Computers & Geosciences 76 (2015) 112–120

Contents lists available at ScienceDirect

Computers & Geosciences

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

Detecting precursory patterns to enhance earthquake prediction in Chile

E. Florido [a,b], F. Martínez-Álvarez [b,*], A. Morales-Esteban [c], J. Reyes [d], J.L. Aznarte-Mellado [a]

[a] Department of Artificial Intelligence, Universidad Nacional de Educación a Distancia – UNED, Spain
[b] Department of Computer Science, Pablo de Olavide University of Seville, Spain
[c] Department of Building Structures and Geotechnical Engineering, University of Seville, Spain
[d] TGT-NT2 Labs, Chile

ARTICLE INFO

Article history:
Received 23 October 2014
Received in revised form
4 December 2014
Accepted 10 December 2014
Available online 11 December 2014

Keywords:
Seismic time series
Earthquake prediction
Pattern discovery
Clustering

ABSTRACT

The prediction of earthquakes is a task of utmost difficulty that has been widely addressed by using many different strategies, with no particular good results thus far. Seismic time series of the four most active Chilean zones, the country with largest seismic activity, are analyzed in this study in order to discover precursory patterns for large earthquakes. First, raw data are transformed by removing aftershocks and foreshocks, since the goal is to only predict main shocks. New attributes, based on the well-known *b*-value, are also generated. Later, these data are labeled, and consequently discretized, by the application of a clustering algorithm, following the suggestions found in recent literature. Earthquakes with magnitude larger than 4.4 are identified in the time series. Finally, the sequence of labels acting as precursory patterns for such earthquakes are searched for within the datasets. Results verging on 70% on average are reported, leading to conclude that the methodology proposed is suitable to be applied in other zones with similar seismicity.

1. Introduction

Chile stands out as the most seismic country in the world, as evidenced by the fact that in the last 450 years in the Chilean mainland a total of 38 earthquakes of a magnitude larger than 7.5 M_s have taken place; 17 of them resulted in tsunamis (Cárdenas-Jirón, 2013).

Moreover, the greatest magnitude earthquake occurred in history around the world and known as *Earthquake of Valdivia* (May 22, 1960) took place in Chile and reached a magnitude of 9.5 M_s, producing a 10 m high tsunami and reaching Hawaii (Cisternas and Atwater, 2005).

The great Chilean seismicity is mainly explained by the subduction between the Nazca and the South American plates at a rate of 8 cm/year. This seismic activity is resulting from four different sources: interplate thrust, medium depth intraplate, cortical and outer-rise (Reyes et al., 2013).

According to Reyes and Cárdenas (2010), Chile can be divided into six seismic regions. This study is focused on the four most active ones through the following cities: Talca (Region 3), Pichilemu (Region 4), Santiago (Region 5), and Valparaíso (Region 6).

This study is meant to be considered within the context of the earthquake prediction algorithms (Martínez-Álvarez et al., 2011; Morales-Esteban et al., 2013; Reyes et al., 2013), aiming at reducing the impact of the damage and avoiding thus the loss of many human lives (Duncan, 2005; Stein and Toda, 2013; Tucker, 2013).

Thus, the authors analyze time series of earthquakes in each of the above-mentioned cities in order to discover precursory patterns for large earthquakes. In particular, it is intended to provide a reasonable probability that an earthquake larger than a certain threshold will be observed in a near future. To fulfil such a task, the analysis of the *b*-value parameter has been carried out as well as the application of unsupervised learning techniques, in a similar way as done in Morales-Esteban et al. (2010). To enhance earthquake prediction accuracy, it is desired that discovered patterns are transformed into information valid for supervised classifiers. That is, the final goal is to transform these patterns into new seismicity indicators (Panakkat and Adeli, 2007; Reyes et al., 2013; Zamani et al., 2013), in order to provide better predictions.

The remainder of the paper is structured as follows. Section 2 reviews the state of the art in earthquake prediction. Section 3 describes the underlying geophysical fundamentals as well as the methodology used to discover precursory patterns. The results for the four cities analyzed are shown in Section 4.

* Corresponding author.
E-mail addresses: eflorido1@alumno.uned.es (E. Florido), fmaralv@upo.es (F. Martínez-Álvarez), ame@us.es (A. Morales-Esteban), daneel@geofisica.cl (J. Reyes), jlaznarte@dia.uned.es (J.L. Aznarte-Mellado).

http://dx.doi.org/10.1016/j.cageo.2014.12.002
0098-3004/

Finally, the conclusions drawn from this study are discussed in Section 5.

2. Related works

Earthquake prediction is a highly complex problem whose hypothetical solution will imply invaluable benefits to society: being able to predict an earthquake means being able to avoid some of its worst consequences. Hence, for at least the last 100 years it has been the center of attention for researchers from many countries and several disciplines and approaches. The modern era of scientific earthquake prediction dates back to the 1970s, with the known case of Haicheng earthquake (Zhu and Wu, 1977).

Over the preceding months of this earthquake, there were changes in land elevation and ground water levels, and a regional increase in seismicity was also registered. This variation in earthquake activity finally triggered an evacuation warning. The earthquake of magnitude 7.3 struck the region days later on February 4, 1975. Thanks to the prediction, and the prompt response from the authorities, only around 1% of the casualties expected for such a magnitude in that area materialized (2041 people died, whereas it was estimated that this number could have reached over 150,000 without the said prediction and subsequent evacuation).

Despite the optimism inspired by this success, the amount of examples of failures to predict earthquakes (which consequently resulted in human catastrophes) is much higher and entails that earthquake forecasting remains an open problem with ample room for improvement.

Following the apparent lack of success of the Parkfield prediction experiment (Bakun and Lindh, 1985), a harsh controversy took place amongst the scientific community, in which the predictability of earthquakes was in dispute (Bakun et al., 2005). This debate marked a milestone in the area and contributed to focus the scientific efforts towards predicting time-dependent earthquake hazards including the associated probabilities and errors.

The International Commission on Earthquake Forecasting for Civil Protection, appointed by the Italian Government following the 2009 L'Aquila earthquake, summarized in a thorough report (Jordan et al., 2011) the state of the art in operational earthquake prediction. According to this report, searching for diagnostic earthquake precursors (physical, chemical or biological changes that can be related to the occurrence of a seismic event) and analyzing the information conveyed by them is one of the most common approaches.

Seismic precursors include changes in strain rates, seismic wave speeds and electrical conductivity; variations of radon concentrations in groundwater, soil and air; fluctuations in groundwater levels; electromagnetic variations near and above Earth's surface; thermal anomalies; anomalous animal behavior; and seismicity patterns.

Foreshocks (smaller earthquakes registered before an important one or main shock occurs) and other patterns of seismicity can be observed by seismic networks and have been extensively investigated as precursors. Taken as individual events, foreshocks cannot be discriminated *a priori* from background seismicity and hence cannot be used as diagnostic precursors. However, statistics, automatic learning and pattern recognition methods have been used to discern any repetitive patterns in seismic activity that might be caused by precursory processes.

The availability of larger earthquake catalogs, including events with smaller magnitudes, together with the computational advances that facilitate complex time series analysis contributes to the blooming of new methods in the last years. A review of seismicity-based earthquake forecasting techniques can be found in Tiampo and Shcherbakov (2012). These authors draw a distinction amongst physical process models (models which assume that a particular physical mechanism is associated with the generation of large earthquakes and their precursors) and smoothed seismicity models (models which apply some kind of filtering technique to seismicity catalog data in order to forecast seismic events). This last family is not related to the results presented in this paper and originated in the works by Frankel (1995).

Amongst the physical process models, accelerating moment release (AMR, see Mignan (2011) for a review) plays an influential role, while the characteristic earthquake hypothesis (Schwartz and Coppersmith, 1984) also stands as a prominent research line. Other proposed models include the M8 family of algorithms (Ismail-Zadeh and Kossobokov, 2011), the Region–Time–Length (RTL) algorithm (Gambino et al., 2014; Sobolev and Tyupkin, 1999), the Load–Unload Response Ratio (LURR) (Yin et al., 1995) or the Pattern Informatics (PI) index (Holiday et al., 2006).

The study of variations in the *b*-value in the wake of an earthquake is also a fertile approach which has been studied intensively over the past 20 years. A comprehensive review over the early research in this issue is performed in Wiemer and Wyss (2002), which demonstrates that the *b*-value is highly heterogeneous in both space and time and on a wide variety of scales whereas persistent variations occur that are correlated with the stress field in fault zones.

3. Methodology

This section introduces the methodology applied to discover precursory patterns in *b*-values for earthquakes in Chile. First, Section 3.1 provides a description of the geophysical fundamentals underlying this study. The description of the methodology itself can be found in Section 3.2.

3.1. Gutenberg–Richter law

Earthquake magnitude distribution has been observed from the beginning of the 20th century. Ishimoto and Iida (1939) and Gutenberg and Richter (1942) observed that the number of earthquakes, N, of magnitude larger than or equal to M follows a power law distribution defined by

$$N(M) = \alpha M^{-B} \tag{1}$$

where α and B are adjustment parameters.

Gutenberg and Richter (1954) transformed this power law into a linear law expressing this relation for the magnitude frequency distribution of earthquakes as

$$\log_{10}(N(M)) = a - bM \tag{2}$$

This law relates the cumulative number of events $N(M)$ with magnitude larger than or equal to M with the seismic activity, a, and the size distribution factor, b. The a-value is the logarithm of the number of earthquakes with magnitude larger than or equal to zero. The b-value is a parameter that reflects the tectonics of the area under analysis (Lee and Yang, 2006) and it has been related to the physical characteristics of the area. A high value of the parameter implies that the number of earthquakes of small magnitude is predominant and, therefore, the region has a low resistance. Contrariwise, a low value shows that the relative number of small and large events is similar, implying a higher resistance of the material.

Gutenberg and Richter used the least squares method to estimate coefficients in the frequency–magnitude relation from (2). Shi and Bolt (1982) pointed out that the *b*-value can be obtained

by this method but the presence of even a few large earthquakes has a significative influence on the results. The maximum likelihood method, hence, appears as an alternative to the least squares method, which produces estimates that are more robust when infrequent large earthquakes happen. They also demonstrated that for large samples and low temporal variations of b, the standard deviation of the estimated b is

$$\sigma(\hat{b}) = 2.30b^2\sigma(M) \tag{3}$$

where

$$\sigma^2(M) = \frac{\sum_{i=1}^{n}(M_i - \overline{M})^2}{n} \tag{4}$$

with n being the number of events and M_i the magnitude of a single event.

It is assumed that the magnitudes of the earthquakes that occur in a region and in a certain period of time are independent and identically distributed variables that follow the Gutenberg–Richter law (Ranalli, 1969). This hypothesis is equivalent to supposing that the probability density of the magnitude M is exponential:

$$f(M, \beta) = \beta\exp[-\beta(M - M_0)] \tag{5}$$

where

$$\beta = \frac{b}{\log(e)} \tag{6}$$

and M_0 is the cutoff magnitude.

Thus, in order to estimate the b-value, a previous estimation of β is necessary. In Utsu (1965), the maximum likelihood method was applied to obtain a value for β, defined by

$$\beta = \frac{1}{\overline{M} - M_0} \tag{7}$$

where $\overline{M}$ is the mean magnitude of all the earthquakes in the dataset.

From all the aforementioned possibilities, the maximum likelihood method has been selected for the estimation of the b-value in this work.

3.2. Metaheuristic to discover seismic patterns in Chile

The methodology proposed in order to discover knowledge from earthquakes time series is described in this section.

First of all, the earthquake's dataset is constructed as follows. Each earthquake is represented by three features: the magnitude, the b-value and the date of occurrence. Thus, the i-th earthquake is defined by

$$E_i = (M_i, b_i, t_i) \tag{8}$$

where M_i is the magnitude of the earthquake, b_i is its associated b-value and t_i is the date when the earthquake took place.

The b-value is determined from (6) and (7) considering the 50 preceding events (Nuannin et al., 2005). As done in Morales-Esteban et al. (2013), the cutoff magnitude is set to 3.0.

Furthermore, data are grouped into sets of five chronologically ordered earthquakes according to the methodology proposed in Nuannin et al. (2005). Thus, a simpler law with easier interpretation is provided. Each group $\hat{E}_j$ is represented by the mean of the magnitude of the five earthquakes, the time elapsed from the first earthquake and the fifth one and the signed variation of the b-values in this time interval for the five earthquakes in $\hat{E}_j$, i.e.,

$$\hat{E}_j = \{E_{k-4}, \ldots, E_k\} \quad \text{with } k = 5j \text{ and } j = 1, \ldots, \lfloor N/5 \rfloor \tag{9}$$

where N is the number of earthquakes in the dataset and $\lfloor N/5 \rfloor$ is the greatest integer less than or equal to $N/5$. Thus, we can define

$$G_j = (\overline{M}_j, \Delta b_j, \Delta t_j) \tag{10}$$

where

$$\overline{M}_j = \frac{1}{5}\sum_{i=k-4}^{k} M_i \quad \text{with } k = 5j, \tag{11}$$

$$\Delta b_j = b_k - b_{k-4} \quad \text{with } k = 5j, \tag{12}$$

$$\Delta t_j = t_k - t_{k-4} \quad \text{with } k = 5j, \tag{13}$$

and the dataset is composed by the temporal sequence of all G_j:

$$\mathrm{DS} = \{G_1, G_2, \ldots, G_{\lfloor N/5 \rfloor}\} \tag{14}$$

The goal is to find patterns in the data that precede the apparition of earthquakes with a magnitude larger than or equal to 4.4. Hence, the k-means algorithm is applied to the dataset, DS, with the aim of classifying the samples into different groups. As a previous step, the optimal number of clusters has to be determined Morales-Esteban et al. (2014) since the K-means algorithm needs this number as input data. For this purpose, the well-known Silhouette index is applied to clustered data for different numbers of clusters, as proposed in Morales-Esteban et al. (2010). Thus, each sample is considered only by the label assigned by the K-means algorithm in further analysis. That is, DS is transformed into a sequence of labels, SL:

$$\mathrm{SL} = \{C_1, C_2, \ldots, C_{\lfloor N/5 \rfloor}\} \tag{15}$$

where each label $C_j = \{1, 2, \ldots, K\}$ is the cluster assigned by k-means to the group G_j.

Once these labels have been obtained, specific sequences of labels are searched for precursory patterns for medium-large earthquakes. In particular, let a label C_j^* identifying a set of earthquakes contain at least one of magnitude larger than the preset threshold. Note that this threshold will be set to 4.5 for Chile, although it could be any other. Later, subsequences of labels, SL_W preceding such C_j^* are selected. The length of these subsequences can be dynamically selected, but it will be set to two as done in Morales-Esteban et al. (2010), $W=2$. Finally, all possible sequences (a total of K^W) are searched for in SL and every time that it is followed by any C_j^*, the number of occurrences for such SL_W is increased by one unit. The subsequences SL_W designated as precursory patterns are those that globally maximize the set of quality measures proposed in this study, and detailed in Section 4.1.

Fig. 1 illustrates the process of transformation that the original dataset is subjected to, for $K=3$ and $W=2$. Note that gray boxes denote that an event with magnitude larger than 4.4 was encountered in this label, that is, these boxes represent C_j^*. It can be observed that there were two $\mathrm{SL}_W = \{C_3, C_2\}$ and one $\mathrm{SL}_W = \{C_2, C_1\}$ preceding a target event C_j^*.

4. Results

This section first describes the quality parameters used to assess the performance of the proposed metaheuristic. It also shows the results obtained from the application of the methodology described in Section 3 to the four most active cities in Chile.

4.1. Quality parameters

There are several parameters used to evaluate the quality of the results. In this work these will be used the same as in Martínez-Álvarez et al. (2013). Hence, true positives (TPs) identify the

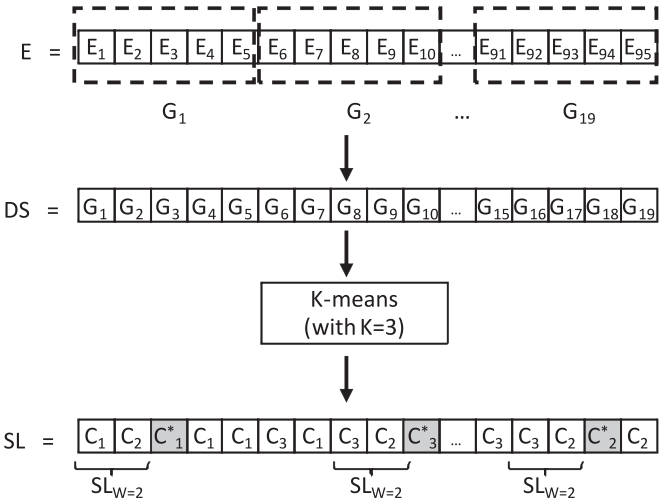

Fig. 1. Transformation of the original dataset into a sequence of labels, for $K=3$.

occurrence of earthquakes with magnitude larger than 4.4 when any of the considered sequence of labels is present. False negatives (FNs) represent the number of cases in which a medium-large earthquake also occurs but no proposed sequences of labels are found prior to its occurrence. True negatives (TNs) and false positives (FPs) refer to the situation in which no earthquakes with magnitude larger than 4.4 occurred. However, the TN denotes that no proposed sequences appear, while the FP makes reference to the appearance of any of the considered sequences.

Additionally, two well-known indices are provided, the sensitivity (S_n) and the specificity (S_p). 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. These indices are defined by the following equations:

$$S_n = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{16}$$

$$S_p = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}} \tag{17}$$

4.2. Results for Talca

Talca is the capital of the Maule region in the province of the same name. It is located in the center of the country with an area of 232 km^2 and an estimated population of 250,000 inhabitants. In terms of seismic activity, it is a very active city and many earthquakes have been recorded throughout history.

During the study of this area between 2003 and 2012, 24 earthquakes of a magnitude larger than 4.4 M_s took place. These data are plotted in Fig. 2 along with the sequences obtained and the variation of the b-value during that period of time. The x-axis on this chart represents the time scale; on the left y-axis, the values for the labels obtained (note that each label represents a set of five earthquakes, as per the methodology described in Section 3). The b-value is also shown; and, finally, on the right y-axis the values for the magnitude of such earthquakes are represented.

Thus, by classifying earthquakes depending on the sequence recorded prior to its occurrence, they can be compared to the total number of occurrences of each sequence. Note that two of the 24 earthquakes took place outside the study period, occurring within the first 50 earthquakes analyzed and required for calculation of the b-value. Therefore, the final study has been conducted on 22 earthquakes of magnitude exceeding 4.4.

Table 1 shows how the earthquakes with a magnitude larger than 4.4 were classified. The left column identifies all the possible sequences encountered in this study. In this sense, the sequence $[C_i–C_j]$ identifies earthquakes that occurred just after a sequence of 10 earthquakes classified into C_i (the first five ones) and into C_j (the remaining five). Note that the number of clusters used was three and the length of the sequence two, as detailed in Section 3. The second column provides information about the number of times an earthquake occurred after that particular sequence. The last column gathers the number of times the sequence was found in the historical data.

The information shown in this table should be interpreted as follows. There were 16 occurrences of the sequence $C_1–C_1$. From these 16 times, only 6 of them were precursor of an earthquake with magnitude larger than or equal to the preset threshold. However, all the six sequences $C_1–C_2$ were found before a candidate earthquake. Therefore, this sequence will be eventually considered as a pattern preceding an earthquake.

Information associated with centroids generated by the k-means algorithm is shown in Table 2. Each row describes one of the three centroids (C_1, C_2 and C_3). The M column denotes the mean magnitude for all the earthquakes grouped in such cluster; Δb is the mean b-value increment; Δt is the mean time elapsed between the first and the last grouped earthquake; and *Membership* stands for the percentage of samples that were grouped in each cluster.

The charts displayed in the first column of Fig. 6 show all the nine possible changes of sequences obtained from the Talca dataset. Each chart starts at point (0, 0)—positioned in the middle of the axis. The first line joins this point to the first centroid of the starting cluster, and the second line travels from this point up to the increment of time and the increment or decrement of the b-value established by the centroid of the second cluster. The x-axis represents the increment of time (Δt) and the y-axis the increment or decrement of the b-value (Δb).

Consider now the sequence $C_1–C_1$. This pattern means that after 0.3910 years, the b-value increased by 0.0300 units. And the recorded mean magnitude was 3.3743.

After examining Fig. 2, Tables 1 and 2, the sequences selected as precursory patterns for the city of Talca are $[C_1–C_2]$, $[C_1–C_3]$ and $[C_2–C_2]$. Table 9 summarizes the results obtained in terms of the quality parameters described in Section 4.1.

The FP rate is shown to be very low due to the choice of candidate sequences that have been made. As for the rest of the cities, it is preferred to have less FP or, in other words, to infer less patterns but with higher reliability. This has a direct effect on the high specificity obtained, that is, with the set of sequences chosen, when a prediction is made that an earthquake will not occur, there is 82% certainty that this will be so. Nonetheless, the goal is to cover as many events as possible and that is the meaning of the high sensitivity that has been obtained, with 64% accuracy.

From a qualitative point of view, the three selected patterns exhibit the following features. First, $[C_1–C_2]$ and $[C_1–C_3]$ are characterized by a smooth and short increase of the b-value (C_1). Then the b-value decreases with a steep slope for a short time (C_2) or with a moderate slope but for a longer time (C_3). The third pattern sequence $[C_2–C_2]$, with a continued decline of the b-value, recorded up to six events (from eight total occurrences). It is noteworthy that the sequences $[C_2–C_3]$, $[C_3–C_2]$ and $[C_3–C_3]$, all of them with continued decreases in the b-value as occurs in $[C_2–C_2]$, did not register any event. The main difference is that the chosen pattern sequence showed a greater slope and in a much shorter time.

4.3. Results for Pichilemu

Pichilemu is the capital of the Cardenal Caro province in the

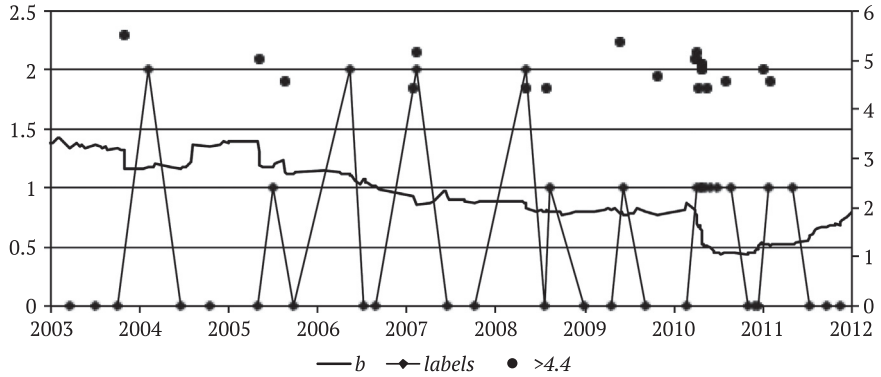

Fig. 2. Talca. *b*-value, labels and earthquakes.

Table 1
Talca – Earthquake classification.

Sequence	♯ Earthquakes	♯ Sequences
$[C_1–C_1]$	6	16
$[C_1–C_2]$	6	6
$[C_1–C_3]$	2	4
$[C_2–C_1]$	1	6
$[C_2–C_2]$	6	8
$[C_2–C_3]$	0	0
$[C_3–C_1]$	1	4
$[C_3–C_2]$	0	0
$[C_3–C_3]$	0	0

Table 2
Centroids of the clusters for Talca.

Cluster	M	Δb	Δt	Membership (%)
C_1	3.3743	+0.0150	0.1955	60.00
C_2	4.0018	−0.0385	0.0838	31.11
C_3	3.6138	−0.0848	0.4688	8.89

region of O'Higgins. It is located in the center of the country with an area of 749 km^2 and an estimated population of 12,800 inhabitants. In terms of seismic activity, it is considered as a very active city, like the rest of the country of Chile.

One of the most significative earthquakes took place in March 2010, and it is known as *the earthquake of the region of O'Higgins*, reached a magnitude of 6.9 M_s. The epicentre of the earthquake was located 15 km to the northeast of the city. Although at the beginning it was considered as a foreshock of the earthquake that occurred in February 27 of 8.8 M_s magnitude and that shocked the whole country, it was finally considered as a different earthquake. In the following hours to the earthquake of Pichilemu, a series of 11 seismic movements with a magnitude larger than 5.0 and two larger than 6.0 took place. Furthermore, the first one produced an early tsunami warning.

During the study on this area between 2010 and 2012, 75 earthquakes of a magnitude larger than 4.4 M_s took place. This data is plotted in Fig. 3 along with the sequences obtained and the variation of the *b*-value during that period of time.

Thus, by classifying earthquakes depending on the sequence recorded prior to its occurrence, they can be compared to the total number of occurrences of each sequence. Table 3 shows how the earthquakes with a magnitude larger than 4.4 were classified. The information shown must be interpreted like that of Table 1 and therefore interpreted as follows.

There were 8 occurrences of the sequence C_1–C_1. From these 8 times, only 3 of them were precursor of an earthquake with magnitude larger than or equal to the preset threshold. This reasoning can be extended to all the sequences C_i–C_j. Table 4 summarizes the centroids generated for Pichilemu, as done in Table 2.

The charts displayed in the fourth column of Fig. 6 show all the nine possible changes of sequences obtained from the Pichilemu dataset.

Consider now the sequence C_1–C_1. This pattern means that after 0.1236 years, the *b*-value increased by 0.0870 units. And the recorded mean magnitude was 3.4608.

After examining Fig. 3, Tables 3 and 4, the sequences selected as precursory patterns for the city of Pichilemu are $[C_1–C_3]$, $[C_2–C_3]$, $[C_3–C_2]$, $[C_3–C_1]$ and $[C_3–C_3]$. Table 9 summarizes the results obtained in terms of the quality parameters described in Section 3.

The FP rate is shown to be very low due to the choice of candidate sequences that has been made. As for the rest of the cities, it is preferred to have less FP or, in other words, to infer less patterns but with high reliability. This has a direct effect on the high specificity obtained, that is, with the set of sequences chosen, when a prediction is made that an earthquake will not occur, there is 85% certainty that this will be so. Additionally, a high sensitivity has been obtained, with 86% accuracy.

A qualitative analysis of the selected sequences in Pichilemu reveals the following conclusions. First, there are three sequences whose second label end with C_3, $[C_x–C_3]$, with $x=\{1, 2, 3\}$. That is, regardless of what happened before C_3, once this tag is detected, the method says that an earthquake of magnitude larger than the preset threshold will take place within the next five days. The other two sequences, $[C_3–C_1]$ and $[C_3–C_2]$, also share a property: the C_3 label is involved in both of them. The difference is that, this time, the occurrence of the earthquake will be foreseen for the next 5 or 10 registrable events.

4.4. Results for Santiago de Chile

Santiago is the capital of Chile. It is located in the metropolitan area of Santiago and in the province of the same name. It is located a little above the center of the country. It is a city with great seismic activity, and many earthquakes have been recorded throughout history.

During the study on this area between 2003 and 2012, 15 earthquakes of a magnitude larger than 4.4 M_s took place. This data is plotted in Fig. 4 along with the sequences obtained and the variation of the *b*-value during that period of time.

Again, by classifying earthquakes depending on the sequence recorded prior to its occurrence, they can be compared to the total number of occurrences of each sequence.

Table 5 shows how the earthquakes with a magnitude larger than 4.4 were classified, as in Tables 1 and 3. The information shown in this table should be interpreted as follows. There were

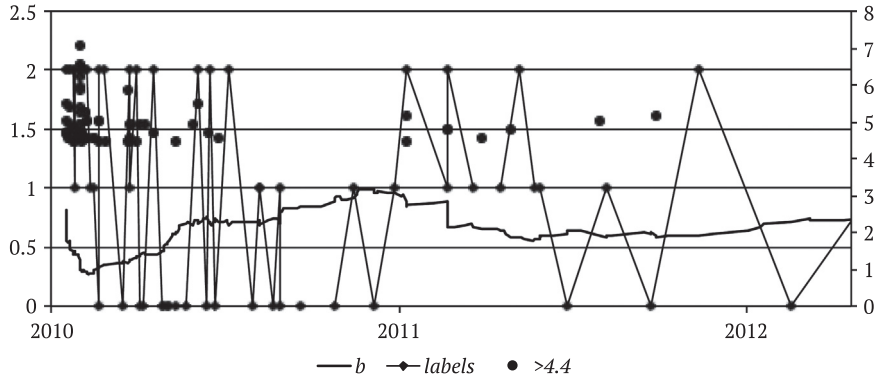

Fig. 3. Pichilemu. *b*-value, labels and earthquakes.

Table 3
Pichilemu – Earthquake classification.

Sequence	♯ Earthquakes	♯ Sequences
$[C_1–C_1]$	3	8
$[C_1–C_2]$	1	7
$[C_1–C_3]$	7	8
$[C_2–C_1]$	2	7
$[C_2–C_2]$	5	8
$[C_2–C_3]$	6	6
$[C_3–C_1]$	6	8
$[C_3–C_2]$	6	6
$[C_3–C_3]$	37	37

Table 5
Santiago – Earthquake classification.

Sequence	♯ Earthquakes	♯ Sequences
$[C_1–C_1]$	6	66
$[C_1–C_2]$	0	2
$[C_1–C_3]$	0	6
$[C_2–C_1]$	0	1
$[C_2–C_2]$	3	6
$[C_2–C_3]$	0	5
$[C_3–C_1]$	1	6
$[C_3–C_2]$	2	5
$[C_3–C_3]$	2	2

Table 4
Centroids of the clusters for Pichilemu.

Cluster	*M*	Δb	Δt	Membership (%)
C_1	3.4608	+0.0435	0.0618	31.51
C_2	3.5950	−0.0134	0.1317	30.14
C_3	4.3179	−0.0339	0.0320	38.36

Table 6
Centroids of the clusters for Santiago.

Cluster	*M*	Δb	Δt	Membership (%)
C_1	3.2374	+0.0194	0.0428	74.00
C_2	3.4003	−0.0067	0.2429	13.00
C_3	3.7266	−0.1037	0.2000	13.00

66 occurrences of the sequence C_1–C_1. From these 66 times, only 6 of them were precursor of an earthquake with magnitude larger than or equal to the preset threshold. Table 6 summarizes the centroids generated for Santiago, as done in Tables 2 and 4.

The charts displayed in the third column of Fig. 6 show all the nine possible changes of sequences obtained from the Santiago de Chile dataset. Consider now the sequence C_1–C_1. This pattern means that after 0.0856 years, the *b*-value increased by 0.0388 units. And the recorded mean magnitude was 3.2374.

After examining Fig. 4, Tables 5 and 6, the sequences selected as precursory patterns for the city of Santiago are $[C_2–C_2]$, $[C_3–C_1]$, $[C_3–C_2]$ and $[C_3–C_3]$. Table 9 summarizes the results obtained in terms of the quality parameters described in Section 3.

The TN rate is shown to be very high due to the choice of candidate sequences that have been made. This fact in junction with the low FP rate obtained has a direct effect on the high specificity obtained, that is, with the set of sequences chosen, when a prediction is made that an earthquake will not occur, there is 87% certainty that this will be so. This time the sensitivity reached 57%.

Perhaps Santiago is the area with the less competitive results. Although a sensitivity rate of 57% was obtained – what means that

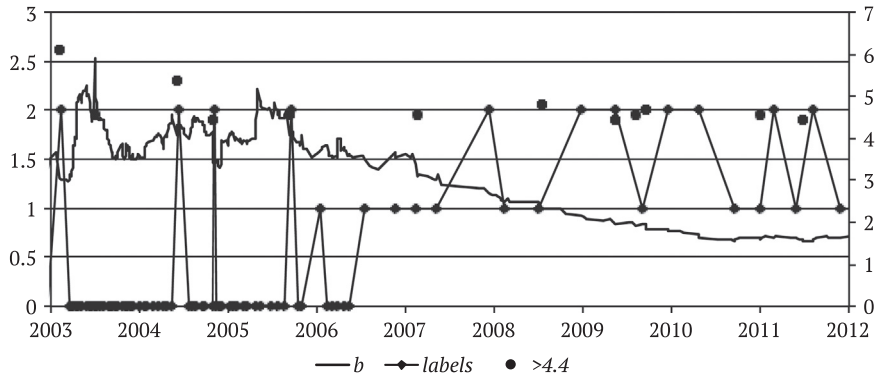

Fig. 4. Santiago. *b*-value, labels and earthquakes.

eight from the 14 earthquakes in the existing dataset were correctly predicted – to get this done 11 false alarms had to be triggered. Except for the case of the sequence $[C_3–C_3]$ where no false positives were included (the pattern was monotonously decreasing, following the trend of the patterns detected in other areas), the remaining selected sequences reached a moderate hit rate. For example, the sequence $[C_2–C_2]$ only added three out of the six events registered (also incorporated three false positives). One possible explanation for these results is that, in general, the found patterns are smoother, and there is no C_i tag identifying a significative decrease of the *b*-value. However, the values for sensitivity and specificity have been set to a satisfactory rate.

4.5. Results for Valparaíso

Valparaíso is the capital of the province and the region of the same name. It is located slightly above the center of the country with an area of 438 km^2 and an estimated population of 294,848 inhabitants. It is one of the three biggest cities in the country. In terms of seismic activity and in junction with all the others cities studied, it is considered as a very active city.

During the study on this area between 2003 and 2012, 53 earthquakes of a magnitude larger than 4.4 M_s took place. This data is plotted in Fig. 5 along with the sequences obtained and the variation of the *b*-value during that period of time.

Thus, by classifying earthquakes depending on the sequence recorded prior to its occurrence, they can be compared to the total number of occurrences of each sequence

Table 7 shows how the earthquakes with a magnitude larger than 4.4 were classified, as done in Tables 1, 3 and 5.

The information shown in this table should be interpreted as follows. There were 72 occurrences of the sequence $C_1–C_1$. From these 72 times, only 7 of them were precursor of an earthquake with magnitude larger than or equal to the preset threshold. Again, Table 8 shows the centroids generated for Valparaíso, as done in Tables 2, 4 and 6.

The charts displayed in the second column of Fig. 6 show all the nine possible changes of sequences obtained from the Valparaíso dataset.

Consider now the sequence $C_1–C_1$. This pattern means that after 0.0648 years, the *b*-value increased by 0.0586 units. And the recorded mean magnitude was 3.3184.

After examining Fig. 5, Tables 7 and 8, the sequences selected as precursory patterns for the city of Valparaíso are $[C_2–C_2]$, $[C_2–C_3]$, $[C_3–C_2]$ and $[C_3–C_3]$. Table 9 summarizes the results obtained in terms of the quality parameters described in Section 3.

The TN rate is shown to be very high compared to the FP rate due to the choice of candidate sequences that have been made. This has a direct effect on the high specificity obtained, that is, with the set of sequences chosen, when a prediction is made that an earthquake will not occur, there is 89% certainty that this will be so.

Table 7
Valparaíso – earthquake classification.

Sequence	♯ Earthquakes	♯ Sequences
$[C_1–C_1]$	7	72
$[C_1–C_2]$	3	8
$[C_1–C_3]$	8	33
$[C_2–C_1]$	1	4
$[C_2–C_2]$	2	3
$[C_2–C_3]$	9	9
$[C_3–C_1]$	5	38
$[C_3–C_2]$	2	4
$[C_3–C_3]$	15	29

Table 8
Centroids of the clusters for Valparaíso.

Cluster	M	Δb	Δt	Membership (%)
C_1	3.3184	+0.0293	0.0324	57.00
C_2	3.7422	−0.0348	0.2346	7.50
C_3	3.6432	−0.0432	0.0253	35.50

Note that the goal is to cover as many events as possible and such is the meaning of the high sensitivity that has been obtained, with 54% accuracy.

The case of the patterns found in Valparaíso is paradigmatic. The four selected patterns represent a monotonically decreasing function of the *b*-value, corroborating the hypothesis supposed at the state of the art. It also matches with the shape of the patterns found in the other cities. The only pattern expected to be more accurate is the sequence $[C_3–C_3]$. Looking at its shape it can be observed that it has a steady and pronounced decline (a steep slope), being the typical behavior for earthquake precursors. And in fact it is so, because 15 earthquakes took place after this pattern, rather more after any other pattern. However, it added 14 false alarms, number which was expected to be lower.

Generally speaking, two main conclusions can be drawn from the analysis of the selected precursory patterns. First of all, reported earthquakes with magnitude larger than 4.4 in Chile are preceded by patterns that usually last six months (average lasted time for patterns is 0.47 years). Therefore, there would be time enough to deploy policies in case these patterns were detected in its early state. Secondly, the shape of these patterns is always the same: an initial increase of the *b*-value followed by an acute decrease. In short, when the *b*-value starts decreasing, there is a serious danger that a large earthquake occurs.

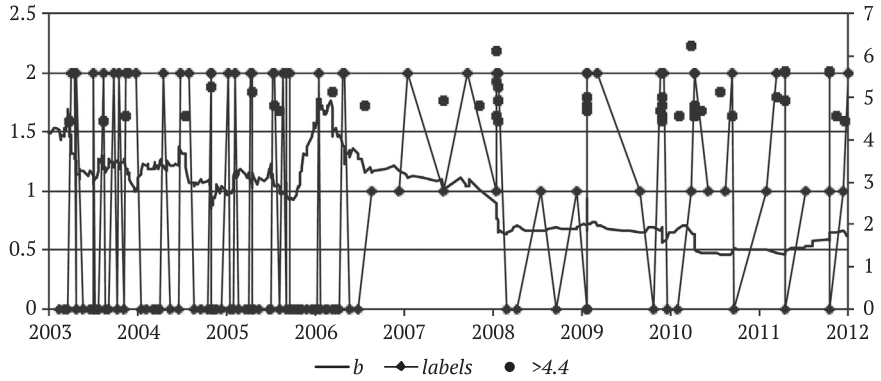

Fig. 5. Valparaíso. *b*-value, labels and earthquakes.

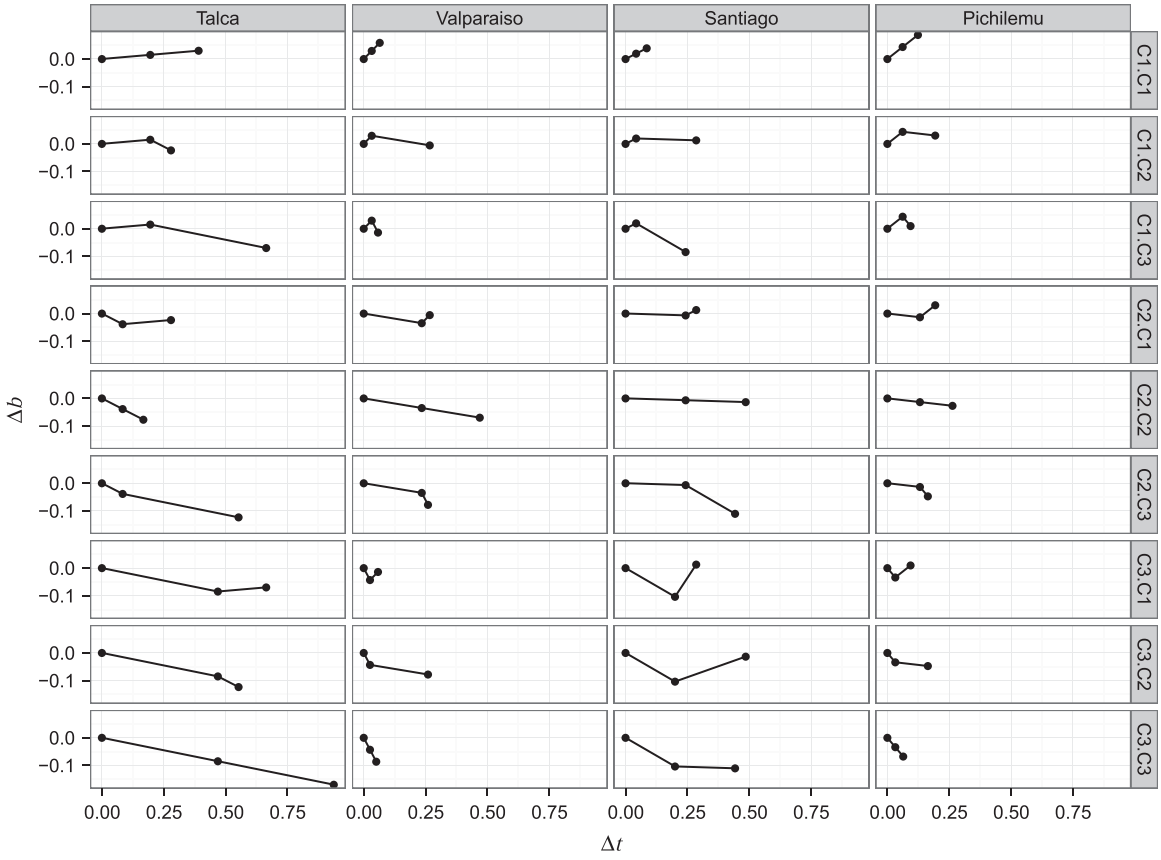

Fig. 6. Patterns discovered in all four cities.

Table 9
Result from the selected sequences.

Location	TP	TN	FP	FN	S_n	S_p
Talca	14	18	4	8	0.64	0.82
Pichilemu	62	19	3	11	0.85	0.86
Santiago	8	74	11	6	0.57	0.87
Valparaíso	28	131	17	24	0.54	0.89

5. Conclusions

The discovery of precursory patterns for earthquakes with magnitude larger than 4.4 in Chile has been addressed in this work. In particular, four of the most active Chilean zones have been studied. The metaheuristic used firstly transformed the raw data into a dataset containing new attributes, mainly based on the *b*-value. Then, a clustering algorithm is applied discretizing thus the transformed dataset. Later, an exhaustive search for all the sequences preceding an earthquake with magnitude larger than 4.4 is performed. From all the existing sequences some of them are selected as precursory candidates, which display some specific patterns (acute decreases of *b*-value as found in Morales-Esteban et al., 2010). The results obtained verged on 70% accuracy, confirming the usefulness of the proposed approach. Future work is directed to the discovery of similar patterns worldwide and to increasing the accuracy of the method, as well as using this information as input for supervised classifiers.

Acknowledgments

The authors would like to thank Spanish Ministry of Science and Technology, Junta de Andalucía and Pablo de Olavide University of Seville for the support under Projects TIN2011-28956-C00, P12-TIC-1728 and APPB813097, respectively. This work has also been partially funded by a Spanish Ramón y Cajal grant, 2012.

References

Bakun, W.H., Aagaard, B., Dost, B., Ellsworth, W.L., Hardebeck, J.L., Harris, R.A., Ji, C., Johnston, M.J.S., Langbein, J., Lienkaemper, J.J., Michael, A.J., Murray, J.R., Nadeau, R.M., Reasenberg, P.A., Reichle, M.S., Roeloffs, E.A., Shakal, A., Simpson, R. W., Waldhauser, F., 2005. Implications for prediction and hazard assessment from the 2004 Parkfield earthquake. Nature 437 (7061), 969–974.

Bakun, W.H., Lindh, A.G., 1985. The Parkfield, California earthquake prediction experiment. Science 229 (4714), 619–624.

Cárdenas-Jirón, L.A., 2013. The Chilean Earthquake and Tsunami 2010. A multidisciplinary Study of M_w8.8, Maule. WIT Press, UK.

Cisternas, M., Atwater, B.F., 2005. Predecessors of the giant 1960 Chile earthquake. Nature 437 (7057), 404–407.

Duncan, A., 2005. Earthquakes: future shock in California. Nature (19), 1476–4687.

Frankel, A., 1995. Mapping seismic hazard in the central and eastern United States. Seismol. Res. Lett. 66 (4), 8–21.

Gambino, S., Laudani, A., Mangiagli, S., 2014. Seismicity pattern changes before the $M=4.8$ Aeolian Archipelago (Italy) Earthquake of August 16, 2010. Sci. World J. ID531212:1–8.

Gutenberg, B., Richter, C.F., 1942. Earthquake magnitude, intensity, energy and acceleration. Bull. Seismol. Soc. Am. 32 (3), 163–191.

Gutenberg, B., Richter, C.F., 1954. Seismicity of the Earth. Princeton University, USA.

Holiday, J.R., Rundle, J.B., Tiampo, K., Klei, W., Donnellan, A., 2006. Systematic procedural and sensitivity analysis of the pattern informatics method for forecasting large ($M > 5$) earthquake events in southern California. Pure Appl. Geophys. 163 (11–12), 2433–2454.

Ishimoto, M., Iida, K., 1939. Observations sur les séismes enregistrés par le microsismographe construit dernièrement. Bull. Earthq. Res. Inst. 17, 443–478.

Ismail-Zadeh, A., Kossobokov, V., 2011. Earthquake Prediction, M8 algorithm. Encyclopedia of Earth Sciences Series, pp. 178–182.

Jordan, T.H., Chen, Y., Gasparini, P., Madariaga, R., Main, I., Marzocchi, W., Papadopoulos, G., Sobolev, G., Yamaoka, K., Zschau, J., 2011. Operational earthquake forecasting. State of knowledge and guidelines for utilization. Ann. Geophys. 54 (4), 315–391.

Lee, K., Yang, W.S., 2006. Historical seismicity of Korea. Bull. Seismol. Soc. Am. 71 (3), 846–855.

Martínez-Álvarez, F., Reyes, J., Morales-Esteban, A., Rubio-Escudero, C., 2013. Determining the best set of seismicity indicators to predict earthquakes. Two case studies: Chile and the Iberian Peninsula. Knowl. Based Syst. 50, 198–210.

Martínez-Álvarez, F., Troncoso, A., Morales-Esteban A., Riquelme, J.C., 2011. Computational intelligence techniques for predicting earthquakes. In: Lecture Notes in Computer Science, Part II, vol. 6679, pp. 287–294.

Mignan, A., 2011. Retrospective on the Accelerating Seismic Release ASR hypothesis: controversy and new horizons. Tectonophysics 505 (1–4), 1–16.

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.

Morales-Esteban, A., Martínez-Álvarez, F., Reyes, J., 2013. Earthquake prediction in seismogenic areas of the Iberian Peninsula based on computational intelligence. Tectonophysics 593, 121–134.

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

Nuannin, P., Kulhanek, O., Persson, L., 2005. Spatial and temporal *b*-value anomalies preceding the devastating off coast of NW Sumatra earthquake of December 26, 2004. Geophys. Res. Lett. 32 (11), 23–36.

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

Ranalli, G., 1969. A statistical study of aftershock sequences. Ann. Geofis. 22, 359–397.

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, 1314–1328.

Schwartz, D.P., Coppersmith, K.J., 1984. Fault behavior and characteristic earthquakes: examples from the Wasatch and San Andreas Fault Zones. J. Geophys. Res.: Solid Earth 89 (B7), 5681–5698.

Shi, Y., Bolt, B.A., 1982. The standard error of the magnitude–frequency *b* value. Bull. Seismol. Soc. Am. 72 (5), 1677–1687.

Sobolev, G.A., Tyupkin, Y.S., 1999. Precursory phases, seismicity precursors and earthquake prediction in Kamchatka. Volcanol. Seismol. 20, 615–627.

Stein, R., Toda, S., 2013. Megacity megaquakes—Two near misses. Science (341), 850–852.

Tiampo, K.F., Shcherbakov, R., 2012. Seismicity-based earthquake forecasting techniques: ten years of progress. Tectonophysics 522–523, 89–121.

Tucker, B.E., 2013. Reducing earthquake risk. Science 341 (6150), 1070–1072.

Utsu, T., 1965. A method for determining the value of *b* in a formula $log(n) = a - bM$ showing the magnitude–frequency relation for earthquakes. Geophys. Bull. Hokkaido Univ. 13, 99–103.

Wiemer, S., Wyss, M., 2002. Mapping spatial variability of the frequency–magnitude distribution of earthquakes. Adv. Geophys. 45, 259–302.

Yin, X., Chen, X., Song, Z., Yin, C., 1995. A new approach to earthquake prediction: the Load/Unload Response Ratio (LURR) theory. Pure Appl. Geophys. 145 (3–4), 701–715.

Zamani, A., Sorbi, M.R., Safavi, A.A., 2013. Application of neural network and ANFIS model for earthquake occurrence in Iran. Earth Sci. Inf. 6 (2), 71–85.

Zhu, F., Wu, G., 1977. Prediction of the Haicheng earthquake. Trans. Am. Geophys. Union 58 (5), 236–272.