Earth Planets Space, **61**, 1223–1232, 2009

Earthquake hazard assessment in seismogenic systems through Markovian artificial neural network estimation: an application to the Japan area

C. Herrera[1] and F. A. Nava[2]

[1]*UABC, San Quintín, B.C., Mexico*
[2]*CICESE, Seismology Department, Ensenada, B.C., Mexico*

(Received March 27, 2008; Revised June 12, 2009; Accepted July 17, 2009; Online published December 21, 2009)

An earlier work (Herrera *et al*.: *Earth Planets Space*, **58**, 973–979, 2006) introduced two new methods for seismic hazard evaluation in a geographic area with distinct, but related, seismogenic regions. These two methods are based on modeling the transition probabilities of *states*, i.e. patterns of presence or absence of large earthquakes, in the regions, as a Markov chain. This modeling is, in turn, based on a straightforward counting of observed transitions between states. The *direct* method obtains transition probabilities among states that include events with magnitudes $M \geq M_r$, where M_r is a specified threshold magnitude. The *mixed* method evaluates probabilities for transitions from a state with $M \geq M_r^{\mathrm{m}}$ to a state with $M \geq M_r^{\mathrm{M}}$, where $M_r^{\mathrm{m}} < M_r^{\mathrm{M}}$. Both methods gave very good results when applied to the Japan area, with the mixed method giving the best results and an improved magnitude range. In the work presented here, we propose other methods that use the learning capacity of an elementary neuronal network (perceptron) to characterize the Markovian behavior of the system; these neuronal methods, *direct* and *mixed*, gave results ∼7 and ∼6% better than the counting methods, respectively. Method performance is measured using grading functions that evaluate a tradeoff between positive and negative aspects of performance. This procedure results in a normalized grade being assigned that allows comparisons among different models and methods.
Key words: Probabilistic seismic hazard assessment, neural networks, Markov chains.

1. Introduction

The term seismic hazard denotes the probability of occurrence of earthquakes in a given time, space, and magnitude range. Seismic hazard assessment is one of the main goals in seismology because it is a key factor in a correct and useful assessment of seismic risk which, in turn, can diminish the social, economic, and political impacts of the devastating effects caused by large earthquakes.

Seismic hazard estimations for large earthquakes can be made from deterministic earthquake cycle models, from a purely statistical analysis of seismicity, or from stochastic models such as ours (Herrera *et al*., 2006) where we combine a very simple physical model (embodied in the concept of the *system* we use) with a statistical Markov chain analysis of seismicity in the system. The main problem with hazard estimations involving statistical analysis is the comparatively short span of available seismic catalogs compared with the relatively longer average recurrence times of large earthquakes. Thus, in order to test a seismic hazard estimation method, the researcher needs a seismogenic region with frequent large earthquakes and a reliable seismic catalog, i.e., a region such as Japan.

The Japan area has a particularly high level of seismic activity and has often experienced large and destructive earthquakes. The tectonic regime of the Japan and surrounding areas is a very complex system, with seismicity and faulting related to the continuous NW motion on the NE and SW Japan arc systems. Seismicity in Japan is categorized as intraplate or interplate depending on its tectonic origin. Intraplate events are shallow events that occur on land, while interplate events occur along major subduction zones, such as those between the Philippine, Pacific, and North American plates. In Japan, earthquake recurrence for intraplate events is much longer than that for interplate events (Shimazaki, 1976). Many large interplate earthquakes have occurred in the Tohoku district of northern Japan, suggesting strong seismic coupling on the plate boundary (Kanamori, 1977). Ito *et al*. (1999, 2000) applied an inversion analysis of GPS measurements in order to find the spatial distribution of the interplate coupling in NW and SE Japan. Large earthquakes have occurred repeatedly along the Nankai through (Thatcher and Rundle, 1984). This high level of seismic activity is of major concern, emphasizing the need for earthquake preparedness and an updated seismic characterization of the area.

A variety of probabilistic models have been proposed for seismic hazard assessment in the Japan area. The well-known time-predictable recurrence model of Shimazaki and Nakata (1980) is based on earthquakes and morphological data from Japan. Utsu (1984) presented results from their application of four renewal models—Weibull, gamma, lognormal, and the double exponential probability density—for the recurrence of earthquakes observed in several seismic regions in Japan. In a number of cases, the lognormal

Copyright © The Society of Geomagnetism and Earth, Planetary and Space Sciences (SGEPSS); The Seismological Society of Japan; The Volcanological Society of Japan; The Geodetic Society of Japan; The Japanese Society for Planetary Sciences; TERRAPUB.

model gave the best results. In his comprehensive review of earthquake prediction efforts in Japan, Mogi (1985) noted that recurrence times for the Tokyo area are exponentially distributed. Wyss *et al*. (2007) and Katsumata and Kasahara (1999) used the statistics of seismicity to evaluate seismic quiescence before the Izu-Oshima 1990 and the Kurile 1994 earthquakes, respectively. Other efforts focused on improving earthquake prediction in Japan are described in Hirata (2004). The model we propose here differs from the abovementioned models mainly in that we introduce the concept of *system* and explore the possibility of it having a Markovian behavior.

Many of the models used for hazard estimation are based on the analysis of the seismic history of a given area; these include hazard analysis from recurrence-time estimates based on the Gutenberg-Richter distribution (Gutenberg and Richter, 1944), the numerous models based on Poissonian seismicity models (e.g,, Brillinger, 1982; Lomnitz and Nava, 1983), which model the earthquake recurrence as an entirely random process, and those from Bayesian methods (e.g., García-Fernández and Egozcue, 1989; Rüttener *et al*., 1996), which provide a mathematical model to estimate the distribution of random variables in the presence of uncertainties and the semi-Markov process applied to linear zones (Patwardahan *et al*., 1980). Almost all seismic hazard models including a causal physical component are based, directly or indirectly, on the elastic-rebound model (Reid, 1910; Richter, 1958). Among these models are those for recurrence times based on seismotectonic arguments, such as the time predictable or slip-predictable models of Shimazaki and Nakata (1980), those based on the seismic gap concept (e.g., Fedotov, 1965; McCann *et al*., 1979; Kagan and Jackson, 1991), and those based on seismic migration (e.g., Richter, 1958; Mogi, 1968). However, these determinations are not reliable enough, given the large number of implicit suppositions and unknown factors they contain, over which there is as yet no control. New methods are needed to overcome the limitations of these conventional methods.

The artificial neural networks (*ANN*) approach has recently been shown to have an enormous potential for solving a variety of problems in various fields, such as image and signal processing (Kashyap, 1976; Mcllraith and Card, 1997), civil, electrical, and mechanical engineering (Chao and Skibniewski, 1994; Karunanithi *et al*., 1994), and seismology (Zhao and Takano, 1999). This approach does not depend upon any assumptions on the distribution of the data, it is capable of handling data having different precision levels, and it has a rapid data processing capability (Dowla *et al*., 1990).

Nava *et al*. (2005) and Herrera *et al*. (2006) proposed statistical methods for seismic hazard evaluation based on modeling the transition probabilities between *states*, i.e., the geographical patterns of occurrence or non-occurrence of large earthquakes in different regions of a given geographic area during a time interval, as a Markov chain. These methods were based on the straightforward counting of observed transitions between states. The *direct* method obtained transition probabilities between states, where both initial and final states consider events with magnitudes $M \geq M_r$, where M_r is a specified threshold magnitude. The *mixed* method evaluates probabilities for transitions from a state with $M \geq M_r^{\mathrm{m}}$ to a state with $M \geq M_r^{\mathrm{M}}$, where $M_r^{\mathrm{m}} < M_r^{\mathrm{M}}$. The motivation for the mixed method is that events with large magnitudes are relatively scarce, so that adequate sampling of large events is severely limited by the length of existing catalogs. This method explores the possibility of smaller magnitudes yielding information on the occurrence of larger magnitude earthquakes. If the threshold magnitude is too large, almost all observed states become the zero state (no activity at all), and forecasting becomes trivial but useless. By using different threshold magnitudes, we extend the forecasting power of the direct method to higher target magnitudes. Both methods gave very good results when applied to the Japan area, with the best results and an improved magnitude range being obtained using the mixed method (Herrera *et al*., 2006).

The form in which the matrices for the Markovian transition probabilities are constructed in the straightforward counting method, with each new considered transition modifying the previous probability estimates, can be thought of as a learning process. Therefore, we considered the possibility of applying other learning methods to the Markovian seismic hazard evaluation problem, and the *ANN* were natural candidates for this objective. Here, we report on the simplest type of neural network, called *perceptron*. Our analysis revealed that the results from perceptron learning are better than those from the counting methods. The assessment criteria are discussed in the following sections.

2. System Seismic Hazard (Review)

We define a *system* as a geographic area that includes R seismogenic regions. Given a seismic catalog and a starting time during each successive time interval of Δt, the state of each region s_r has one of two values, 0 or 1, corresponding, respectively, to absence or presence of earthquakes with a magnitude larger or equal than some threshold value M_r. The total state of the system s is the sum of the regional states:

$$s = \sum_{r=0}^{R-1} 2^r s_r, \tag{1}$$

and there are $S = 2^R$ possible system states. In binary format, s is simply the concatenation of the binary regional states; it ranges from 00...00 to 11...11 and shows at a glance which regions have earthquakes and which have not for each state. Thus, the *system seismic hazard* is the probability of the system having a given state during a particular interval. In the following sections, we heuristically propose that the system is Markovian and assess whether the assumption is correct by comparing the results with those of memory-less models, such as the uniform or Poissonian ones.

3. The Perceptron and State Coding

A Markovian model, for which the state probability over a time interval depends only on the immediately previous state, allows use of the simplest type of *ANN*, the *perceptron*, which consists of a set of input units $[z] = \{z_i; i =$

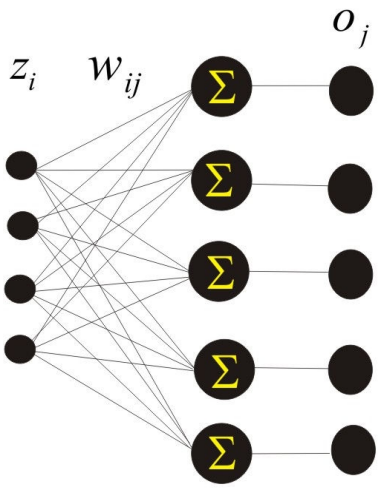

Fig. 1. Perceptron. The circles with the summation sign represent the neurons.

$1, 2, ..., n\}$, followed by a layer of *neurons* that calculate a weighted sum of the inputs and a set of output units $[o] = \{o_j;\ j = 1, 2, ..., n\}$ where the result of the computation is read off (Fig. 1).

To the input $[z]$, corresponds the output $[o]$, whose jth element is

$$o_j = f\left(\sum_{i=1}^{n} w_{ij}\, z_i + b_i\right), \tag{2}$$

where the w_{ij} are weighted paths connecting each input unit to each node, and f is an activation function that determines the form of each output. A constant input $[b]$, known as the "bias", is frequently added to the set of weights in order to give greater stability to the system.

The weights are chosen so that, for a given input, the network produces a given desired output, but these weights generally are not known *a priori*. The most important characteristic of an *ANN* is that it can learn, i.e., given pairs of input/target output values, the perceptron can use a learning scheme, such as the simple Hebb's rule (Rosenblatt, 1958, 1962), based on the residuals of target minus actual outputs, to modify an initial set of weights and thus achieve a better match to the desired output:

$$\begin{aligned} w_{ij}^{k+1} &= w_{ij}^{k} + \Delta w_{ij}^{k} \\ \Delta w_{ij}^{k} &= \eta\left(d_j^k - o_j^k\right) z_i^k, \end{aligned} \tag{3}$$

where η, the *learning rate*, is a damping factor that controls how strongly the residuals modify the weights.

For our Markovian system, we will ask the perceptron to output the state for interval $j+1$, as response to the input of the state for interval j. Thus, each couple of successive system states (S_j, S_{j+1}) corresponds to an input/target output pair.

For the perceptron to function correctly, we need an orthogonal representation for the various states of the system. Accordingly, each state will be coded as a binary word of S bits, with only bit number $j+1$ (where j corresponds to the state number) different from zero. We will denote a state thus coded by z. For example, for a system with four regions and $S = 16$, state 10 (activity in the regions 1 and 3, $s = 1010$) is coded as $z = 0000000000100000$. This orthogonal representation ortho-normalizes the states in S dimensions.

4. Markovian Perceptron: Direct and Mixed Methods

From the catalog, given t_0 and Δt, we tried both the direct and the mixed methods described above.

For the direct method, one list of coded states z_j is obtained for $M \geq M_r$, and the perceptron is made to learn from input-output pairs $\{z_j, z_{j+1}\}$.

For the mixed method, two lists of coded states are used—z_n^{m} for $M \geq M_r^{\mathrm{m}}$, and z_n^{M} for $M \geq M_r^{\mathrm{M}}$, where $M_r^{\mathrm{m}} < M_r^{\mathrm{M}}$—and the perceptron is made to learn from input-output pairs $\{z_j^{\mathrm{m}}, z_{j+1}^{\mathrm{M}}\}$, so that once the perceptron has been trained, the current system state for $M \geq M_r^{\mathrm{m}}$ can be input to obtain a forecast of the coming state for $M \geq M_r^{\mathrm{M}}$.

5. Parameter Choice

A first, tentative choice of the system parameters is made so as to obtain optimal definition and coverage within the limits set by the catalog length (and, in some degree, by its accuracy). This choice is conducted in manner described below.

The seismic hazard spatial bounds are the boundaries of interrelated seismogenic regions, which together constitute a *system*. The term interrelated regions denotes regions which may be considered separate from a tectonic or structural point of view, yet are close enough to each other so that the stress changes caused by large earthquakes and the corresponding plate motions in one region may influence seismicity in other regions. Because of location uncertainties, regions are separated by "no man's land" strips with widths corresponding to the uncertainties.

The choice of the threshold magnitudes, denoted by M_r, is governed by three factors: (1) M_r should be large enough for useful hazard estimations, i.e., should correspond to potentially damaging earthquakes; (2) M_r should be large enough so that its occurrence should be largely influenced by the overall regional stress state; (3) M_r should be small enough to allow a sufficient statistical sample, but it should not be so small as to appear in all or most of the considered intervals, since then its forecast would be of no interest.

The choice of the time interval is governed by five factors: (1) Δt should be small enough for hazard estimations to be useful; (2) for too small Δt, state 0 (no earthquakes in any region) will be the most frequent one, so that the 0 to 0 transition will be dominant, and other probabilities different from p_{00} may be so small as to have no forecasting value; (3) for too large Δt, state $S-1$ (earthquakes in all regions) will be dominant, and all probabilities different from p_{S-1S-1} may be so small as to have no forecasting value; (4) for a given catalog length, increasing Δt diminishes the number of sampled transitions and makes estimates of p_{ij} less robust; (5) Δt should be large enough to allow interaction among regions. Thus, once the regions and threshold

magnitudes have been chosen, a preliminary estimation of Δt is based on the zeroes of $\theta_{00} - \theta_{S-1S-1}$ (the number of transitions from state 0 to state 0 minus the number of transitions from state $S-1$ to state $S-1$) and of $\xi_0 - \xi_{S-1}$ (the total number of occurrences from state 0 minus the total number of occurrences to state $S-1$).

Once Δt has been chosen, the initial time t_0 is fixed within the interval $t_{\min} - \Delta t < t < t_{\min} + \Delta t$, where $t_{\min}$ is the time of the first earthquake with $M \geq M_r$ and $t_{\min} - \Delta t$ is covered by the catalog. The use of different initial times allows the stability of the method to be assessed by measuring the variation in results from one realization to another (Nava *et al*., 2005).

The final choice of regions, time interval, and threshold magnitudes is made empirically and is pragmatically determined by the combination of the parameters that results in the best performance.

6. Performance Evaluation

For any method of hazard estimation, one direct measure of its performance is the assessment of the probability it assigned to the actual outcome or outcomes. In our case, prior to transition n, the system is in state i, and the probabilities p_{ij} constitute the hazard estimates for the next state; if state k occurs, then the observed transition occurred with likelihood $\hat{p}_n = p_{ik}$. The mean likelihood is

$$\hat{p} \equiv \frac{\sum_{n=1}^{n_t} \hat{p}_n}{n_t}, \tag{4}$$

where n_t is the number of realized transitions. According to this measure, the best of several hazard estimate methods would be the one yielding the highest $\hat{p}$. To assess whether a given hazard estimate by itself is good, we can measure its $\hat{p}$ against the "natural" reference level, which is the maximum entropy probability corresponding to the null hypothesis of uniform probability, where all states are equally likely to occur:

$$p_{ij}^{U} \equiv u = S^{-1} = \langle p_{ij} \rangle. \tag{5}$$

However, the $\hat{p}$ measure does not take into account the fact that the main object of hazard estimation is to *forecast* earthquake occurrences so that society may be prepared for them. A useful evaluation must take into account factors such as the probability level of forecasts, multiple (and contradictory) forecasts, false alarms, missed forecasts, among many others. We will now define quantitative measures for results that characterize the performance of a given method or model.

We define a *forecast* as the statement that a given outcome has a *high* probability of occurring. A probability is considered to be high if it lies above a given threshold

$$p_x = f_x\, u, \tag{6}$$

where f_x is a *success factor* which expresses the threshold in terms of the uniform probability u. A forecast is *successful* when an outcome with $p_{ij} > p_x$ occurs, and the number of successful forecasts is n_x. When a forecast is not successful, then it is a *false alarm* (type I error), and the number of false alarms will be denoted by n_f.

The *multiplicity*, m_i, is the number of elements larger than p_x in row i of the transition probability matrix, i.e., the number of simultaneous forecasts. When $m_i = 0$, there is no forecast (no success or false alarm) and outcome i is a *missed event* (type II error). The number of missed events is $n_S = n_t - n_p$, where n_p is the number of transitions for which there was a forecast, and n_t is the total number of considered transitions. In order to be very strict with successes, we divided each forecast success by the corresponding multiplicity; for example, for $m_i = 2$, a successful forecast would count as half a success plus one false alarm, while an unsuccessful one would count as two false alarms.

The *regional error* is the number of regions whose activity (occurrence or non-occurrence of earthquakes) was erroneously forecast.

All of the above counts are normalized by n_t so that performances with different lengths may be compared.

There is one last factor to consider, and this is the usefulness of the forecasts—i.e., the (non-)triviality of the forecasts. There are two trivial cases. The first is when the threshold magnitude is larger than the largest observed one; in this case, the probability of no earthquakes at all, $p_{00} = 1$, and all other probabilities are null. The second trivial case is when the threshold magnitude is very small, and $p_{S-1S-1} = 1$; however, this case is not so important to us since we are interested in large earthquakes only. For both trivial cases, the measures described above would yield optimum values, but the forecasts would be completely useless because they would carry no information at all. Thus, to avoid trivial, or close to trivial, cases, we need to penalize a hyperabundance of transitions ending in the 0 and $S-1$ states. This will be done by considering the diminishing information content of the corresponding transition probabilities.

The information (in bits) contained in a forecast with probability p is commonly defined as $I(p) = -\log_2(p)$ (cf. Goldman, 1953), where 1 bit corresponds to $p = 0.5$, which in turn corresponds to uniform probability in the binary case. Here, we will use *ubit* information units that assign a value of 1 to the information in the uniform probability, i.e., $I(p = u) = 1$ ubit. If a total of N_t transitions have been used to evaluate the transition probabilities, then the probability of any transition ending in state 0 is

$$p_0 = \frac{\sum_i p_{i0}}{N_t}, \tag{7}$$

and, for $p_0 > u$, all $\hat{p}_{i0}$ probabilities and forecast successes are multiplied by $I(p_0)$ (in ubits) before being counted. Observed $\hat{p}_{iS-1}$ probabilities and successes are qualified in a similar way.

All measures except $\hat{p}$ depend on the choice of success factor; a large f_x is desirable because we want forecasts to be made for high probabilities. A large will minimize false alarms, but a too large can so reduce the number of forecasts (increasing the number of missed events) and, hence, of successes, as to make the model almost useless. A low f_x will result in high multiplicity and yield a large number of false alarms and increase the regional error (both undesirable). So, the optimum value of f_x has to be found in order to obtain the best performance out of a given model

(a given combination of system, Δt, M_r^{m}, and M_r^{M}; for the direct method $M_r^{\mathrm{m}} = M_r^{\mathrm{M}} = M_r$).

Choosing the method or model which yields the best performance is not straightforward because there are usually tradeoffs between desirable and undesirable traits that make direct inspection and comparison impractical. Therefore, we decided to make use of grading functions, i.e., mathematical functions which take into account all relevant factors, weighted according to their relative size and importance, and combine them in such a way that desirable features increase their value and undesirable ones decrease it. Adjusting for relative size is necessary because we may be comparing quantities with different orders of magnitude (e.g., average probability versus normalized number of successes or false alarms). Weighting for importance is largely subjective, but it reflects a consensus of desirability for different traits; for instance, false alarms are quite undesirable (for many obvious reasons), and it is usually preferable to have fewer false alarms than more successes.

There is no rule to say which form a grading function should take. We tried linear, non-linear, product, and mixed grading functions, but in this work we show the results of only two:

$$d_0 = 0.8 + 5\hat{p} + \left(\frac{10 n_x - n_{\mathrm{f}} - e - n_{\mathrm{s}}}{n_{\mathrm{t}}} \right) + 0.00001 f_x, \quad (8)$$

$$d_1 = 1.0 + \left(\frac{(400 + 0.00002 f_x)\hat{p}^2 n_x^2}{n_t (n_{\mathrm{f}} + e + n_{\mathrm{s}})} \right). \quad (9)$$

Of course, the absolute values from a grading function are quite arbitrary and can be changed by modifying some of the baseline or scaling parameters (chosen here so both grades can be clearly seen when plotted using the same scales). We are mostly interested in the relative values, or relative optimum values (using always the same parameters in the grading function, of course). However, some idea of what the actual values do represent can be had by comparison with grades obtained for the reference "null hypothesis" Poisson (memory-less) and uniform (random guess) models.

7. Forecasts and *Aftcasts*

In actual forecasting, the probability of the system state for a time interval beginning at some given time is estimated from all available data up to this time. A serious problem is that a minimum number of transitions must be used to achieve robust probability estimates, and this may not leave enough transitions to have a representative sample for assessing the model's performance if the catalog is not long enough. This was the case the application discussed below. Consequently, a stopgap measure was adopted until enough forecast transitions were observed, to *aftcast* all 384 available transitions, i.e., do "forecasts" for data already used in estimating the probabilities and compare these results with the true forecasts of 20 transitions (the forecast for transition number 365 was made from probabilities estimated using the first 364 transitions; number 366 using transitions up to 365, and so on). As will be shown, aftcast and forecast performance evaluations roughly agree, but comparisons will be based mainly on aftcast results.

8. Application and Results

The models described in the preceding sections were applied to the Japan area, using data from the Japan Meteorological Agency (JMA) as reported by the International Seismological Centre (ISC) for the period spanning January 1964 to May 2002.

The way in which the system, time interval, and threshold magnitudes were chosen has already been discussed in Section 5. Four regions, shown in Fig. 2, were chosen: Kurile Islands (0), Central Japan (1), SE Japan (2), and Ryukyu Islands (3), defining 16 system states (Table 1). This is the same system for which Nava *et al.* (2005) applied the direct counting method to a combination of $\Delta t = 0.10$ yr (384 transitions), and Herrera *et al.* (2006) applied the mixed counting method. Here, we use the same parameters with the aim of being able to directly compare the results obtained by the different methods.

The perceptrons were trained according to Eq. (3) using diverse values for the initial weights w_{ij} and the learning rate η. We found that, for our data, bias inputs rapidly diminished to negligible values so that their presence did not significantly influence the results; hence, in what follows, we use no bias. The activation function we used is the identity (*purelin*) function $f(x) = x$, which simply transmits the output of each neuron.

For our application, the performance of these perceptrons is limited by the inconsistencies in the transition history; therefore, we cannot expect exact prediction. Consequently, we normalize the output to allow a probability interpretation; this normalization does not change the performance. In order to conserve the probabilistic interpretation of the W matrix, negative w_{ij} values were changed to zero after each correction step, and the matrix was then normalized by rows.

Perceptron performance depends heavily on the learning rate; *underlearning* with a very small $\eta \sim n_{\mathrm{t}}^{-1}$ (where n_{t} is the total number of transitions) leads to weights almost equal to the Markovian transition probabilities. This similarity indicates that the same probabilistic estimation can be obtained for two methods with different philosophies. Given that the estimation of the perceptron with a pure line activation function is equivalent to a least-squares estimation (Bishop, 1995), it is particularly interesting that the perceptron gives values similar to those of a model that simply counts the occurrences. Naturally, the forecasting performance of this perceptron is identical to that of the direct counting method. The immediate question is whether the weights from the slow-learning perceptron are the ones that yield the best performance. The answer is a categorical no, because the results shown below indicate that other models of perceptron, with different combinations of initial values and learning rate, give quite better results. Perceptron performance improves as η increases, reaches a peak, and then starts degrading for too large values of η (overlearning).

A search was made for the optimal parameter values, i.e., the ones that resulted in the highest grades for each model. These optimal grades were compared to see how the mixed perceptron method compared with the direct one and with the direct counting methods and to choose the best model for issuing useful forecasts. The following figures show

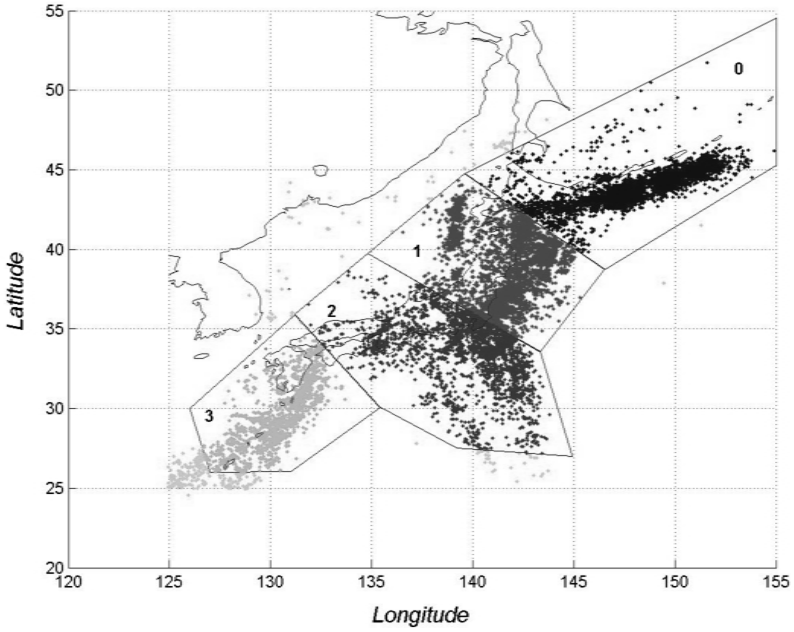

Fig. 2. Seismicity in the study zone reported in the ISC/JMA (January 1964–May 2002), and the four regions of the system.

Table 1. System states: decimal (left) and binary (right) showing which regional states (0 or 1) it comprises.

Status	Region				Status	Region			
	3	2	1	0		3	2	1	0
0	0	0	0	0	8	1	0	0	0
1	0	0	0	1	9	1	0	0	1
2	0	0	1	0	10	1	0	1	0
3	0	0	1	1	11	1	0	1	1
4	0	1	0	0	12	1	1	0	0
5	0	1	0	1	13	1	1	0	1
6	0	1	1	0	14	1	1	1	0
7	0	1	1	1	15	1	1	1	1

results obtained from our analysis.

Figure 3 shows the performance of the direct perceptron as a function of M_r, for the particular case of identity function $f(x) = x$, initial weight values $w_{ij} = 0.00001\,\forall\, i, j$ and learning rate $\eta = 0.10$.

For aftcasts (Fig. 3, left) the grading functions indicate that the direct perceptron performs better than the direct counting method for various M_r values (the dashed lines indicate the optimum values for the direct counting method); the d_0 grades show considerable improvements for $M_r \geq 6.1$, with a peak at $M_r = 6.1$; the f_x. The grading function values for these threshold magnitudes are shown in Table 2.

For forecasts (Fig. 3, right), the direct perceptron does give better results than the direct counting method—in all aspects and for all magnitudes of M_r considered in this analysis. The values of both grading functions are higher than those obtained with the direct counting method (dashed lines).

Given that the best results of the aftcast performance of the direct perceptron are for $M_r = 6.1$, we decided to explore the performance of the mixed perceptron for $M_r^{\mathrm{M}} = 6.1$. We found that, as for the counting methods, the mixed perceptron does not perform better than the direct perceptron. These results lead to the speculation that when the information at a given threshold magnitude is sufficient for a good performance of the direct methods, the smaller magnitudes do not contribute a significant quantity of information. However, for greater threshold magnitudes, when the direct method is no longer quite efficient, the mixed perceptron does significantly improve the results.

Figure 4 shows the grades of the mixed perceptron for $M_r^{\mathrm{M}} = 6.2$ as a function of M_r^{m}; the horizontal dashed lines indicate the (optimum) values for the direct perceptron for $M_r = 6.2$. It can be clearly seen that for aftcasts (left), the mixed perceptron gives better results than the direct perceptron for several values of M_r^{m}; both grading functions show improvements for $M_r^{\mathrm{m}} = 5.7, 5.8, 6.0$, with the maximum at 6.1. For forecasts, the mixed perceptron does give better results than the direct perceptron in all aspects and for all the considered values of M_r^{m}.

Figure 5 shows the (color scale-coded for the online version and gray scale-coded for the print version) transition probability matrices $\boldsymbol{W}$ from the direct perceptron for $M_r = 6.2$ (left), and from the mixed perceptron for $M_r^{\mathrm{m}} = 6.1$ and

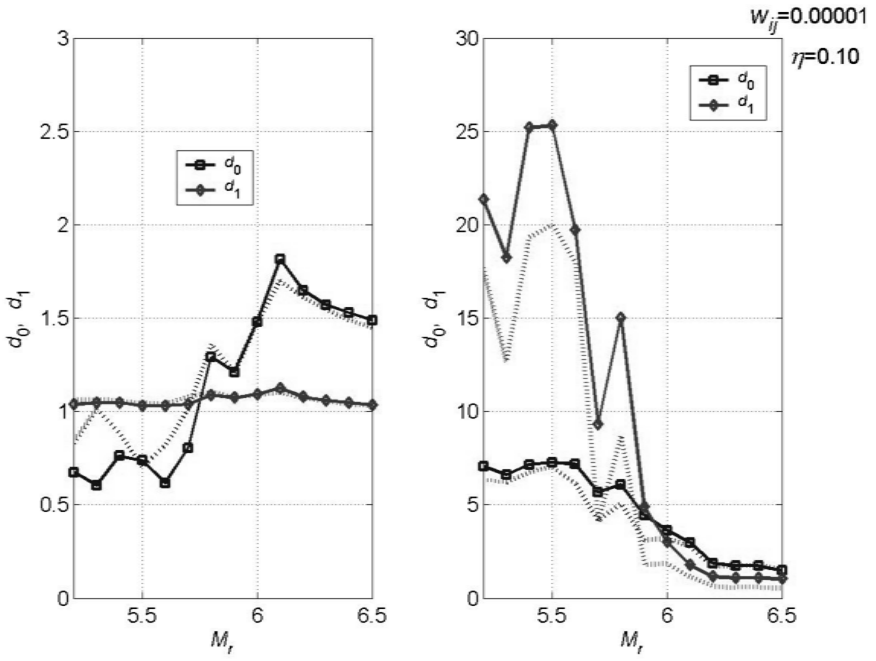

Fig. 3. Optimum grades for the direct perceptron as a function of M_r for aftcasts (left) and forecasts (right). Dashed lines are the best values for the direct counting method.

Table 2. Best results of state aftcasting for both perceptrons (direct and mixed) for high magnitudes. n_t is the number of transitions, f_x is the success factor, n_x is the number of successes, p^b and is the Bernoulli binomial probability of observing n_x successes in n_t transitions with uniform probability u.

Perceptron	M_r	n_t	f_x	n_x	d_0	d_1	p^b
Direct	$M_r = 6.1$	384	5.0	54	1.814	1.124	$1.7 \cdot 10^{-8}$
Direct	$M_r = 6.2$	384	6.2	43	1.647	1.077	$8.9 \cdot 10^{-5}$
Direct	$M_r = 6.3$	384	6.0	39	1.569	1.059	$9.6 \cdot 10^{-4}$
Direct	$M_r = 6.4$	384	6.5	37	1.527	1.047	$2.7 \cdot 10^{-3}$
Direct	$M_r = 6.5$	384	7.5	32	1.487	1.035	$2.0 \cdot 10^{-3}$
Mixed	$M_r^M = 6.1$ & $M_r^m = 5.7$	384	4.5	51	1.670	1.098	$2.3 \cdot 10^{-7}$
Mixed	$M_r^M = 6.2$ & $M_r^m = 6.1$	384	5.0	49	1.816	1.112	$1.2 \cdot 10^{-6}$
Mixed	$M_r^M = 6.3$ & $M_r^m = 5.7$	384	6.5	44	1.733	1.084	$4.6 \cdot 10^{-5}$
Mixed	$M_r^M = 6.4$ & $M_r^m = 5.8$	384	6.2	39	1.654	1.063	$9.6 \cdot 10^{-4}$
Mixed	$M_r^M = 6.5$ & $M_r^m = 6.1$	384	6.5	33	1.554	1.043	$1.4 \cdot 10^{-2}$

Table 3. Best results of state forecasting for both perceptrons for high magnitudes. Quantities are the same as in Table 2.

Perceptron	M_r	n_t	f_x	n_x	d_0	d_1	p^b
Direct	$M_r = 6.1$	20	6.0	4	2.980	1.769	$2.6 \cdot 10^{-2}$
Direct	$M_r = 6.2$	20	6.0	3	1.864	1.130	$9.3 \cdot 10^{-2}$
Direct	$M_r = 6.3$	20	9.0	2	1.737	1.069	$2.3 \cdot 10^{-1}$
Direct	$M_r = 6.4$	20	9.0	2	1.737	1.069	$2.3 \cdot 10^{-1}$
Direct	$M_r = 6.5$	20	4.0	2	1.492	1.035	$2.3 \cdot 10^{-1}$
Mixed	$M_r^M = 6.1$ & $M_r^m = 5.5$	20	6.0	6	4.469	5.194	$9.4 \cdot 10^{-4}$
Mixed	$M_r^M = 6.2$ & $M_r^m = 5.5$	20	6.0	7	4.407	5.715	$9.4 \cdot 10^{-4}$
Mixed	$M_r^M = 6.3$ & $M_r^m = 5.8$	20	9	4	3.520	3.092	$2.6 \cdot 10^{-2}$
Mixed	$M_r^M = 6.4$ & $M_r^m = 5.8$	20	9	4	3.520	3.092	$2.6 \cdot 10^{-2}$
Mixed	$M_r^M = 6.5$ & $M_r^m = 5.8$	20	9	4	3.598	4.185	$2.6 \cdot 10^{-2}$

$M_r^M = 6.2$ (right). It can be seen that, although the overall shape is the same, maxima differ between them; the differences between them are due to the information contributed by the earthquakes with $M_r^m \leq M < M_r^M$.

Figure 6 shows the behavior of the grading function for $M_r^M = 6.5$ as a function of M_r^m. The mixed perceptron performs consistently better than the direct perceptron for both aftcasts and forecasts, with a very conspicuous maximum for forecasts at $M_r^m = 6.2$.

The ratio $r = (d_0^{(\text{mixed})} - d_0^{(\text{direct})})/d_0^{(\text{direct})}$ for aftcasts

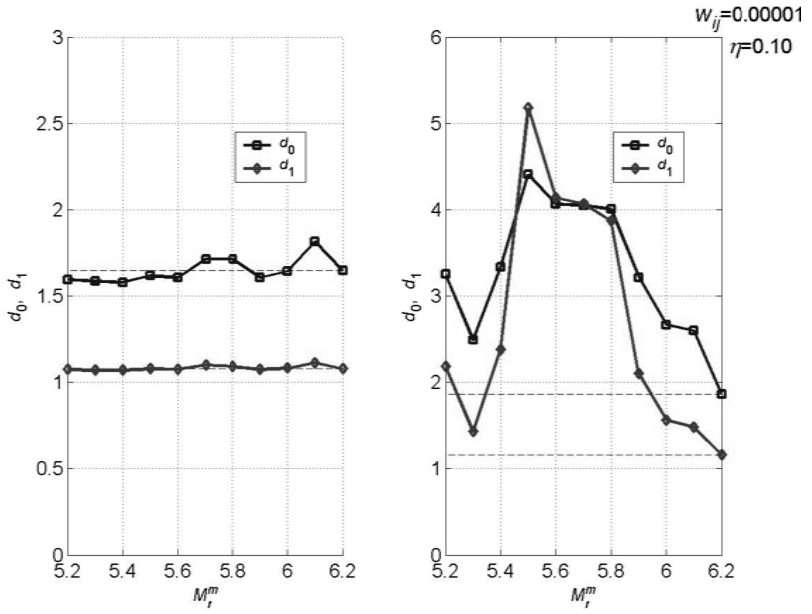

Fig. 4. Optimum grades for the mixed perceptron with $M_r^{\mathrm{M}} = 6.2$ as a function of M_r^{m}, for aftcasts (left) and forecasts (right). Dashed lines are the best values for the direct perceptron.

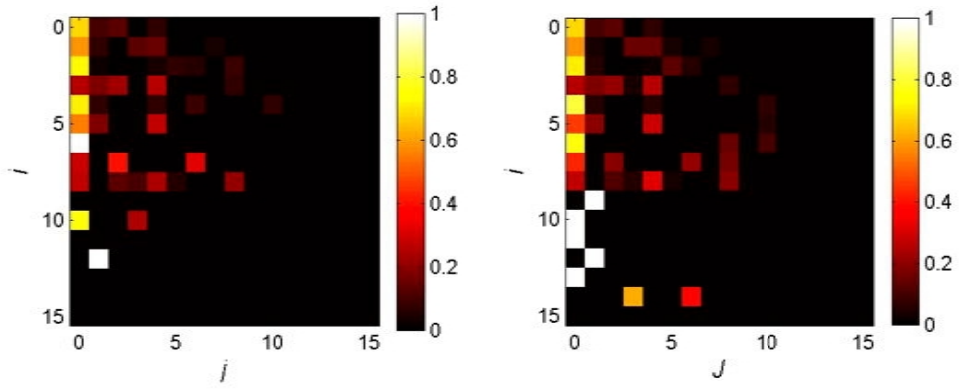

Fig. 5. Transition probability matrices. Left: Direct perceptron for $M_r = 6.2$. Right: Mixed perceptron for $M_r^{\mathrm{m}} = 5.7$ and $M_r^{\mathrm{M}} = 6.2$.

of the whole catalog (384 transitions) and forecasts of 20 transitions shows that the mixed method gives hazard estimations that are 10% better than the estimates obtained using the direct method (for aftcast and threshold magnitudes $M_r = 6.2$ and $M_r^{\mathrm{M}} = 6.2$, respectively). For forecasts and threshold magnitudes of $M_r = 6.1$ and $M_r^{\mathrm{M}} = 6.1$, respectively, r indicates an improvement of 50% in hazard estimations of the mixed perceptron over the direct one.

9. Discussion

The new methods proposed here yield better results than the straightforward counting methods (Nava *et al.*, 2005; Herrera *et al.*, 2006). These new methods yield useful and reliable seismic hazard estimates from the statistical analysis of seismicity catalogs, and they are adaptable to different seismogenic areas.

In order to compare our results with those from Herrera *et al.* (2006), we used their grading functions. The results in Tables 2 and 3 show that d_1 behaves like d_0 in all cases, but it is less sensitive; therefore, we will use only d_0 to discuss the performance of our methods. In order to quantify the relative performances, we will use the ratio $r_{\mathrm{p,c}} = (d_0^{\mathrm{perceptron}} - d_0^{\mathrm{count}})/d_0^{\mathrm{count}}$ for the best cases.

(1) Direct estimates:

a) Aftcasts: for $M_r = 6.1$, $d_0^{\mathrm{count}} = 1.699$ (Herrera *et al.*, 2006, table 2), and $d_0^{\mathrm{perceptron}} = 1.814$, which yields $r_{\mathrm{p,c}} = 0.068$; i.e., the perceptron aftcasts are 6.8% better than those from direct counting.

b) Forecasts: for $M_r = 6.1$, $d_0^{\mathrm{count}} = 2.794$ (Herrera *et al.*, 2006, table 3), and $d_0^{\mathrm{perceptron}} = 2.980$,

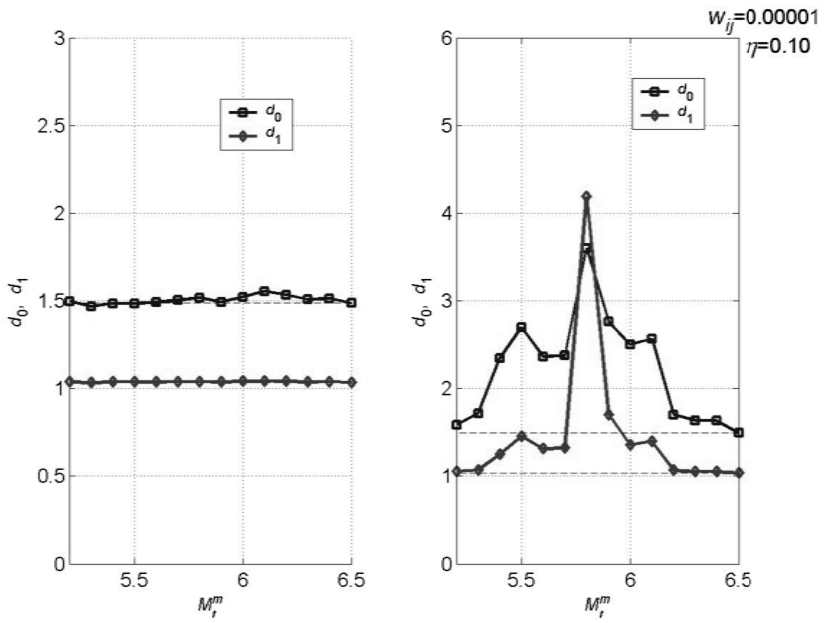

Fig. 6. Optimum grades for the mixed perceptron as a function of M_r^{m} for afcasts (left) and forecasts (right). Dashed lines are the best values for direct perceptron for $M_r^{\mathrm{m}} = 6.5$.

which yields $r_{\mathrm{p,c}} = 0.067$.

(2) Mixed estimates:

a) Aftcasts: for $M_r^{\mathrm{m}} = 5.7$ and $M_r^{\mathrm{M}} = 6.3$, $d_0^{\mathrm{count}} = 1.680$ (Herrera *et al*., 2006, table 2), and $d_0^{\mathrm{perceptron}} = 1.733$, which yields $r_{\mathrm{p,c}} = 0.032$.

b) Forecasts: for $M_r^{\mathrm{m}} = 5.8$ and $M_r^{\mathrm{M}} = 6.3$ or 6.4, $d_0^{\mathrm{count}} = 3.327$ (Herrera *et al*., 2006, table 3), and $d_0^{\mathrm{perceptron}} = 3.520$, which yields $r_{\mathrm{p,c}} = 0.058$.

It is important to underline that, although for the higher magnitudes the values of the grading functions observed by the mixed perceptron are not as large as in the optimum cases ($M_r^{\mathrm{M}} = 6.2$ for *aftcast* and $M_r^{\mathrm{M}} = 6.1$ for *forecast*), they are better than those observed by the direct methods. For example, the mixed perceptron for the forecast for $M_r^{\mathrm{m}} = 6.1$ and $M_r^{\mathrm{M}} = 6.5$ is 4.5% higher than the direct perceptron for $M_r = 6.5$. A major result is that the mixed method permits the forecasting magnitude limit of a given catalog to be extended, thereby enabling the seismic hazard for slightly larger, more important events to be estimated.

10. Conclusions

Our application of all proposed methods to the Japan area for *aftcasts* of the whole catalog and *forecasts* of 20 transitions yielded extremely satisfactory results that have negligible probabilities of being obtained by pure random guessing or by a memory-less model. Of all the methods proposed, the mixed perceptron gives the best results, particularly in terms of estimating the hazard for earthquakes with $M_r^{\mathrm{M}} = 6.5$. This means that for the Markovian process under consideration perceptron learning is more efficient in extracting the statistical information in catalogs than direct counting methods.

Acknowledgments. We are thankful to the JMA and ISC networks for the data used in this paper. Our sincere thanks are extended to Ewa Glowacka, Cecilio Rebollar, Luis Munguía, and Cinna Lomnitz for useful criticism and comments. We are grateful to two anonymous referees for constructive criticism and comments. We acknowledge the financial support of the CONACYT credit scholarship no. 142091.

References

Bishop, C. M., *Neural Networks for Pattern Recognition*, 488 pp., Calendon Press Oxford, UK, 1995.

Brillinger, D., Seismic risk assessment: some statistical aspects, *Earthq. Predict. Res.*, **1**, 183–195, 1982.

Chao, L. C. and M. J. Skibniewski, Estimating construction productivity: Neural network-based approach, *J. Comput. Civil Eng.*, **8**(2), 234–251, 1994.

Dowla, F. U., S. R. Taylor, and R. W. Anderson, Seismic discrimination with artificial neural networks: preliminary results with regional spectral data, *Bull. Seismol. Soc. Am.*, **80**, 1346–1373, 1990.

Fedotov, S., Regularities of the distribution of strong earthquakes in Kamchatka, the Kurile Islands, and northeast Japan, *Trudy Isnt. Fiz. Zemli. Acad. Nauk. SSSR*, **36**, 66–94, 1965.

García-Fernández, M. and J. J. Egozcue, Seismic hazard assessment in TERESA test areas based on a Bayesian technique, *Nat. Haz.*, **2**(3–4), 249–265, 1989.

Goldman, S., *Information Theory*, 385 pp., Dover Publs. Inc., USA, 1953.

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

Herrera, C., F. A. Nava, and C. Lomnitz, Time-dependent earthquake hazard evaluation in seismogenic systems using mixed Markov Chains: An application to the Japan area, *Earth Planets Space*, **58**, 973–979, 2006.

Hirata, N., Past, current and future of Japanese national program for earthquake prediction research, *Earth Planets Space*, **56**, xliii–l, 2004.

Ito, T., S. Yoshioka, and S. Miyazaki, Interplate coupling in southwest Japan deduced from inversion analysis of GPS data, *Phys. Earth Planet. Inter.*, **115**, 17–34, 1999.

Ito, T., S. Yoshioka, and S. Miyazaki, Interplate coupling in northeast Japan deduced from inversion analysis of GPS data, *Earth Planet. Sci. Lett.*, **176**, 117–130, 2000.

Kagan, Y. and D. Jackson, Seismic gap hypothesis: Ten years after, *J. Geophys. Res.*, **96**, 21419–21431, 1991.

Kanamori, H., The energy in grat earthquakes, *J. Geophys. Res.*, **82**, 2981–2987, 1977.

Karunanithi, N., W. J. Grenney, D. Whitley, and K. Bovee, Neural networks for river flow prediction, *J. Comput. Civil Eng.*, **8**, 201–220, 1994.

Kashyap, R. L., Speaker recognition from an unknown utterance and speaker-speech interaction, *IEEE Trans on Acoustics, Speech and Signal Processing*, **24**, 481–488, 1976.

Katsumata, K. and M. Kasahara, Precursory seismic quiescence before the 1994 Kurile Earthquake ($M_w = 8.3$) revealed by three independent seismic catalogs, *Pure Appl. Geophys.*, **155**, 443–470, 1999.

Lomnitz, C. and F. Nava, The predictive power of seismic gaps, *Bull. Seismol. Soc. Am.*, **73**, 1815–1824, 1983.

McCann, W. R., S. P. Nishenko, S. P. Sykes, and J. Krause, Seismic gap and plate tectonics: seismic potential for major boundaries, *Pure Appl. Geophys.*, **117**, 1082–1147, 1979.

Mcllraith, A. L. and H. C. Card, Birdsong recognition using backpropagation and multivariate statistics, *IEEE Trans on Signal Processing*, **45**, 2740–2748, 1997.

Mogi, K., Migration of seismic activity, *Bull. Earthq. Res. Inst.*, **46**, 53–74, 1968.

Mogi, K., *Earthquake Prediction*, 355 pp., Academic Press, Japan Inc., 1985.

Nava, F. A., C. Herrera, J. Frez, and E. Glowacka, Seismic hazard evaluation using Markov chains; Application to the Japan area, *Pure Appl. Geophys.*, **162**, 1347–1366, 2005.

Patwardahan, A. S., R. B. Kulkarni, and D. Tocher, A semi Markov model for characterizing recurrence of great earthquakes, *Bull. Seismol. Soc. Am.*, **70**, 323–347, 1980.

Reid, H. F., The mechanism of the Earthquake, The California Earthquake of April 18, 1906, in *Report of the State Earthquake Investigation Commission*, edited by Carnegie Institution, **2**, 16–18, Washington D.C., 1910.

Richter, C., *Elementary Seismology*, 521 pp., W. H. Freeman, San Francisco, 1958.

Rosenblatt, F., The perceptron: A probabilistic model for information storage and organization in the brain, *Psychological Rev.*, **65**, 386–408, 1958 (Reprinted in Anderson & Rosenfeld, 92–114, 1988).

Rosenblatt, F., *Principles of Neurodynamics*, 215 pp., Spartan, New York, 1962.

Rüttener, E., J. J. Egozcue, D. Mayer-Rosa, and S. Mueller, Bayesian estimation of seismic hazard for two sites in Switzerland, *Nat. Haz.*, **14**, 165–178, 1996.

Shimazaki, K., Intra-plate seismicity and inter-plate earthquakes: historical activity in southwest Japan, *Tectonophysics*, **33**, 33–42, 1976.

Shimazaki, K. and T. Nakata, Time predictable recurrence model for large earthquakes, *Geophys. Res. Lett.*, **7**, 279–282, 1980.

Thatcher, W. and J. B. Rundle, A viscoleastic coupling model for the cyclic deformation due to periodically repeated earthquakes at subduction zones, *J. Geophys. Res.*, **89**, 7631–7640, 1984.

Utsu, T., Estimation of parameters for recurrence models of earthquakes, *Bull. Earthq. Res. Inst., Univ. Tokyo*, **59**, 53–56, 1984.

Wyss, M., K. Shimazaki, and T. Urabe, Quantitative mapping of a precursory seismic quiescence to the Izu-Oshima 1990 (M6.5) earthquake, Japan, *Geophys. J. Int.*, **127**, 735–743, 2007.

Zhao, Y. and K. Takano, An artificial neural network approach for broadband seismic phase picking, *Bull. Seismol. Soc. Am.*, **89**, 670–680, 1999.

C. Herrera (e-mail: cherrera@uabc.mx) and F. A. Nava