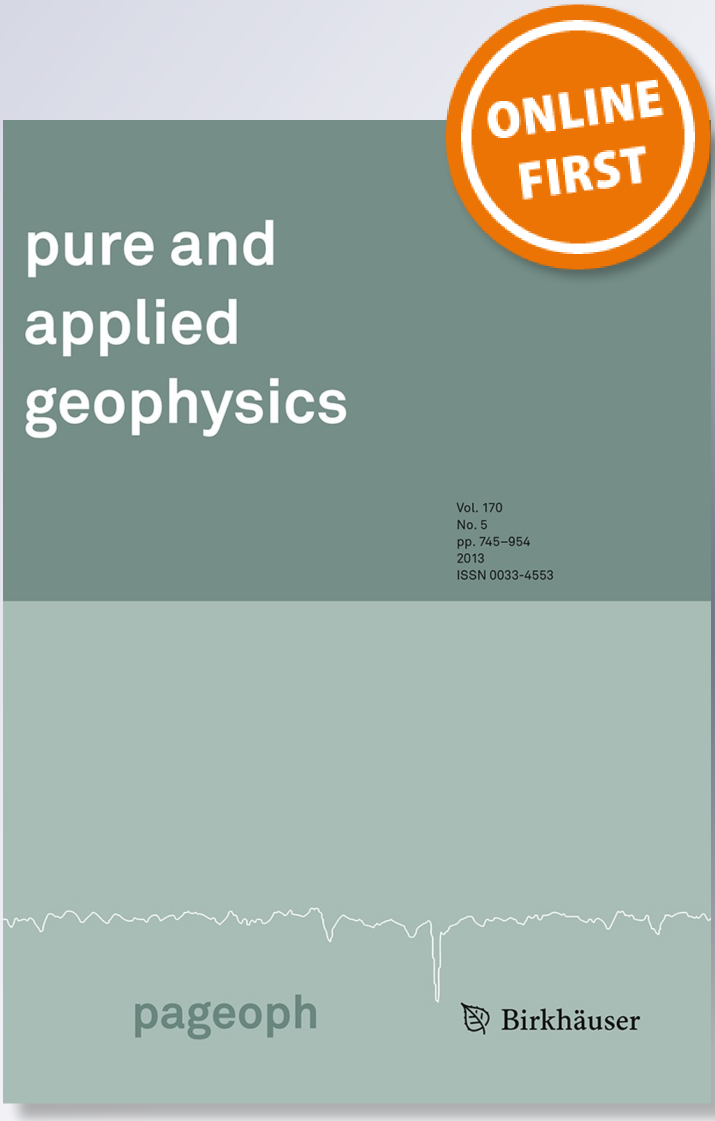
Semi-Periodic Sequences and Extraneous Events in Earthquake Forecasting: I. Theory and Method, Parkfield Application

Fidencio Alejandro Nava Pichardo, Claudia Beatriz Quinteros Cartaya, Ewa Glowacka & José Duglas Frez Cárdenas

Pure and Applied Geophysics
pageoph

ISSN 0033-4553

Pure Appl. Geophys.
DOI 10.1007/s00024-013-0679-5



Springer

Pure Appl. Geophys.
© 2013 Springer Basel
DOI 10.1007/s00024-013-0679-5

Pure and Applied Geophysics

Semi-Periodic Sequences and Extraneous Events in Earthquake Forecasting: I. Theory and Method, Parkfield Application

Fidencio Alejandro Nava Pichardo,[1] Claudia Beatriz Quinteros Cartaya,[1] Ewa Glowacka,[1] and José Duglas Frez Cárdenas[1]

Abstract—We present a new method to identify semi-periodic sequences in the occurrence times of large earthquakes, which allows for the presence of multiple semi-periodic sequences and/or events not belonging to any identifiable sequence in the time series. The method, based on the analytic Fourier transform, yields estimates of the departure from periodicity of an observed sequence, and of the probability that the sequence is not due to chance. These estimates are used to make and to evaluate forecasts of future events belonging to each sequence. Numerous tests with synthetic catalogs show that the method is surprisingly capable of correctly identifying sequences, unidentifiable by eye, in complicated time series. Correct identification of a given sequence depends on the number of events it contains, on the sequence's departure from periodicity, and, in some cases, on the choice of starting and ending times of the analyzed time window; as well as on the total number of events in the time series. Some particular data combinations may result in spectra where significant periods are obscured by large amplitudes artifacts of the transform, but artifacts can be usually recognized because they lack harmonics; thus, in most of these cases, true semi-periodic sequences may not be identified, but no false identifications will be made. A first example of an application of the method to real seismicity data is the analysis of the Parkfield event series. The analysis correctly aftcasts the September 2004 earthquake. Further applications to real data from Japan and Venezuela are shown in a companion paper.

Key words: Seismic hazard, earthquake sequences, semi-periodicity.

1. *Introduction*

Reid's elastic rebound model of earthquake generation (Richter 1958) postulates that strain accumulates in the ground until the associated stress surpasses the rock strength and the ground breaks suddenly in an earthquake that releases the accumulated strain, after which the strain begins again to accumulate and the earthquake recurs after the stress again surpasses the rock strength. The tectonic plate paradigm (e.g., Morgan 1968; Cox 1973; Richardson *et al*. 1979) furnishes an acceptable strain source, inter-plate motion, for the elastic rebound model, and since this motion can be considered to have a constant rate over thousands of years, the earthquake cycle could be expected to be periodic.

This expectation led to forecasts based on studies of recurrence times (e.g., Lomnitz 1966; Rikitake 1976; and references therein) and on the popular time predictable model (Shimazaki and Nakata 1980). However, earthquakes have not generally shown the expected periodicity, as illustrated by the Parkfield, California, experiment in which, on the basis of a six earthquake sequence with an apparent $\sim$21.9 years periodicity, an $M \geq 6$ earthquake was expected to occur in 1993 with 0.95 probability, but the expected earthquake was 11 years late (Bakun and Lindh 1985; Bakun *et al.* 2005; Savage 1993; Lomnitz 1994; Kagan 1997; Jackson and Kagan 2006). Later on we will show a forecast for the Parkfield earthquake using our method.

The lack of periodicity does not mean that the elastic rebound model or the constant stress rate source assumption are wrong, but rather that seismogenic zones are complex so that the model should only be applied locally. Indeed, rocks are heterogeneous and strength varies spatially, cracks can range from microscopic to regional faults, and the strength and stress of a seismogenic region is determined, at each point, by characteristics which may vary in time and by stresses that depend in a non-linear fashion on

[1] Seismology Department, CICESE, Carretera Tijuana-Ensenada 3918, Ensenada, BC 22860, Mexico. E-mail: fnava@cicese.mx; cquinter@cicese.mx; glowacka@cicese.mx; jofrez@cicese.mx

Published online: 25 May 2013

Birkhäuser

the history of rupturing, giving rise to self-organized criticality (Bak *et al*. 1988; Bak and Tang 1989; Bak and Chen 1991; Turcotte 1992). This self-organized criticality together with the elastic rebound model, indicate that semi-periodicity (also known as quasi-periodicity) can be expected to occur, at least locally.

We do not expect repetitions of so-called characteristic earthquakes (Jackson and Kagan 2006), but systems with self-organized criticality, while having essentially random activity for small events, do exhibit semi-periodic behavior for large events, and there may be characteristic rupture lengths in a given region (Lomnitz 1994).

Thus, we can expect large to mega earthquakes to be the best candidates to exhibit semi-periodic behavior; and it is precisely these earthquakes that are the ones that cause more damage and the ones that it is important to forecast.

Usual periodicity studies simple-mindedly assume that all earthquakes above a given magnitude threshold are due to a single process, and do not consider the possibility of the seismicity being caused by more than one semi-periodic processes and/or having *extraneous* earthquakes, i.e., events unrelated to the observed semi-periodic processes, whose occurrence times may be considered random. It is evident that recurrence (inter-event) times from a series incorporating events from different co-existing sequences plus extraneous events will be absurdly short. Besides, the usual studies do not use the observed departures from periodicity to estimate the probabilities of their periodicity results being due to chance.

We propose a semi-periodicity analysis method that takes into account all the above mentioned factors and may be useful as a factor in seismic hazard estimates.

In the present paper we present the theory and the computational scheme, illustrated by the analysis of a synthetic catalog. Tests with synthetic catalogs are very important because, with these catalogs, it is possible to tell if sequences and extraneous events are correctly identified; also, numerous analyses of synthetic catalogs shed light upon the limitations and capabilities of the method. We also present the application of our method to the Parkfield data and the corresponding forecast.

In the companion paper by Quinteros *et al.* (2013), we present our scheme for preparing real catalogs for analysis, which diminishes the problem of ignoring earthquakes below some threshold magnitude, and applications of the method to data from Japan and Venezuela.

2. *Method*

2.1. Theory

Earthquake occurrence can be considered a point process (Daley and Vere-Jones 2002) in which the times of observed earthquakes constitute a series

$$t_{\mathrm{E}} = \{t_j;\ \ j = 1, \ldots, K\}.$$

If the series corresponds to a periodic process, with *recurrence period* τ, then the occurrence times, which constitute a *sequence*, are expressible as

$$t_j = t_0 + j\,\tau,$$

where t_0 is some initial time, and in this case there is no problem at all in determining $\tau_j = t_j - t_{j-1} = \tau$.

A process and its related sequence are *semi-periodic* if:

$$t_j = t_0 + j\,\tau + \theta_j, \tag{1}$$

where θ_j is a realization of a random variable θ such that $|\theta| << \tau$ (we will assume it smaller than $\sim 0.2\,\tau$,) in this case it is still possible to estimate τ through the mean and standard deviation of inter-event times. $\tau_j = t_j - t_{j-1} = \tau + \theta_j - \theta_{j-1}$ which, depending on the signs of the random variations, may differ from τ by as much as ± the sum of the θ absolute values.

In practice the semi-periodic t_{E} sequence may be "contaminated" by R extraneous events, whose times are arbitrary, so that the resulting observed series is

$$t_{\mathrm{E}} = \{t_k;\ \ k = 1 : N\},$$

where $N = K + R$ and K of the t_k's will correspond to t_j's and R of them will not, but we do not know which are which.

Another possibility is that the observed series could be a combination of two (or more) sequences from semi-periodic processes having different recurrence and initial times, plus extraneous events, in which case $N = K_1 + K_2 + \cdots + R$.

Fourier analysis is an obvious tool to look for periodicities or, in the present case, semi-periodicities.

Using the discrete Fourier transform (DFT) to analyze a discrete time series constructed from the observed seismicity presents the serious problem that the DFT only recognizes periods which are submultiples of the total observation time, and the appropriate *T*, multiple of the unknown period(s) we are looking for is not known! If the right *T* is not chosen, the periodic component will be "smeared" all over the spectrum; the problem is particularly difficult because sequences corresponding to large earthquakes are usually short, so that the frequency intervals $\Delta s = T^{-1}$ are large. Thus, for the DFT, various total times must be tried to look for the best definition, and in the case where two or more different recurrence periods are involved, it may be impossible to find a given *T* that will allow correct identification of all of them.

The answer is to obtain the analytical Fourier transform (FT) by building a function

$$f(t) = \sum_{j=j_1}^{j_2} \delta(t - t_j), \qquad (2)$$

and recognizing it as a section of the function corresponding to the infinite series:

$$f_\infty(t) = \sum_{j=-\infty}^{\infty} \delta(t - t_j),$$

so that

$$f(t) = f_\infty(t)\, \Pi\left[\frac{t - t_c}{T}\right], \qquad (3)$$

where $\Pi(t)$ is the boxcar function centered at time $t_c = (t_b + t_e)/2$ and $T = t_e - t_b$, where t_b and t_e are the times where our catalog begins and ends, respectively.

The analytical FT of (3) is

$$F(s) = \int_{-\infty}^{\infty} f(t) e^{-i2\pi t s}\, dt = \sum_{j=j_1}^{j_2} e^{-i2\pi t_j s}$$
$$= F_\infty(s) * T\,\mathrm{sinc}(T s)\, e^{-i2\pi t_c s}, \qquad (4)$$

and it is possible to evaluate *F*(*s*) for any frequency *s* whatsoever. As will be shown below, the important components of the spectrum $F_\infty(s)$ can usually be identified properly in spite of the convolution with the sinc function.

What should we expect to find in the frequency domain? For a strictly periodic series,

$$f(t) = \Sigma(t/\tau) \quad \supset \quad F(s) = \tau\, \Sigma(\tau s)$$

where the shah function $\Sigma(x) = \sum_{j=-\infty}^{\infty} \delta(x - j); j \in Z$ (e.g., Bracewell 1965). For semi-periodicity the spectrum will differ from the shah depending on the actual values of the θ_i, and extraneous events make *F*(*s*) depart further from the periodic case. Relatively small random variations from periodicity cause large phase shifts for high frequencies, but small shifts for low ones, so the periods of interest, commensurable with *T*, are recognizable in most cases. Extraneous events cause phase shifts at all frequencies, but since they are not periodic, the Fourier transform is usually able to identify, although sometimes only approximately, the underlying semi-periodicities. These approximate identifications are then refined as described below.

A sample periodic sequence with unit amplitudes and $\tau = 100$ years, spanning $T = 400$ years (top left) is shown in Fig. 1; also shown is a segment of its analytic Fourier amplitude spectrum (middle left) which is the shah convolved with the sinc($T\,s$) function shown at bottom left. Figure 1 also shows three examples of amplitude spectra for the same sequence after random variations of θ_j, to show the effect of variations on the spectrum; normal deviations with $\sigma_\theta = 5$ years (top right) yield a perfectly identifiable spectral replicated peak; for deviations with $\sigma_\theta = 10$ years (middle right) the peak is still identifiable; and for deviations with $\sigma_\tau = 15$ years (bottom right) the peak is still identifiable, but is barely replicated and a larger, spurious, peak can be seen on its right.

2.2. *Spectral Analysis*

In this section we will describe the proposed method of semi-periodic sequence identification, using as illustration a synthetic series (Fig. 2) consisting of two semi-periodic sequences, one having five events with $\tau^{(1)} = 75$ years and normal random

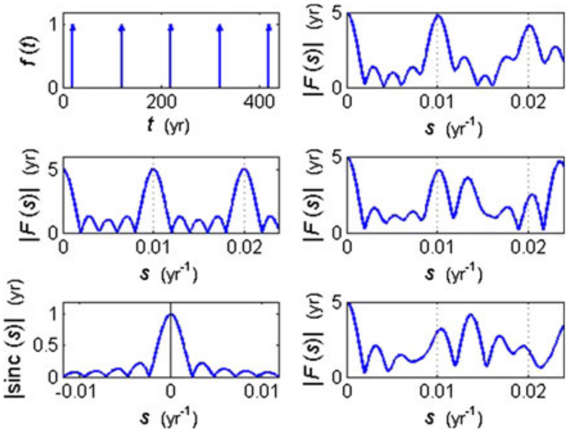


Figure 1
Example showing a periodic sequence with $\tau = 100$ years (*top left*), its analytic FT (*middle left*), and the convolved sinc function (*bottom left*). Spectra of the same series after random normal variations θ, with standard deviations $\sigma = 5$ years (*top right*), $\sigma = 10$ years (*middle right*), and $\sigma = 15$ years (*bottom right*). *Dotted lines* indicate the frequency $1/\tau$, and its first harmonic

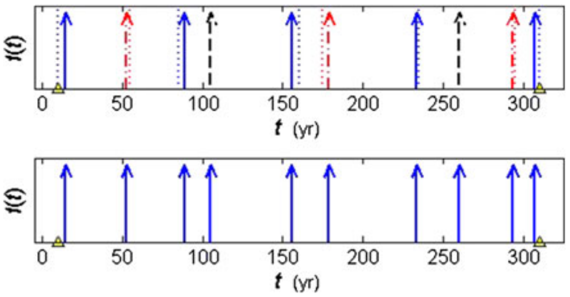


Figure 2
Synthetic series for example of an application. *Top* the *five solid* (*blue*) and the *three dot-dash* (*red*) *arrows* indicate the members of the first and second semi-periodic sequences, respectively, while the *dotted lines* close to each arrow denote the occurrence times if the sequences were completely periodic. The *dashed* (*black*) *arrows* represent the two extraneous events. Triangles show t_{b} and t_{e}. *Bottom* example series as seen without identifying characteristics; note how difficult it is to estimate any kind of periodicities in this series

variations with $\sigma^{(1)} = 4.0$ years, and the other having three events with $\tau^{(2)} = 120$ years and variations with $\sigma^{(2)} = 3.5$ years, and two extraneous events with random occurrence times; we will suppose the catalog spans from $t_{\mathrm{b}} = 10$ to $t_{\mathrm{e}} = 310$ years.

That sequences within a given time series can be very difficult to identify by eye is illustrated in Fig. 2. Indeed, Poisson-distributed seismicity has exponentially distributed intervals $\Delta t_i = t_{i+1} - t_i$, so that, for Poisson seismicity, the ratio $\rho = E[\Delta t]/\text{Standard deviation}[\Delta t] \approx 1$ (for very large N). For $N = 10$ events the most probable ρ value (estimated by Monte Carlo methods) is $\rho_{\max} \sim 1.093$ for our example series $\rho_{\mathrm{E}} = 1.170$ and $\Pr(\rho_{\max} \le \rho \le \rho_{\mathrm{E}}) = 0.102$ which illustrates that series composed of semi-periodic sequences plus extraneous events can appear quite Poissonian, which is why the sequences may be indiscernible to the eye.

The analysis is made through several passes of spectral peak, and corresponding sequence, identification. After each sequence identification (except for the last one), events which do not belong to the sequence are kept or rejected as possible options for the next pass, according to a criterion which grows more stringent with each pass. The last pass is made using only the events belonging to the sequence identified in the previous pass in order to have refined estimates of period and phase. For the example we are using four passes with acceptance criteria being 1/4, 1/4.5, 1/5, and 1/6 of the measured periods, respectively. We tried different criteria, and the one chosen here gave good results and was easy to remember; but other criteria may be used.

The function $f(t)$ (Eq. 2) is built from the total N event occurrence times ($N = 10$ in this example) and the total observation time is calculated from the times declared as the beginning and ending times of the catalog: $t_{\mathrm{b}} = 10$ and $t_{\mathrm{e}} = 310$ years. Mean inter-event times and the corresponding standard deviation are evaluated and printed. The FT is computed, using the sum of exponentials in Eq. (4), and its absolute value is plotted (Fig. 3); rough guides for acceptable semi-periodicity frequencies $s_{\min}$ and $s_{\max}$ are indicated on the spectrum. Since there must be an absolute minimum of three events corresponding to a given period in order to identify it, the minimum frequency to be considered is $s_{\min} \approx 2/T$. On the other hand, the smallest period that can be identified from the observed series cannot be much smaller (considering possible departures from periodicity, not known yet) than the largest observed inter-event time $\Delta t_{\max}$ so the largest observable frequency is approximated by $s_{\max} \approx 1.25/\Delta t_{\max}$. As mentioned above, both $s_{\min}$ and $s_{\max,}$ are only rough guides to the acceptable frequency range, and the results do not depend on their exact values. The part of the

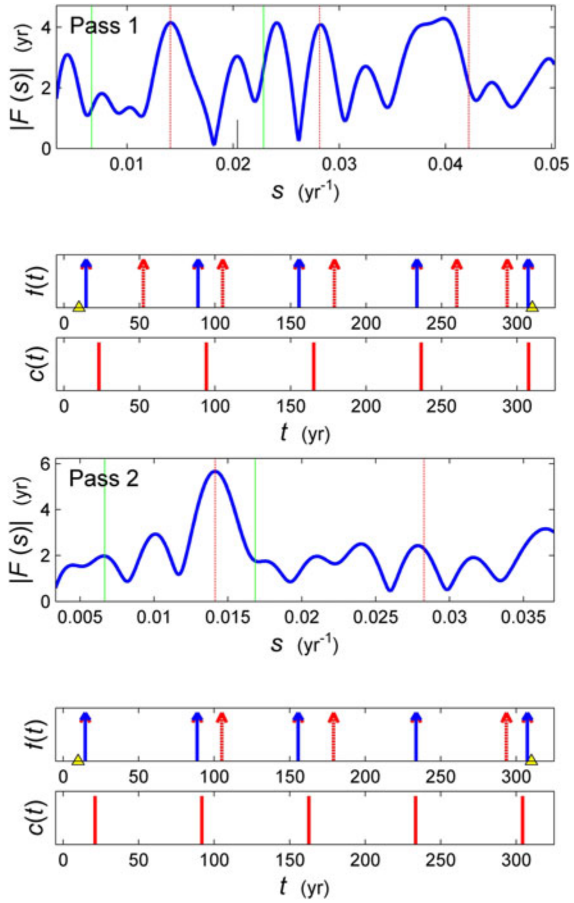


Figure 3

First two passes of sequence identification. For each pass (*top*) Fourier spectrum of the series shown in the middle picture, the *solid vertical* (*green*) *lines* are the guideline frequencies $s_{\min}$ and $s_{\max}$, the *vertical dashed lines* indicate the chosen frequency and its multiples; (*middle*) Earthquake series with events shown as *dashed* (*red*) *arrows*, with those identified as corresponding to the sequence with period $\tau_p = 1/s_p$ overdrawn by *solid* (*blue*) *arrows*; the corresponding comb teeth are shown by *solid lines* in the bottom picture. The *short vertical line* near 0.02 years^{-1} in the first pass indicates that the corresponding peak was rejected as a viable option

spectrum shown will range from $s_1 = 0.5\,s_{\min}$ to $s_2 = 2.2\,s_{\max}$ in order to be able to identify long periods, frequencies somewhat larger than $1/\Delta t_{\max}$ and possible replications of spectral peaks.

To determine possible sequence periods, a spectral peak located between $s_{\min}$ and $s_{\max}$ (or not too far from one of them) is chosen, and the corresponding frequency s_{p} (indicated by the first dashed vertical line on the left) is noted (Fig. 3, top). From this frequency, a "comb", $c(t)$, having K "teeth" evenly separated by $\tau_{\mathrm{p}} = 1/s_{\mathrm{p}}$ and starting time $t_{0\mathrm{p}} = -\phi_{\mathrm{p}}/2\pi\, s_{\mathrm{p}}$, where ϕ_{p} is the spectral phase corresponding to s_{p}, is built (Fig. 3, bottom). For each tooth the closest observed earthquake time is identified, in order to determine which events could be part of a sequence having these particular periods and starting times; if for some tooth, no earthquake time is found that differs by less than the pass criterion, then that particular τ_{p} is rejected as being a construct of the FT, and some other spectral maximum is tried. Medium to high frequency peaks should be tried first, because when there are many events in a series the comb corresponding to a spurious spectral peak may be artificially well fit.

The goodness of fit is estimated from the differences between the time of each teeth t_i^{c} and the occurrence time of the corresponding earthquake t_i, as

$$\varepsilon = \sqrt{\frac{\sum_{i=1}^{K}\left(t_i^{\mathrm{c}} - t_i\right)^2}{K-2}} = \sigma, \tag{5}$$

which takes into account that the data have been used to estimate two parameters. This standard error is also designated as σ because, if we suppose the deviations from periodicity to be normally distributed (and there is no basis to suppose otherwise) then this error is the best estimate of the distribution's standard deviation.

The upper half of Fig. 3 shows the first sequence identification pass from the time series shown in the bottom of Fig. 2. The chosen frequency is $s_{\mathrm{p}} = 0.0141$ year^{-1}, corresponding to period $\tau_{\mathrm{p}} = 1/s_{\mathrm{p}} = 71.09$ years; a comparison with Fig. 2 shows that one five-element sequence has been correctly identified. The fit standard deviation is $\sigma = 8.35$ years.

If the comb is adequately fit, the identification is refined by repeating the process after eliminating those events that could not possibly belong to the comb, being more than $\tau_{\mathrm{p}}/4$ away from any tooth. The spectrum thus obtained is usually clearer than the first one (Fig. 3, lower half); again, a peak is chosen and events belonging to the new comb are identified, with $\tau_{\mathrm{p}}/4.5$ as the new acceptance criterion. The new identified period in the example is $\tau_{\mathrm{p}} = 70.75$ years with $\sigma = 6.10$ years.

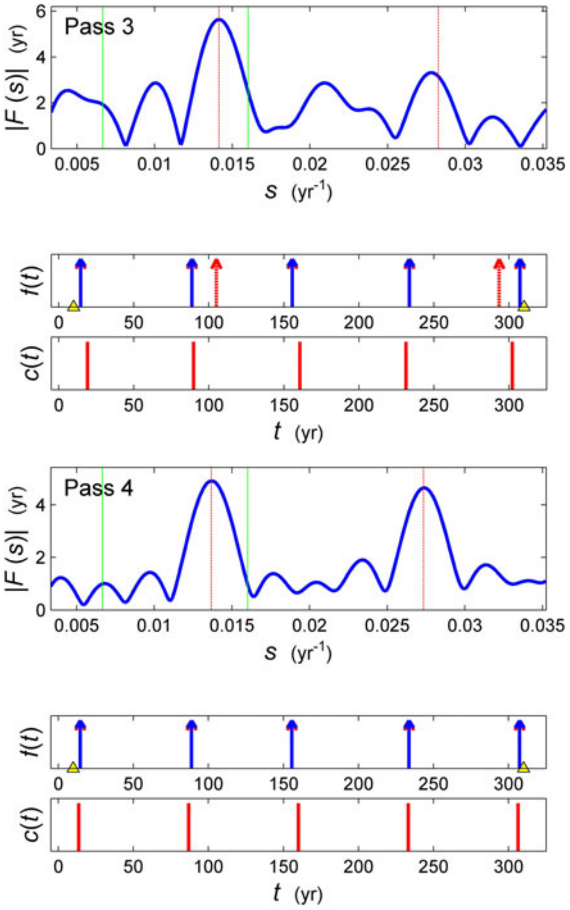


Figure 4
Third and fourth passes of sequence identification. Same conventions as in Fig. 3

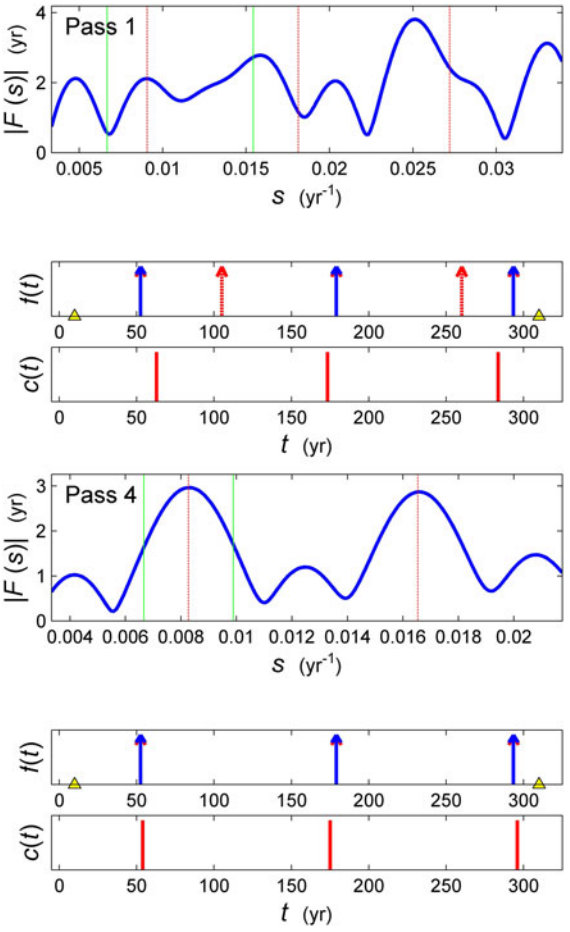


Figure 5
First and fourth passes in the identification of the second sequence in the synthetic catalog. Same conventions as in Fig. 3

After events separated by more than $\tau_p/4.5$ are eliminated; a third pass is made (Fig. 4, upper half) with $\tau_p/5$ acceptance criterion; for the example this pass yields a new period $\tau_p = 70.76$ years with $\sigma = 6.10$ years. Finally, a fourth pass is made using only the events which were identified as corresponding to the comb (Fig. 4, lower half). Usually, at this stage the spectral peaks are clearly identifiable; the comb resulting from these final, refined, estimates of τ_p and τ_{0p} is accepted or rejected using $\tau_p/6$ as the final acceptance criterion. The final identified period is $\tau_p = 73.17$ years with $\sigma = 2.91$ years.

Once a sequence and the corresponding comb have been determined, the significance estimation and forecast are carried out as described in the next sections.

Next, the events belonging to the identified sequence are eliminated, and the analysis is applied to the remaining events, as shown in Fig. 5 (for reasons of space, we show only the first and fourth steps in the analysis. The first-pass identified period is $\tau_p = 110.21$ years with $\sigma = 15.50$ years, and the final result is $\tau_p = 120.97$ years with $\sigma = 4.95$ years).

The events belonging to the new identified sequence are eliminated, and the process continues until no more semi-periodicities are present in the remaining series, at which point the analysis ends.

Analysis of numerous realizations of synthetic time series, from strictly periodic where we know which events belong to which sequences and which are extraneous, to completely random, show that the FT is surprisingly capable of identifying underlying semi-periodicities within a series. Correct identification of sequences depends on the number of events in each sequence, on each sequence's deviations from strict periodicity, and, in some cases, on the choice of starting and ending times of the analyzed time window; identification also depends, of course, on the amount of "noise", i.e., events that do not belong to the particular sequence to be identified. Rarely, some particular values of the variations, combined with the relative times of different sequences and extraneous (random) events may result in spectra where significant periods are obscured by large amplitudes artifacts of the transform; however, artifacts can be usually recognized because they lack harmonics; thus, in most of these cases, true sequences may not be identified, but no false identifications will be made.

2.3. *Significance*

Events occurring at random can, of course, generate semi-periodic sequences spanning a given observation time, and these cannot be distinguished from sequences generated by a semi-periodic process. However, the probability that the identified sequence could result from purely random occurrence of the observed events, P_ϕ can be estimated, as described below. Thus, $P_c = 1 - P_\phi$ is the probability that the sequence did not result by chance, and this estimate is a measure of how significant a given identification is; clearly, a sequence having a large probability of occurring by chance is not a good basis for any sort of forecast.

Let us assume that we have identified, from among N events occurring over a time T, a sequence having K elements with σ rms deviation from the comb teeth; the actual values of the period and starting time, while extremely important for forecasting, can be considered as a byproduct of having found a sequence with the above mentioned characteristics. Thus, we will consider P_ϕ to be the probability of N Poisson-distributed events resulting in a similar sequence (same T and K) with rms deviation $\varepsilon \leq q\,\sigma$, where q (>1) is a safety factor, introduced because for forecast purposes it is better to be pessimistic (we use $q = 1.25$). We estimate P_ϕ using a Monte Carlo method, where we generate a large number of random series, discard those that cannot result in a similar sequence, for each eligible sequence adjust comb period and starting time by least-squares, compute the rms deviation between events and comb teeth, and consider one success each time this rms deviation is $\leq\varepsilon$; finally, P_ϕ is estimated by dividing the number of successes by the total number of series.

Figure 6 shows how P_ϕ varies with σ/T, and K, for typical values of K and N. For given T, all P_ϕ probabilities increase significantly as the uncertainty σ increases; but sequences with only three events can have significant values of P_ϕ even for small σ.

Now, since a semi-periodic sequence consists of a periodic part (the comb) plus random variations around the associated comb teeth, characterized by σ, it follows that P_c also is the probability that the next comb tooth, tooth number K since teeth are numbered from 0 to $K - 1$, will indicate, within the uncertainty associated with the semi-periodic behavior, the occurrence time of the next (future) earthquake in the sequence. $P_c = 1$ would mean absolute certainty as to the occurrence of the next earthquake around the

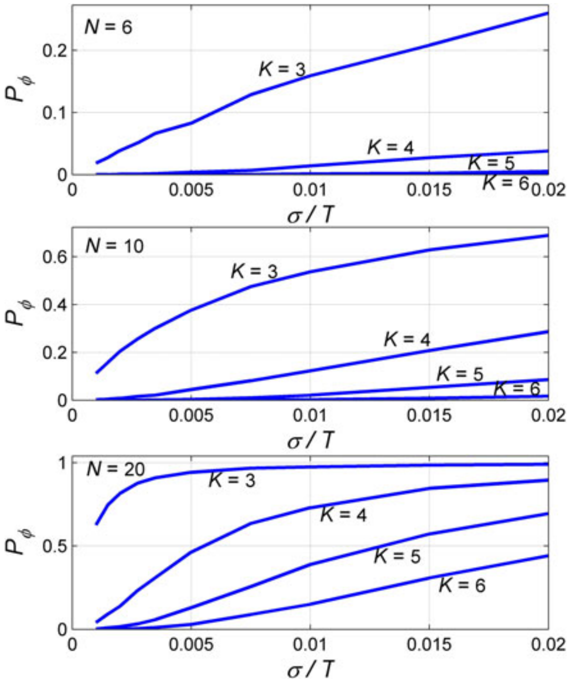


Figure 6
Illustration of the dependence of P_ϕ on σ/T and K for three typical values of N

time of the next comb tooth, while the usual $P_c < 1$ indicates that there is a possibility that the forecast of a future event may be a false alarm (because the observed sequence has occurred by chance). For the first sequence in our sample, using $q = 1.25$ and 50,000 realizations, for $T = 300$ years, $N = 10$, $K = 5$, and $\sigma = 2.91$ years, $P_\phi = 0.021$ and $P_c =$ 979.

2.4. Forecasts

The whole object of identifying semi-periodic sequences is to be able to forecast the future occurrence of comparable events. According to a given identified comb, with period τ_p, initial time τ_{0p}, K teeth (each identified with an observed event, and the first one occurring at the initial time), and θ characterized by its standard deviation σ, the next event should occur at time

$$t_f = t_{0p} + K\,\tau_p \pm q\,\sigma, \tag{6}$$

where q is a factor that can be set to give a desired confidence interval to the forecast. This forecast time for our example, for $q = 2$ and 95.45 % confidence level, is $t_f = 379.53 \pm 5.82$ years.

Of course, the actual distribution of variations from periodicity cannot be assessed from the number of occurrences to be expected in real-life applications of the method; however, considering that a large amount of processes and factors are involved in bringing about the rupture initiation, i.e., the occurrence time, of a large event, we can invoke the Central-limit theorem to approximate the unknown probability density distribution by a normal one, centered at t_f, with standard deviation σ, truncated at the time of the most recent event and normalized to have total area p_c. The forecast method is not dependent on the variation distribution being normal, so that anyone who considers some other distribution more appropriate can substitute it for the normal one in what follows.

Other measures of the significance of our results are given by the probability and information gains (Vere-Jones 1998; Harte and Vere-Jones 2005) over the background Poisson probability estimate of having at least one earthquake during a given interval $(t_f - q\sigma,\, t_f + q\sigma)$ around the forecast time. For such an interval, the probability of having a comb event is $P_{cq} = P_c\,\mathrm{erf}(q)$ the Poisson probability of having at least one event is $\pi_{1+} = 1 - e^{-\lambda 2 q \sigma}$ with $\lambda = N/T$, and the Poisson probability of having an event not belonging to the comb is $\pi^*_{1+} = 1 - e^{-\lambda^* 2 q \sigma}$ with $\lambda^* = (N - K)/T$; thus, the probability of having a comb event and/or a non-comb event is $P_{cq} + \pi^*_{1+} - P_{cq}\pi^*_{1+}$, so that the probability gain is

$$P_G = \frac{P_{cq} + \pi^*_{1+} - P_{cq}\pi^*_{1+}}{\pi_{1+}}.$$

The information gain is the difference between the self-information (Fano 1961) or entropy score (Harte and Vere-Jones 2005) of the comb forecast and that of the background probability,

$$I_G = \log_2(P_G),$$

expressed in bits.

For the first sequence of our example, with $N = 10$, $T = 300$ years, $K = 5$, $\tau_p = 73.171$ years, the comb probability is $p_c = 0.979$; the occurrence rates are $\lambda = 0.0333$/years, and $\lambda^* = 0.0167$/years, so for three intervals centered on the forecast time, forecast and Poisson probabilities, and the resulting probability and information gains are as shown in Table 1.

The forecast pdf, $p(t)$, is presented graphically in Fig. 7 (top) together with the identified sequence, the comb $c(t)$, and $\pm 2\sigma$ ($q = 2$) uncertainty ranges.

Figure 7 (middle) shows a close-up of $p(t)$, together with other functions that help to visualize some consequences of the forecast. The survivor function

$$S(t) = \Pr(x > t) = 1 - P(t) = \int_t^\infty p(x)\,\mathrm{d}x,$$

where x is occurrence time, $P(t)$ is the cumulative of $p(t)$, is the probability of not having had an earthquake at a given time t. For $P_c < 1$, $S(t)$ does not tend

Table 1

Probabilities and probability and information gains for the first sequence in the example

q	P_{cq}	π_{1+}	π^*_{1+}	P_G	I_G (bits)
1	0.668	0.176	0.092	3.967	1.988
2	0.935	0.3215	0.1765	2.944	1.558
3	0.976	0.441	0.252	2.228	1.156

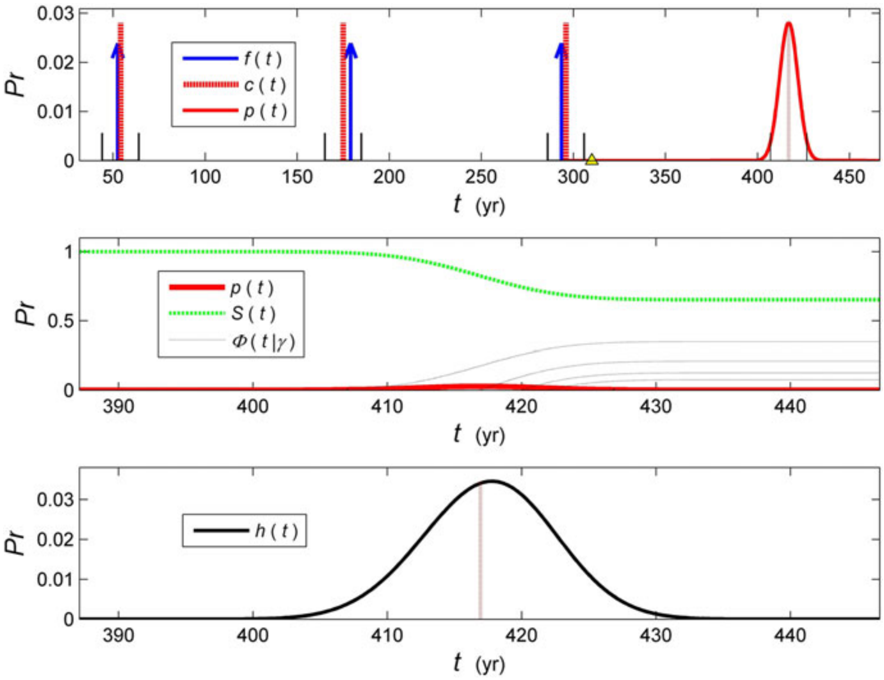


Figure 7
Top example of forecast based on the sequence identification shown in Fig. 5; earthquakes are (*blue*) *arrows* and comb teeth are *dashed thick* (*red*) *lines* (line heights have no meaning); the forecast pdf $p(t)$ is shown by the *thick* (*red*) curve and the *dotted line* at its center indicates t_f while the short vertical lines on both sides of the teeth and the forecast indicate $\pm 2\sigma$. *Middle* close-up of $p(t)$ *thick* (*red*) *line* with t_f indicated by a *short dotted vertical line*; also shown are $S(t)$ *dot-dash* (*green*) *line* and $\Phi(t|\gamma)$ for several γ values *dotted* (*black*) *lines*

to zero, but tends to P_ϕ, the probability of a false alarm.

Also shown is the future lifetime

$$\Phi(t|\gamma) = \Pr(x \le t \,|\, x > \gamma) = \frac{\Pr(\gamma < x \le t)}{\Pr(x > \gamma)} = \frac{P(t) - P(\gamma)}{S(\gamma)},$$

which gives, for times t greater than a given time γ, the probabilities that the earthquake will occur at some time x before or at t, given that the earthquake has not yet occurred by time γ. The dotted lines show $\Phi(t|\gamma)$ for various γ; for γ before the significant part of $p(t)$ the future lifetimes grow more rapidly as γ increases, attaining similar maximum values $\sim P_c$. For $P_c = 1$ the same limiting value is attained for all γ, but for $P_c < 1$ the maximum value decreases with increasing γ, and tends to zero for survival times later than the significant part of $p(t)$; i.e., for these times the earthquake occurrence probabilities tend to zero.

Figure 7 (bottom) shows the hazard function

$$h(t)\,\mathrm{d}t = \Pr(t < x \le t + dt \,|\, x > t),$$

where x is occurrence time, which describes how, given that the event has not occurred at time t, (the remaining part of) $p(t)$ should be renormalized:

$$h(t) = \frac{p(t)}{1 - P(t)} = \frac{p(t)}{S(t)}.$$

For $P_c < 1$ the survivor function $S(t)$ does not tend to zero, so the hazard function does not increase indefinitely, but instead attains a maximum at some time later than that of the maximum of $p(t)$, and then decreases to zero.

3. *Parkfield*

We will now apply our method to analyze the time series (Table 2) used by Bakun and Lindh

Table 2

The first six lines (1857 to 1966) contain the set of occurrence times used by Bakun *and* Lindh *(1985) to predict an earthquake near Parkfield around December 1987; the last line gives the time of actual earthquake occurrence*

Year/month/day	*M*	*t*
1857/01/09	7.9	1,857.02339
1881/02/02	~6.0	1,881.08904
1901/03/03	~6.0	1,901.16849
1922/03/10	6.0	1,922.18767
1934/06/08	6.0	1,934.43425
1966/06/28	6.0	1,966.48767
2004/09/28	6.0	2,004.74240

(1985) to predict the occurrence of an $M \geq 6.0$ earthquake near Parkfield, California, on December 1987 ±0.7 year.

Figure 8 shows the semi-periodicity analysis of the time series consisting of the first six times shown in Table 1, with $t_b = 1,850$ and $t_e = 1,970$ ($T = 120$ years). For the first pass (top), the small vertical line around 0.0365 years^{-1} indicates that this frequency was rejected by the program; the 0.0268 years^{-1} is acceptable and yields a comb with four events, the solid arrows above the tentative comb lines show the best-fitting events. The 1881 event is discarded and the second pass (middle) shows now a single dominant spectral peak around 0.0277 years^{-1}, and the 1922 event is again identified as a bad fit. The 1922 event is discarded and the third and fourth passes (third not shown, because it is identical with the fourth) identify a clear spectral peak at 0.0275 years^{-1}, corresponding to $\tau_p = 36.36$ years with $\sigma = 4.55$ years which results in the aftcast of an earthquake around $t_f = 2005.63 \pm 9.10$ years with a high $P_c = 0.894$ (Fig. 9). The corresponding probability and information gains are presented in Table 3.

The large aftcast probability is due to the relatively small $N = 6$ for $K = 4$, in spite of the large σ. In any case, as can be seen in Fig. 9, the aftcast predicts very well the occurrence time of the September 2004 earthquake (dotted arrow).

Savage (1993) states that the problem with the Parkfield prediction was that it did not consider alternative hypotheses, and presents other two hypotheses based on the same time series. We propose the new alternative hypothesis that not all earthquakes in a given series need belong to the same sequence.

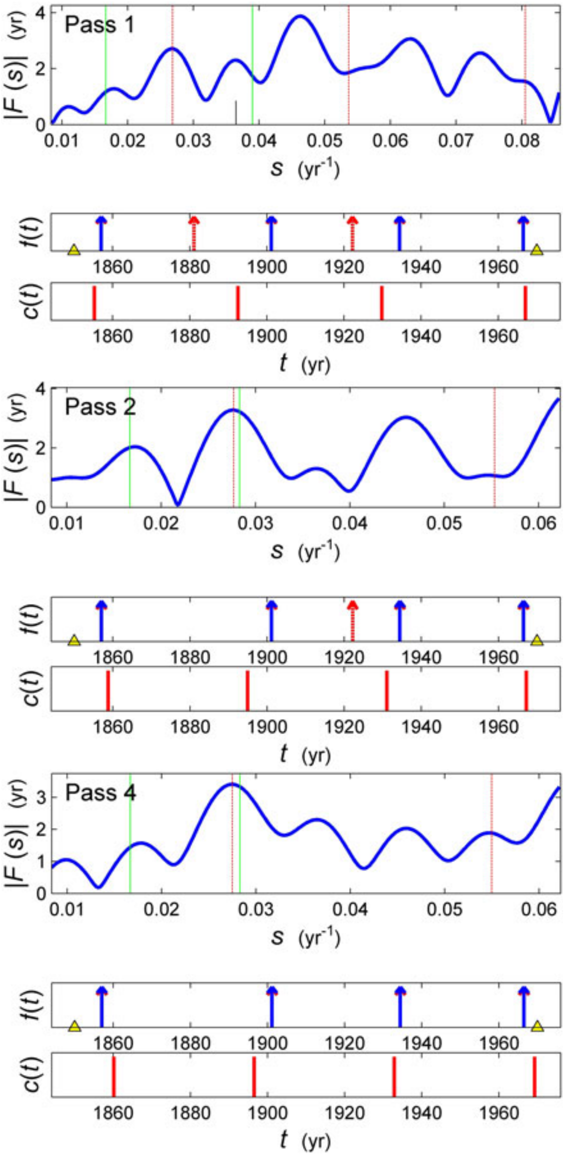


Figure 8
Semi-periodicity analysis of the Parkfield series. Same conventions as in Fig. 3

4. Discussion

We present a method to look for semi-periodic sequences within a series of occurrence times, which takes into account the possible presence of more than one sequence and of event times, which do not belong to the sequences. Many trials with synthetic data sets, show that the method is surprisingly capable of identifying semi-periodicities, although some combinations of events result in spectra without recognizable maxima

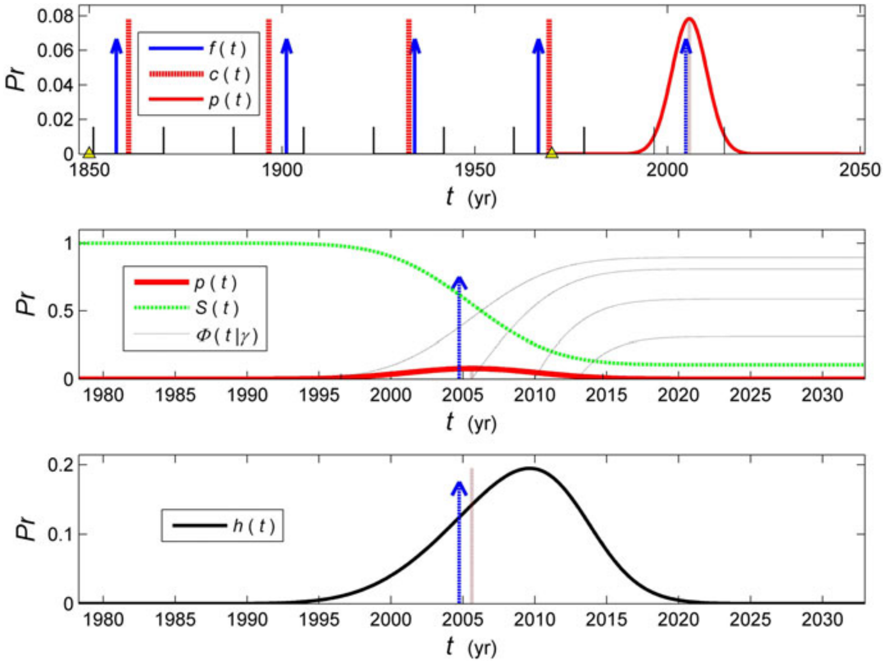


Figure 9

Forecast for the Parkfield sequence. Same conventions as in Fig. 7, plus actual occurrence of the 2004.74 earthquake shown as a *dotted arrow*. The forecast pdf in the *middle plot* can barely be seen because of the very large σ , but how close the actual earthquake occurrence time is to the forecast time t_f can be clearly appreciated in the top and *bottom plots*

Table 3

Probabilities and probability and information gains for the Parkfield sequence

q	P_{cq}	π_{1+}	π^*_{1+}	P_G	I_G (bits)
1	0.611	0.366	0.141	1.820	0.864
2	0.854	0.598	0.262	1.493	0.578
3	0.892	0.745	0.366	1.251	0.323

or with maxima that are artifacts of the transform, but these maxima are usually recognizable because they lack harmonics. We also propose measures to evaluate the significance of the identifications, i.e., the probability that they are not occurring by chance, and use these to forecast future events.

A very large success ratio in identifying known sequences within synthetic data sets, prompted the application of the method to real seismicity data sets; the first application considers the Parkfield prediction data set, and results in a quite accurate aftcast. Other results from application to data from Japan and Venezuela, are given in the companion paper by Quinteros *et al.* (2013).

The main limitation in the application of the method is that, given that it is a purely statistical method, there is no assurance that the identified sequences correspond in effect to physical semi-periodic processes that may be expected to produce, semi-periodically, future earthquakes. Thus, the forecasts from the method cannot be considered as a basis for concern or preventive action until their reliability is confirmed by further studies; however they could be useful as a factor when considering hazard estimates.

Acknowledgments

This research was partially funded by a CONACYT scholarship 242919 (C. Quinteros) and was partially carried out within project CGL2011-29474-C01-01. Our sincere thanks to Cinna Lomnitz and to an unknown reviewer for helplful criticism and comments; we also would like to express our thanks to editor Andrzej Kijko.

References

Bak, P., Tang, C., and Wiesenfeld, K. (1988) *Self-organized criticality*. Phys. Rev. A *38*(1), 364–374.

Bak, P., and Tang, C. (1989) *Earthquakes as a self-organized critical phenomenon*. J. Geophys. Res. *94*(B1), 15635–15637.

Bak, P., and Chen, K. (1991) *Self-organized criticality*. Sci. Am. 26–33.

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

Bakun, W., Aagard, B., Dost, B., Ellsworth, W., Hardebeck, J., Harris, R., Ji, C., Johnston, M., Langbein, J., Lienkaemper, J., Michael, A., Murray, J., Nadeau, R., Reasenberg, P., Reichle, M., Roeloffs, E., Shakal, A., Simpson, R., and Waldhauser, F. (2005) *Implications for prediction and hazard assessment from the 2004 Parkfield earthquake*. Nature *437*, 969–974. doi 10.1038/nature04067.

Bracewell, R., The Fourier transform and its applications. (McGraw-Hill Book Co., USA, 381 pp., 1965).

Cox, A. (1973) Plate tectonics and geomagnetic reversals. (W. H. Freeman and Co.).

Daley, D., and Vere-Jones, D., An introduction to the theory of point processes, (Springer, USA, 469 pp., 2002).

Fano, R., Transmission of information. (M.I.T Press & J. Wiley & Sons, USA, 389 pp. 1961).

Harte, D., and Vere-Jones, D. (2005) *The entropy score and its uses in earthquake forecasting*. Pure Appl. Geophys. *162*, 1229–1253.

Jackson, D., and Kagan, Y. (2006) *The 2004 Parkfield earthquake, the 1985 prediction, and characteristic earthquakes: Lessons for the future*. Bull. Seism. Soc. Am. *96*, S397–S409.

Kagan, Y. (1997) *Statistical aspects of Parkfield earthquake sequence and Parkfield prediction experiment*. Tectonophysics *270*, 207–219.

Lomnitz, C. (1966) *Statistical prediction of earthquakes*. Reviews of Geophysics *4*, 377–393.

Lomnitz, C., Fundamentals of earthquake prediction. (JohnWiley & Sons Inc., USA, 326 pp. 1994).

Morgan, W. J. (1968). *Rises, trenches, great faults, and crustal blocks*. J. Geophys. Res. *73*(6), 1959–1982.

Quinteros, C., Nava, F., Glowacka, E., and Frez, J. (2013) Semi-periodicity and extraneous events in earthquake forecasting. II: Application, forecasts for Japan and Venezuela. Pageoph accepted.

Richardson, R. M., Solomon, S. C., and Sleep, N. H. (1979) *Tectonic stress in the plates*. Reviews of Geophysics, *17*, 981–1019.

Richter, C., Elementary Seismology, (W. H. Freeman and Co., USA, 768 pp., 1958).

Rikitake, T., Earthquake prediction. (Elsevier, 357 pp., 1976).

Savage, J. (1993) *The Parkfield prediction fallacy*. Bull. Seismol. Soc. Am. *83*, 862–881.

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

Turcotte, D., Fractals and Chaos in Geology and Geophysics. (Cambridge University Press, Second Edition, New York. 221 pp, 1992).

Vere-Jones, D. (1998) *Probabilities and information gain for earthquake forecasting*. Comput. Seismol. *30*, 248–263.

(Received January 28, 2013, revised April 25, 2013, accepted April 29, 2013)