Detection of Seismic Dual-Zones with Application to Earthquake Prediction

A. Bali, M. Zaré & A. Andalib
International Institute of Earthquake Engineering and Seismology, Iran

15 WCEE
LISBOA 2012

SUMMARY
We introduce an efficient approach to manage the forecasting of long-term, mid-term and short-term earthquakes which will be assumed to have magnitudes higher than a threshold level. Dividing the entire global plane into well-defined sub-regions (zones), this method creates an "event matrix" whose different cells correspond to different spatial-temporal seismic attitudes, with each cell identifying the total number of events occurred in that sub-region within that specified period of time. The event matrix is then searched for similarities among rows by means of statistical methods. If the similarity measure between two rows is more than a specific threshold level, then the corresponding sub-regions are referred to as "dual zones". While a straightforward search algorithm is computationally extensive, we use a numerical algorithm instead, known as sparse matrix computational technique. To test the method, the world's seismicity catalogue is received and divided to two parts for searching similarities and performance evaluation.

Keywords: Seismic probability, dual zone, precursor earthquake, sparse matrix, forecasting.

1. INTRODUCTION

Geologists believe that the Earth is a complex system and its physical parameters happen to show various nonlinear, chaotic and stochastic behaviors (Plagianakos and Tzanaki 2001) many of which yet to be discovered and some of them not totally justified thus far.

Let's assume the Earth as a system which follows the global behavioral patterns (Keilis-Borok and Soloviev 2003), that any variation observed in one zone is related to some other specific areas of this network. In a causal system like this, a relative pattern can be achieved by analyzing the current system status and then comparing the results to its situation in the past. This procedure would finally pave the way to describe the seismic behavior of the earth and the probable inter-dependencies among the earthquakes.

In this research, we are looking forward neither to discovering any global basis nor to identifying any global connection amongst the structural parts of this complex network; However, we do strongly wish to come by an outcome which can help us to better understand the diverse effects of the diverse origins of seismic activities via following the revolutions already happened with gigantic earthquakes in different areas of our world. In this process, once some certain and relatively straight rules are achieved through observations, then we can hope that further events analysis will be done more easily and reliably such that the interpretation on the causes and description of the details will more truly assist our later cases on more reliable forecasting.

Therefore, the main goal of this paper is to locate the imminent short-term quakes likely to happen in the future, based upon a new method that is comparing the data already achieved by the past records in order to decide on the future. We call these locations as dual zones (duals) to the reference zone. We also named our suggested method as Bali-Zaré Earthquakes Forecasting Method (BZEFM). This method, which is introduced in Bali (2010) in detail, can be used to evaluate occurrence pattern of the events. The objective of the current research is neither to represent a new mechanism of the process, nor to express the relationship between two subsequent earthquakes -happened in different areas of the World- in terms of plate tectonic theory. However, this subject can also be studied further in another survey. Here, and by means of statistical analysis, we try to find the probable spatial-temporal dependencies for earthquake events globally, including different regions and different time intervals. In fact, we describe a new method for finding dual zones where an earthquake in one zone acts as a precursor to other event(s) in some other zone(s).

In the next sections of this paper, we introduce our method. In Section 3, we provide the experimental results and practically prove the validity of this new technique by means of several reliable simulations. Section 4, concludes the work.

2. BZEFM APPROACH

Let E indicate the set of vectors $E = [a_1, a_2, \dots, a_i, \dots, a_n]$, where a_i with $i = 1, 2, \dots, n$ is the vector corresponded to the i'th time interval of the catalogue, and $n = (T_F - T_S)/\tau$. Also, T_S and T_F show the starting and ending date of the events recorded in the earthquake catalogue, respectively. τ is a desired constant parameter, which indicates the length of the time interval allocated to each vector a_i. According to BZEFM, the global map is divided into m cells with $\hat{l} \times \hat{l}$ degrees in size, and the event centers are at $\pi_j, j = 1, 2, \dots, m$. Then, only those events from the earthquake catalogue which have magnitudes greater than or equal to a threshold level $(M \geq M_T)$ and which have occurred within the desired time interval i, will be taken into consideration. The vector a_i includes $a_i = [n_1^i, n_2^i, \dots, n_j^i, \dots, n_m^i]^T$ where n_j^i is the number of earthquakes larger than M_T in the j-th zone of π_j and the i-th time interval, and $[\cdot]^T$ is the same vector, yet transposed. Therefore, the matrix E named here as "spatial-temporal matrix of earthquake events" or simply "event matrix", has m rows and n columns, corresponding to the seismicity of m different cells at n different time intervals. This event matrix E might be referred to as E^τ, should the need rise in order to emphasize the length of time intervals which are considered to set up the event matrix. Fig 2.1 shows a typical event matrix. Our method uses such a matrix to extract the useful yet latent spatial-temporal patterns.

In this manner, one can study different attributes of the matrix from two different aspects: 1) row (cell)-based and 2) column-based approaches; both of which could be used efficiently to predict the probability of an imminent earthquake event in a specific geographical zone in the future

Row-based approach: this method makes use of the recorded data related to the seismic behavior of one "specific zone", e.g., Bandar-e-Abbas, Iran within "different time intervals", e.g., as of the year 1973 through 2009,

Column-based approach: unlike the first procedure, here one should use the available dataset in order to extract items of information recorded from a specific time event (e.g., time interval of Jan. 1973 through the end of March in the same year) around the world or within "several different zones".

seismicity of one special cell in different time

$$E = \begin{bmatrix} 1 & 2 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

seismicity of different cells in a special time

Figure 2.1. A typical 9-by-5 event matrix.

By means of such a big two-dimensional spatial-temporal dataset, which has been embedded throughout the event matrix, there is a high chance for scientists to take a new analytical look at the recorded seismic behaviors, which is the main goal of the approach introduced in this paper. For this to happen, each and every row of the matrix is compared one by one to the other rows conveying other items of information related to the past events, and this process continues until every desired entry of this matrix has been thoroughly compared to all the given entries from the other rows. Therefore, one row of the matrix that represent the seismic activity of one specific city, say, Bandar-e-Abbas, Iran, within different time intervals is picked up as the "target row" and then, the data embedded in the rest of the rows will be compared to the target row. This comparison takes effect via the seismic difference measure A

$$A_{target,j} = \sum_{k=1}^{n} \left(E_{target,k} - E_{j,k}\right)^2, (j = 1,2,\cdots,m), \tag{2.1}$$

where k indicates time, $E_{target,k}$ and $E_{j,k}$ indicate respectively a specific entry k from the target row, and k-th entry from the j-th row of the spatial-temporal event matrix. This way $A_{target,j}$ or shortly put $A_{t,j}$ gives a measure for comparing the seismic similarities between the target city and another city represented by the j-th row; in such a way that the less the $A_{t,j}$, the higher the similarity of occurring the events between the two cities. By similarity, we mean simultaneous seismically-active time intervals with similar number of events higher than the threshold. Meanwhile, there is also another parameter referred to as measure B which indicates the seismic stillness similarity and is defined as follows

$$B_{t,j} = \frac{\alpha_{t,j}}{2n}, (j = 1,2,\cdots,m), \tag{2.2}$$

where $\alpha_{t,j}$ is the number of non-zero entries in the two rows t and j. Hence, $B_{t,j}$ shows the ratio of none-zero entries to the total number of rows in a matrix; e.g., if each row includes 30 entries, then comparing the two rows results in the assessment of data conveyed through the entire sum of 60 entries. Now, if only 50 entries out of the rest, do have none-zero values, then B will be equal to $1/4 = 0.25$. Concerning the explicit definitions for A and B it is concluded that as far less and far more become the two measures A and B, then the seismicity behavior of the studied rows will be much more alike, and the amount of positive correlation between the two rows will also grow up. Therefore, the third measure C is obtained by combining the two previous definitions for A and B, as follows

$$C_{t,j} = \frac{B_{t,j}}{A_{t,j}}, (j = 1,2,\cdots,m-1, j \neq t). \tag{2.3}$$

Needless to mention, the bigger values for $C_{t,j}$ imply higher similarity between the two rows, which represent the events similarities between the target city and the other one under consideration. The descending vector $C_t = [C_{t,1}, C_{t,2}, \cdots, C_{t,m-1}]$ indicates the similarity amongst all different cells regarding the target cell. Here, we use the vector C_t to define some of main concepts discussed in this paper

Definition Given an event matrix E with the target cell t, the j-th cell with a value $C_{t,j}$ greater than a specific threshold level is referred to as a "seismic dual zone to the target zone", or a "dual", in brief.

Definition Given an event matrix E^τ, if the target row t is entirely shifted by an amount of $\tau' = a\tau, (a = \pm1, \pm2, \pm3, \cdots)$, and if afterward it happens for the j-th cell to make a dual to the newly shifted target cell, then the j-th cell is referred to as a "seismic dual to the target zone with a time-shift of τ'", or a "dual with a time-shift of τ'", in brief.

Definition Given an event matrix E^τ, the j-th cell is a "precursor" of the target cell, if the j-th cell is a dual to the target zone with a time-shift of $\tau' = a\tau$, and $a > 0$.

Definition Given an event matrix E^τ, the j-th cell is a "postcursor" of the target cell, if the j-th cell is a dual to the target zone with a time-shift of $\tau' = a\tau$, and $a < 0$.

Definition If in the event matrix $E^{\tau_1}_{m\times n}$, the cell A is known as a dual to the cell B with a time-shift of $\tau'_1 = a_1\tau_1$, and if in the event matrix $E^{\tau_2}_{m\times n'}$, the cell B is known as a dual to the cell A with a time-shift of $\tau'_2 = a_2\tau_2$, then the cells A and B are "resonant duals to each other, with respective time-shifts of τ'_1 and τ'_2", or as "resonant duals, with time-shifts of τ'_1 and τ'_2", in brief.

It should be mentioned that the procedure of finding dual zones includes an exhaustive search (brute-force search) algorithm, i.e., for each reference zone, all possible candidates must be enumerated systematically to find the possible dual zone(s). Here, we have employed special data mining algorithms to manage the computational complexity of the problem.

3. EXPERIMENTAL RESULTS AND VALIDITY OF BZEFM METHOD

In order to experiment the performance of the new method and to test its validity, first the world's seismicity catalogue was received for the time span 1.1.1973 through 6.30.2010 from USGS/NEIC. This catalogue is publicly accessible from the website of American National Earthquake Information Center (http://earthquake.usgs.gov/regional/neic/). The data extracted from the interval 1.1.1973-6.30.2009 was then used to look for any dual zones, and the last year's period of time, i.e. the interval 6.30.2009-6.30.2010 was considered to serve as an interval to validate the duality which were determined already. In this paper, we refer to these two periods as identification and validation periods, respectively. Here, the world's map or the global surface is segregated into cells of 2-by-2 degrees in size, and events with magnitudes greater than the threshold level 5.5, which have occurred within the desired time interval in every specific cell, have been taken into account. Regarding the 2-by-2 division of the map, the ensuing spatial-temporal matrix of earthquake events would convey 16200 rows (cells). Meanwhile, regarding the starting and ending dates of the catalogue, as well as the three monthly divisions, this matrix is expected to include 152 columns presenting the 152 seasons. In order to avoid any redundant zero values as for the entries of the event matrix, and also to ensure that the algorithm would be acting correctly, those cells with less than 3 events with magnitudes $M \geq 5.5$ within the previous 38 years, were fairly indicated as *seismologically inactive areas* and hence were removed. Via this strategy, the entire number of rows representing the earthquake events were largely reduced from 16200 into 398; a fact that in turn influenced the procedure with a significant reduction in the computational complexity for the algorithms.

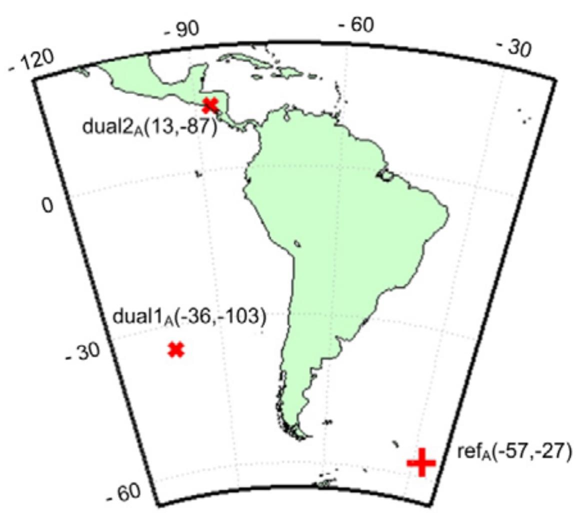

Figure. 3.1. The positions of the reference region A (-57,-27), and its first best and second best dual zones.

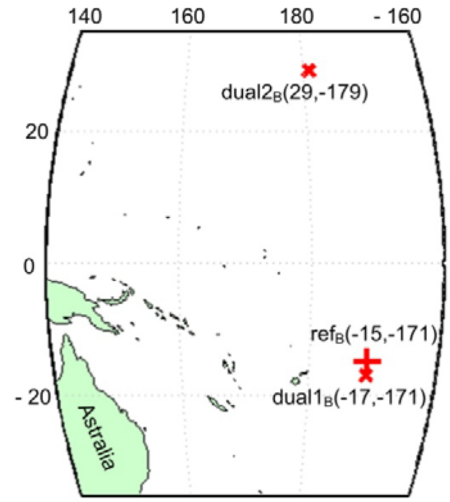

Figure. 3.2. The positions of the reference region B(-15,-171) and its first best and second best dual zones.

Following this, values of parameters A, B, and C were calculated for each reference cell and then, the outcomes were sorted out in accord with the observed maximum probabilities. This way, some cells of the matrix happened to demonstrate noticeable spatial correlations. However, a more startling point appeared when these dual cells which were identified in time intervals prior to 6.30.2009 showed similar attitudes within the evaluation time interval and this could further assert the accuracy of the approach introduced by new method. Table 3.1 demonstrates the performance of the new method over dual zones of two reference regions A (lat:-57, long:-27) and B (lat:-15, long:-171). The positions of the reference cell A and its dual zones, i.e., $dual_{1A}$, and $dual_{2A}$ are depicted in Fig. 3.1. Similarly, the positions of cell B and its dual zones, i.e. $dual_{1B}$ and $dual_{2B}$, are depicted in Fig. 3.2. The high percent of duality for the set of cells A and B, indicates that an earthquake event in each of the dual cells may be considered as an alarm for the corresponding reference cell. It is also interesting that the dual zones of the reference cell A are located far from one another.

Following this step, and in order to make use of the event matrix as a tool conveying the precursory data for the future events, the reference row basically needs to get a 1-coloumn time shift (i.e. a 3-month time interval) at a time, and the newly-shaped shifted matrix row will then be compared to the rest of the rows. For instance, according to the results of the new algorithm applied to the seismic catalogue's data gathered until 6.30.2009, it was revealed that the 2-by-2 cells with longitudinal and latitudinal degrees (lat:-23, long:-177) as for the center point, served well as a precursor for a second cell with coordinates (lat:17, long:-101). Fig 3.3 depicts how these two cells are taking a geographical stand before each other. Other successful precursory cases as for the succeeding events of a 3-month interval imminence (expected to occur within the target cell) are shown in Table 3.2. It is noteworthy that the target cell, here, is located on the Pacific Ring of Fire in the neighborhood of Mexico City, Mexico.

Strictly speaking, these cases are the direct hints for calling the two mentioned areas as dual zones with a time difference of 3 months. In other words, the reference cell is fairly considered a precursor to the second cell.

In order to evaluate the algorithm of the new method, the items of information extracted from the last year part of the catalogue which shows off in the two cells mentioned above have been examined within the evaluation time interval.

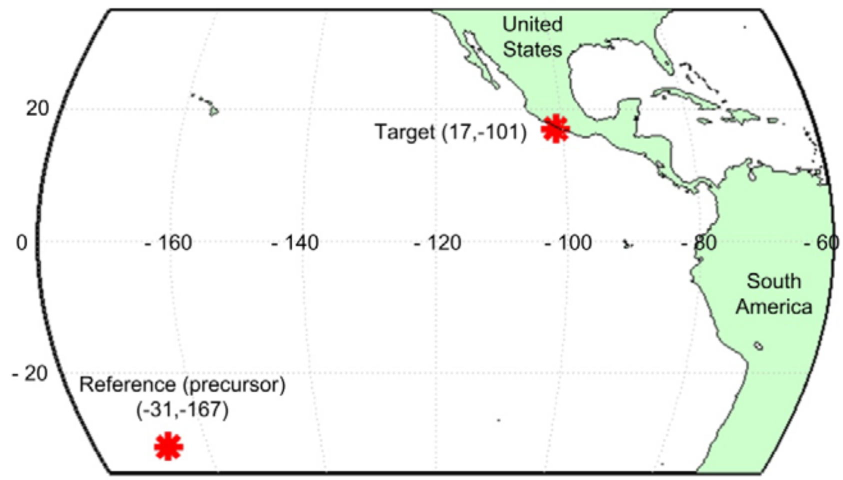

Figure 3.3. The positions of the reference (precursor) cell (-31,-167), and the target cell (17,-101).

According to this procedure and in regard to an earthquake event with magnitude of 6 (in Richter scale) dating back to 2.22.2010 which has occurred in the reference cell, another earthquake event with a magnitude greater than 5.5 has simply been anticipated within a 3-month interval as of 3.1.2010 to 6.1.2010. This prediction did prove to be true since an earthquake of 5.5 Richter occurred in 4.14.2010 within the second cell.

It is worth to mention that the reference cell for this survey included some events with magnitudes of 5.5 and bigger, within the specific time interval; and part of them, got to serve successfully as a precursor to the second cell. In fact, some events fail to act as a precursor, because they were very slightly out of the specific time interval. Therefore, if we choose the boundary of the cells fuzzy, it is expected to achieve more promising results. This may be the subject of further research on the method presented in this paper for finding seismic dual zones.

Table 3.1. Performance Over Identification and Validation Periods for Some Dual Cells.

Reference cell	ref_A (-57,-27)	ref_B (-15, -171)
First best cell in duality	$dual1_A$(-35, -103)	$dual1_B$ (-17,-171)
Second best cell in duality	$dual2_A$ (13, -87)	$dual2_B$ (29, -179)
Rate of contemporary events in the three cells (identification period)	38%	38%
Rate of contemporary events in the three cells (validation period)	75%	96%

Table 3.2. An instance of a successful precursory case for the succeeding events of a 3-month interval imminence in the target cell, found by BZEFM.

	Reference (precursor) cell: (-23,-177)	Target cell: (17, -101)
Identification period	4.8.1973-$M = 5.5$	7.16.1973-$M = 6.2$
	11.11.1974-$M = 5.6$	2.22.1975-$M = 5.8$
	2.14.1976-$M = 5.9$	6.7.1976-$M = 6.7$
	1.13.1988-$M = 5.7$	2.8.1988-$M = 5.8$
	4.11.1998-$M = 6.2$ 4.11.1998-$M = 5.5$ 4.12.1998-$M = 5.5$	7.11.1998-$M = 5.5$ 7.12.1998-$M = 5.5$
Validation period	2.22.2010-$M = 6$ 3.18.2010-$M = 5.6$	4.14.2010-$M = 5.5$

In another experiment, we reduced the size of cells to 1-by-1 degrees and obtained an event matrix with 64800 rows and 152 columns. To challenge the complexity problem again, we removed those rows of the matrix with less than five events $M \geq 4$. This way, we reduced the rows of the matrix to 4714. We then searched thorough the matrix to find the dualities. We found at least three dual cells for 4648 target cells out of 4714 (98.6%) over the identification period from 1.1.1973 to 6.3.2009. Using these dual cells, we managed to do successful forecast over evaluation period, 6.30.2009-6.30.2010. This period includes 591 earthquakes with $M \geq 5.5$, which are happened in 272 days. Considering the 1-by-1 geographical divisions and the 3-month time divisions, BZEFM may raise alarm in $360 \times 180 \times (12/3)$ occasions, corresponding to the elements of the event matrix on validation period. This sparse matrix only has 438 nonzero elements.

Here, the performance of the algorithm is presented using the well-know matrix of confusion (Provost *et al.* 1998, Provost and Fawcett 2001). This matrix provides the rate of True/False forecasts for 192600 Positive/Negative occasions. BZEFM correctly forecasts $d = 283$ earthquakes with $M \geq 5.5$, and a=191348 non-active elements of the event matrix. The model also incorrectly raises an earthquake alarm for $c = 1252$ cells, and also misses $b = 155$ events. The resulting confusion matrix is shown in Table 3.3, from which some standard statistical measures may be concluded. The BZEFM *accuracy* defined as the proportion of the total number of predictions that were correct is $(a + d)/(a + b + c + d) = 0.9926$. The *recall* or *true positive rate (TP)* as the proportion of positive cases that are correctly identified is $d/(c + d) = 0.1844$. Similarly, the *true negative rate* (*TN*) is defined as the proportion of negatives cases that are classified correctly: $a/(a + b) = 0.9992$. Finally, *precision* (P) is the proportion of the predicted positive cases that are correct: $d/(b + d) = 0.6461$.

Table 3.3. Confusion Matrix for Forecasting earthquakes with $M \geq 5.5$ over the evaluation period, 6.30.2009-6.30.2010, Using BZEFM.

		Prediction Outcome	
		Negative	Positive
Actual Value	Negative	True Negative: a=190910	False Negative: b=155
	Positive	False Positive: c=1252	True Positive: d=283

The BZEFM seems to have great accuracy, however, it may not be an adequate performance measure (Provost and Fawcett 2001), because the number of negative cases, in this experiment is much greater than the number of positive cases. In fact, there are 191065 negative cases out of 192600 elements of the event matrix. If the model classifies them all as negative, the accuracy would be 99.77%, even though the classifier missed all positive cases. Therefore, other performance measures should be employed, e.g., *geometric mean* g_{mean}, as defined by Kubat and Matwin (1997) as

$$g_{mean1} = \sqrt{TP.P}, \qquad g_{mean2} = \sqrt{TP.TN}, \tag{3.1}$$

and $F\_measure$, as defined by Lewis and Gale (1994) as

$$F = \frac{(\beta^2 + 1).P.TP}{\beta^2.P + TP}, \tag{3.2}$$

where β is a value from 0 to infinity and is used to control the weight assigned to *TP* and *P*.

Any classifier evaluated using Eqn. 3.1 or Eqn. 3.2 or will have a measure value of 0, if all positive cases are classified incorrectly. Here, $g_{mean1} = 0.3452$, and $g_{mean2} = 0.4292.$ For $\beta = 1$, which assigns equal weights to *precision* and *recall*, $F\_measure = 2P.TP/(P + TP) = 0.2869$.

Table 3.4. Standard statistical measures for BZEFM results, forecasting earthquakes $M \geq 5.5$ over the evaluation period, 6.30.2009-6.30.2010.

accuracy	99.26%
true positive rate	18.44%
true negative rate	99.92%
precision	64.61%
g_{mean1}	0.3452
g_{mean2}	0.4292
$F\_measure$	0.2869

The above statistical measures are listed in Table 3.4. These values indicate the acceptable forecasting power of BZEFM in our experiment.

4. CONCLUSION

The main purpose of this survey was basically to introduce a new approach for finding spatial-temporal seismic dependencies around the world. However, the results may be used for short-term, mid-term, and long-term earthquake precursory.

In this regard, some significant questions arose and some interesting views ensued for which to be solved the researcher may come up with new ideas or even theories, and those new ideas may simply make the topics for further analysis and future research, as well.

Two main ideas that go along with this survey are as following:

Those areas which follow one another as precursors and post-cursors are referred to as seismic dual zones. Most of these zones are located not nearby but fairly distant from one another. Explaining the results in the framework of plate tectonics theory may be considered as the objective of another research paper.

In some cases, it happens for a particular cell to initially be a precursor to one specific zone within a specific time interval, and as time passes it eventually turns out to be a precursor to another zone within a different period of time. Changes of this kind will need to be scrutinized.
Finally, it should be noticed that reducing the size (in latitudinal/longitudinal degrees) of the cells and also reducing the threshold level of magnitude are of high significance. Both cases were paid due attention during the complementary phase of this project and the algorithm of this paper was applied to them.

We have used the idea presented here, in another paper for earthquake forecasting. This method, namely as Bali-Zaré earthquake forecasting method (BZEFM) employs the information of seismic dual zones to provide alarms for an upcoming event.

REFERENCE

Plagianakos, V. P. and Tzanaki, E. (2001). Chaotic analysis of seismic time Series and short term forecasting using neural networks. *the IEEE International Joint Conference on Neural Networks, In Proc.*, **3**, 1598-1602.

Keilis-Borok V. I. and Soloviev, A. A. (Eds.) (2003). Nonlinear dynamics of the lithosphere and earthquake prediction, Springer-Verlag, Heidelberg, ch. 1.

Bali, A. (2010). Probabilistic model of earthquake forecasting based on earthquake catalogue of Iran, Ph.D. thesis, International Institute of Earthquake Engineering and Seismology, Tehran, Iran.

Provost, F. and Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning*, **42(3)**, 203-231.

Provost, F., Fawcett, T., and Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. *in Proceedings of the 15th International Conference on Machine Learning*, 445-453, Madison, WI. Morgan Kauffmann.

Kubat, M., and Matwin, S. (1997). Addressing the curse of imbalanced training sets: one sided selection. *in Proceedings of the 14th International Conference on Machine Learning*, 179-186, Nashville, Tennesse. Morgan Kaufmann.

Lewis, D., and Gale, W. (1994). A sequential algorithm for training text classifiers", *in Proceedings of ACM-SIGIR 1994*.