e. the number of rows in G with the class $C^* = 1$).

The precursor tree is a K-ary tree, i.e. each node of the precursor tree can have up to K child nodes, as it is shown in Fig. 4. Given a node of the tree, each child adds a different number (between 1 and K) at the end of the sequence s of the parent. The tree can have up to $A + 1$ levels of depth. Therefore, this data structure can represent all precursors of a dataset in an hierarchical way.

3.3.4. *Precursor pattern extraction*

In order to select the most characteristic and general precursors that involves large further earthquakes, in this phase of the methodology, the K best precursors were extracted from the precursor tree. Specifically, one precursor is extracted from each subtree $PT_i, \forall i : 1 \leq i \leq K$. Note that one of the main drawbacks of the previous versions consisted in discovering overlapped patterns. By extracting just one precursor from each subtree this situation is avoided. As an example, in Fig. 4, the PT_1 of the precursor tree is highlighted in blue.

The best precursor extracted for each subtree PT_i is the one that has the maximum ratio h_s/g_s (most characteristic precursor). If there are more than one precursor with the maximum ratio h_s/g_s, the one that has the shorter sequence s is extracted (most general precursor).

3.3.5. *Precursor pattern selection*

Once the K best precursors were extracted, in this phase the best subset of them is selected, because not all precursors together will produce necessarily the best prediction results. For this purpose, an exhaustive search over all possible subsets of precursors within the best K was performed.

Due to the evaluation of goodness of a set precursors is very fast, in terms of computation time, the space of subsets is addressed in a reasonable time. Specifically, such evaluation measurement is defined in Section 4.2 and it involves the calculation of TP, TN, FP and FN values.

Finally, the best subset of precursors selected in this phase are returned as the precursor model, finishing the training procedure.

3.4. *Prediction procedure*

The procedure to produce predictions from the previously generated precursor model is shown in Algorithm 2. The algorithm receives the test attributes, the clustered training (generated in the clustering phase of the training procedure), the best value for parameter A and the best subset of precursors M^* (achieved after the precursor pattern selection phase). Regarding the last input, M_l is the number of precursors obtained in the training phase, $M_l = |M^*|$. The algorithm returns the predicted values $\widehat{C}$ for the test classes.

The procedure is repeated for each test instance a_i. First, a training cluster is assigned to the instance a_i. For this assignment, the Euclidean distance between a_i and each cluster centroid is used. The cluster whose centroid has the lower distance is assigned to the instance, and the corresponding cluster number is appended at then of the vector a_i, resulting in a_i^*.

After cluster assignment, the grouping process performs the same operation explained in Section 3.3.2 only applied to the test instance a_i and producing one group of A^* elements. The grouped test instance G is then used to find precursors within the precursor subset M^*, which are returned in P.

Finally, if no precursors were found ($P = \varnothing$), the prediction will be $\widehat{C}_i = 0$, which means that no earthquake with large magnitude will occur in the next days. On the other hand, if precursors were found, the predicted will be positive ($\widehat{C}_i = 1$).

Algorithm 2: Prediction procedure

Input :

$a \in \mathbb{R}^{n \times m}$ (test attributes)
$L^* \in \mathbb{R}^{n \times (m+1)}$ (clustered training)
$A^* \in A_{grid}$ (best value for parameter A)
$M^* \in < S_j^i, H, G >^{M_l}$ (precursors)

Output:

$\widehat{C} \in \{0, 1\}^n$ (predicted classes)

```
foreach a_i ∈ a do
    a_i^* := assigncluster(a_i, L^*);
    G := grouping(a_i^*, a, A^*);
    P := findprecursors(G, M^*);
    if P = ∅ then
        Ĉ_i := 0;
    else
        Ĉ_i := 1;
    end
end
```

4. Results

This section presents the results achieved in this work, after application of the methodology described in the previous section. First, the data used is introduced in 4.1. Section 4.2 describes the quality parameters used to assess the method performance.

4.1. Data description

Data from Spain, Chile and Japan have been used in this paper. In particular, the datasets used in (Morales-Esteban et al., 2010) are first analyzed. In them, data from two different Iberian Peninsula seismogenic zones can be found. Data from four different zones in Chile are also analyzed in this work, and correspond to those firstly studied in (Reyes et al., 2013). Finally, data from Japan studied in (Asencio-Cortés et al., 2017) are also here analyzed.

Table 2 summarizes information about these datasets. Note that the minimum magnitude of interest for all the datasets is 4.5. That is, patterns associated with earthquakes with magnitude larger than 4.5 are searched for every zone. ZMAP was used to decluster catalogs (Wiemer, 2001).

4.2. Quality parameters to assess the model

True positives (TP) identify the occurrence of earthquakes with magnitude greater or equal to 4.5 when any of the considered sequences of labels are present. On the other hand, the false negatives (FN) represent the number of cases in which a medium-large earthquake also occurs but no proposed sequences of labels are found. True negatives (TN) and false positives (FP) refer to the situation in which no earthquakes occurred. However, the TN denotes that no proposed sequences appear, while the FP makes reference to the apparition of any of the considered sequences.

Table 2
Datasets used in the study.

Zone	Country	Events	Cell size	Date	Declustered
Alboran Sea	Iberian Peninsula	532	1°	1980–2007	Yes
Western Azores-Gibraltar fault	Iberian Peninsula	443	1°	1980–2007	Yes
Santiago	Chile	531	0.5°	2003–2011	Yes
Talca	Chile	353	0.5°	2003–2011	Yes
Valparaíso	Chile	530	1°	2006–2011	Yes
Pichilemu	Chile	353	1°	2005–2011	Yes
Tokyo	Japan	464	2°	2013–2015	No

In addition, six well-known indices are provided: sensitivity, specificity, predictive positive value (PPV), negative positive value (NPV), Matthew's correlation coefficient (MCC) and accuracy. In this context, the sensitivity quantifies the grade of reliability of the method when real events take place while the specificity measures the reliability of the method when sequences of labels are discarded. PPV measures how reliable positive predictions are, whereas NPV measures how reliable negative predictions are. Finally, MCC and accuracy stand for global measures. MCC is in essence a correlation coefficient between the observed and predicted binary classifications, whereas the accuracy accounts for all the actual predictions of the algorithm, irrespective of positive or negative predictions.

All measures range from 0 to 100%, except for MCC whose range is $[-1, +1]$, where $+1$ represents a perfect prediction, 0 no better than random prediction and -1 indicates total disagreement between prediction and observation.

These indices are defined by the following equations:

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{FP + TN} \tag{3}$$

$$PPV = \frac{TP}{TP + FP} \tag{4}$$

$$NPV = \frac{TN}{TN + FN} \tag{5}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6}$$

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{7}$$

4.3. Results for the Iberian Peninsula

Results for the Iberian Peninsula are reported in this section. Several seismogenic zonings are proposed in the literature for the Iberian Peninsula (Martínez-Álvarez et al., 2015; Morales-Esteban et al., 2014).

Table 3
Results performance for the Iberian Peninsula: Western-Azores Gibraltar Fault and Alboran Sea.

Parameters	Western-Azores Gibraltar Fault		Alboran Sea	
	Former version	New version	Former version	New version
TP	23	25	9	9
FN	6	4	1	1
FP	5	3	15	8
TN	47	49	71	78
Sensitivity	79.31%	86.21%	90.00%	90.00%
Specificity	90.38%	94.23%	82.56%	90.70%
PPV	82.14%	89.29%	37.50%	52.94%
NPV	88.68%	92.45%	98.61%	98.73%
MCC	0.7026	0.8109	0.5119	0.6458
Accuracy	86.42%	91.36%	83.33%	90.63%

However, the original work considered the zoning proposed by Martín (1984). In particular, datasets selected to assess the performance of the proposed methodology are those published in (Morales-Esteban et al., 2010), where two zones were analyzed: Western-Azores Gibraltar Fault and the Alboran Sea.

Results achieved with the former methodology were already satisfactory. However, with the new strategy two new TP have been detected and two FP have been avoided for Western-Azores Gibraltar Fault, as summarized in Table 3. These findings lead to improving global measures values from 0.7026 to 0.8109 (MCC) or from 86.42% to 91.36% (accuracy). These facts are also reflected in improved PPV and specificity, parameters directly related to the number of FP, reaching 89.29% and 94.24%, respectively. The increase of TP also influence other parameters such as sensitivity and NPV with new values set to 86.21% and 92.45%.

Results for the Alboran Sea present similar behavior since they are overall improved. FN remain equals to 1, that is, there is one actual occurrence that remains hidden to the new approach. However, the number of FP is significantly decreased from 15 to 8 and TN is increased from 71 to 78. As a consequence, the new specificity reaches 90.70% and PPV 52.94%. Global measures experiment a major improvement as well, with new MCC and accuracy equals to 0.6458 and 90.63%, respectively.

In conclusion, the results have been improved by 5.72% and 8.76% for Western-Azores Gibraltar Fault and the Alboran Sea, respectively, if only the total number of instances properly classified is taken into account.

4.4. *Results for Chile*

Results for four main cities in Chile and surroundings are discussed in this section. It is worth noting that the original seismogenic zoning was proposed in (Reyes and Cárdenas, 2010) and refined in (Reyes et al., 2013; Morales-Esteban et al., 2014), from which the datasets have been extracted.

The former methodology achieved moderately satisfactory results for the city of Santiago, as can be seen in Table 4. MCC was slightly less than 0.5, which is the minimum threshold for a classification to be considered satisfactory. In this sense, both sensitivity and PPV must be clearly outperformed (57.14% and 42.11%, respectively) in order to make use of the patterns discovered with the former methodology. The main achievement with the new methodology lies on the drastic decrease of FP (reduced by 50%) and the significant increase of TP (incremented from 8 to 11), along with their associated improvements in FN and TN rates. These facts lead to an improved 78.57% sensitivity and 61.11% PPV, as well as remarkable values for specificity (95.07%) and NPV (87.83%). Due to this overall improvement, global measures such as MCC reached 0.6588 and accuracy 93.59% (see Table 5).

Results for Pichilemu, on the contrary, were already quite good in terms of PPV and there was no much room for improvement. It can be seen that only 3 FP had been triggered (in a total of 14 triggers), having already reached PPV equals to 95.38%. With the new methodology one less FP was triggering which involves slight improvement in PPV (96.88%). As for FN, unfortunately, it cannot be reduced and therefore the sensitivity remained at 84.93%, which was already satisfactory. Finally, MCC and accuracy reached 0.6823 and 86.32%, respectively.

Table 4
Results performance for Chile: Santiago, Pichilemu, Valparaíso and Talca.

	Santiago		Pichilemu	
Parameters	Former version	New version	Former version	New version
TP	8	11	62	62
FN	6	3	11	11
FP	11	7	3	2
TN	131	135	19	20
Sensitivity	57.14%	78.57%	84.93%	84.93%
Specificity	92.25%	95.07%	86.36%	90.91%
PPV	42.11%	61.11%	95.38%	96.88%
NPV	95.62%	97.83%	63.33%	64.52%
MCC	0.4317	0.6588	0.6470	0.6823
Accuracy	89.10%	93.59%	85.26%	86.32%
	Valparaíso		Talca	
Parameters	Former version	New version	Former version	New version
TP	28	36	14	19
FN	24	16	8	3
FP	17	13	4	3
TN	131	135	18	19
Sensitivity	53.85%	69.23%	63.64%	86.36%
Specificity	88.51%	91.22%	81.82%	86.36%
PPV	62.22%	73.47%	77.78%	86.36%
NPV	84.52%	89.40%	69.23%	86.36%
MCC	0.4450	0.6165	0.4623	0.7273
Accuracy	79.50%	85.50%	72.73%	86.36%

Table 5
Results performance for Japan: Tokyo.

Parameters	Former version	New version
TP	7	9
FN	12	10
FP	12	5
TN	61	68
Sensitivity	36.84%	47.37%
Specificity	83.56%	93.15%
PPV	36.84%	64.29%
NPV	83.56%	87.18%
MCC	0.2040	0.4567
Accuracy	73.91%	83.70%

In Valparaíso, there were already 28 TP and 131 TN. However, the number of FP was 17 which led to a 62.22% PPV. Since the number of FN was 24, the sensitivity obtained a poor 53.85%. Overall, global measures threw 0.4450 and 79.50% values for MCC and accuracy, respectively. The most relevant achievement in this dataset is reduction of FN by 33% (from 24 to 16). Additionally, FP has also been decreased up to 13 and TN increased up to 135 accordingly. In this new situation, values for specificity, NPV or accuracy highlight, reaching 91.22%, 89.40% and 85.50%, respectively.

The last Chilean city analyzed in this study is Talca. Reported results showed only 4 FP but 8 FN, which resulted in a 77.78% PPV and a 63.64% sensitivity. MCC almost reached 0.5 as happened in Santiago and Valparaíso, which indicated that results must be improved. In this sense, the new methodology reported only 3 FN (increasing FP from 14 to 19) and decreased FP from 4 to 3 (increasing TN by one unit as well). Unsually, all the observed quality parameters reached exactly the same value: 86.36% and MCC reached a remarkable 0.7273 value.

Finally, the results have been improved by 5.04%, 1.23%, 7.55% and 18.75% for Santiago, Pichilemu, Valparaíso and Talca, respectively, if only the total number of instances properly classified is taken into account.

4.5. Results for Japan

Results for the city of Tokyo are reported and discussed in this section. Although several seismogenic zonings have been proposed in the literature (Hashimoto et al., 2009), the dataset introduced in (Asencio-Cortés et al., 2017) has been selected since it is only focused in Tokyo and neighborhood.

It can be noticed that results with the former methodology were not that totally satisfactory. Although accuracy reached 73.91%, which could be considered good enough in some contexts, too many FP were triggered, in particular, 12. This value is high if compared to the number of TP, 7, since it leads to a PPV rate slightly over 35% or a MCC value equals to 0.2040. Moreover, sensitivity was not particularly high either (36.84%) whereas rates associated with negatives remained at 83.56% (both specificity and NPV). These values do not meet the minimum requirements to be considered useful in the field of earthquake.

Nevertheless, the improved methodology's output draws a new scenario. The main achievement lies in the significant reduction of FP: from 12 to 5. Therefore the value for TN is increased from 61 to 68. This situation involves new hit rates, i.e., PPV and specificity new values are 64.29% and 93.15%, respectively. The number of TP has been increased as well, being able to accurately discover patterns preceding 9 earthquakes. Analogously, the number of FN has been decreased by 2 units. The effect is clear: sensitivity and NPV has been increased up to 47.37% and 87.18%, respectively. As for global measures, MCC is now 0.4567 and accuracy 83.70%.

It can be concluded that the results have been improved by 13.25% for the Tokyo dataset, if only the total number of instances properly classified is taken into account.

5. Conclusions

A novel methodology to detect earthquake precursory patterns is proposed in this work. In particular, the approach improves its previous version in several aspects. First, new features have been considered in order to discover patterns with diverse shapes. Second, patterns are not overlapped and those found ensure unique shapes. Third, the number of clusters generated is not necessary equals to 3. Finally, patterns can be of arbitrary shape. To assess the performance of the novel approach, seven different datasets from three seismic zones –Iberian Peninsula, Chile and Japan– have been analyzed. Reported results are promising and lead to the conclusion that similar patterns could be found across the Earth due to the different nature of the data used.

Acknowledgements

The authors would like to thank Spanish Ministry of Science and Technology and Junta de Andalucía for the support under projects TIN2011-28956-C00 and P12-TIC-1728, respectively. This work has also been partially funded by a Spanish Ramón y Cajal grant, RYC-2012-11984.

References

Anzidei, M., Pondrelli, S., 2016. The Amatrice seismic sequence: preliminary data and results. Ann. Geophys. 59 (5) ag–7373.

Asencio-Cortés, G., Martínez-Álvarez, F., Morales-Esteban, A., Troncoso, A., 2017. Medium-large earthquake magnitude prediction in Tokyo with artificial neural networks. Neural Comput. Appl. 28 (5), 1043–1055.

Barberio, M.D., Barbieri, M., Billi, A., Doglioni, C., Petitta, M., 2017. Hydrogeochemical changes before and during the 2016 Amatrice-Norcia seismic sequence (central Italy). Sci. Rep. 7 (1), 11735.

Baskoutas, I., Popandopoulos, G.A., 2014. Precursory seismicity pattern before strong earthquakes in Greece. Res. Geophys. 4, 7–11.

Cicerone, R.D., Ebel, J.E., Britton, J., 2009. A systematic compilation of earthquake precursors. Tectonics 476, 371–396.

Curilem, M., Huenupan, F., Beltrán, D., San Martín, C., Fuentealba, G., Franco, L., Cardona, C., Acuña, G., Chacón, M., Khan, M.S., Yoma, N.B., 2016. Pattern recognition applied to seismic signals of Llaima volcano (Chile): an evaluation of station-dependent classifiers. J. Volcanol. Geoth. Res. 315, 15–27.

Florido, E., Aznarte, J.L., Morales-Esteban, A., Martínez-Álvarez, F., 2016. Earthquake magnitude prediction based on artificial neural networks: a survey. Croat. Oper. Res. Rev. 7 (2), 159–169.

Florido, E., Martínez-Álvarez, F., Morales-Esteban, A., Reyes, J., Aznarte, J.L., 2015. Detecting precursory patterns to enhance earthquake prediction in Chile. Comput. Geosci. 76, 112–120.

Gutenberg, B., Richter, C.F., 1944. Frequency of earthquakes in California. Bull. Seismol. Soc. Am. 34, 185–188.

Hashimoto, C., Noda, A., Sagiya, T., Matsu'ura, M., 2009. Interplate seismogenic zones along the Kuril-Japan trench inferred from GPS data inversion. Nat. Geosci. 2, 141–144.

Ishibashi, K., 1998. Two categories of earthquake precursors, physical and tectonic, and their roles in intermediate-term earthquake prediction. Pure Appl. Geophys. 126 (2), 687–700.

Last, M., Rabinowitz, N., Leonard, G., 2016. Predicting the maximum earthquake magnitude from seismic data in Israel and its neighboring countries. PLoS One 11 (3), e0151751.

Li, Y., Gan, W., Wang, Y., Chen, W., Liang, S., Zhang, K., Zhang, Y., 2016. Seismogenic structure of the 2016 Ms6.4 Menyuan earthquake and its effect on the Tianzhu seismic gap. Geodesy and Geodynamics 7 (4), 230–236.

Lloyd, Stuart, 1982. Least squares quantization in PCM. IEEE Trans. Inf. Theor. 28 (2), 129–137.

Martín, A.J., 1984. Riesgo sísmico en la Península Ibérica. Ph. D. Politechnical University of Madrid, Spain.

Martínez-Álvarez, F., Gutiérrez-Avilés, D., Reyes, J., Amaro-Mellado, J.L., Rubio-Escudero, C., Morales-Esteban, A., 2015. A novel method for seismogenic zoning based on triclustering. Application to the Iberian Peninsula. Entropy 17 (7), 5000–5021.

Morales-Esteban, A., Martínez-Álvarez, F., Troncoso, A., de Justo, J.L., Rubio-Escudero, C., 2010. Pattern recognition to forecast seismic time series. Expert Syst. Appl. 37 (12), 8333–8342.

Morales-Esteban, A., Martínez-Álvarez, F., Scitovski, S., Scitovski, R., 2014. A fast partitioning algorithm using adaptive mahalanobis clustering with application to seismic zoning. Comput. Geosci. 73, 132–141.

Nuannin, P., Kulhanek, O., Persson, L., 2004. Spatial and temporal b value anomalies preceding the devastating off coast of nw sumatra earthquake of december 26. Geophys. Res. Lett. 32, 2005.

Panakkat, A., Adeli, H., 2007. Neural network models for earthquake magnitude prediction using multiple seismicity indicators. Int. J. Neural Syst. 17 (1), 13–33.

Reyes, J., Cárdenas, V., 2010. A Chilean seismic regionalization through a Kohonen neural network. Neural Comput. Appl. 19, 1081–1087.

Reyes, J., Morales-Esteban, A., Martínez-Álvarez, F., 2013. Neural networks to predict earthquakes in Chile. Appl. Soft Comput. 13 (2), 1314–1328.

Sammonds, P.R., Meredith, P.G., Main, I.G., 1992. Role of pore fluid in the generation of seismic precursors to shear fracture. Nature 359, 228–230.

De Santis, A., Balasis, G., Pavón-Carrasco, F.J., Cianchini, G., Mandea, M., 2017. Potential earthquake precursory pattern from space: the 2015 Nepal event as seen by magnetic Swarm satellites. Earth Planet Sci. Lett. 461 (1), 119–126.

Shanker, D., Singh, H.N., Paudyal, H., Kumar, A., Panthi, A., Singh, V.P., 2010. Searching for an earthquake precursor–a case study of precursory swarm as a real seismic pattern before major shocks. Pure Appl. Geophys. 167 (6), 655–666.

Smith, W.D., 1981. The b-value as an earthquake precurso. Nature 289 (5794), 136–139.

Unglert, K., Radic, V., Jellinek, A.M., 2016. Principal component analysis vs. self-organizing maps combined with hierarchical clustering for pattern recognition in volcano seismic spectra. J. Volcanol. Geoth. Res. 320, 58–74.

Wiemer, S., 2001. A software package to analyze seismicity: ZMAP. Seismol Res. Lett. 72, 373–382.