Société de Calcul Mathématique, S. A.
Algorithmes et Optimisation

Probabilistic Methods in Seismology

Critical Analysis of the Models

Document addressed to the

Commissariat à l'Energie Atomique

Direction de l'Energie Nucléaire

by the

Société de Calcul Mathématique S. A.

in application of contract no 4000288990/P5B61, May 26th, 2007

redaction : Bernard Beauzamy and Olga Zeydina

July 21, 2007

Siège social et bureaux : 111, Faubourg Saint Honoré, 75008 Paris. Tel. : 01 42 89 10 89. Fax : 01 42 89 10 69
Société Anonyme au capital de 56 200 Euros. RCS : Paris B 399 991 041. SIRET : 399 991 041 00035. APE : 731Z

Executive Summary

We analyze here the usual model, describing the number of earthquakes above a certain magnitude in a certain zone : it uses Poisson Laws. We give two descriptions of this model, which allow us to discuss in depth the underlying assumptions.

We show that, under the assumptions of this model, the total number of earthquakes in a global territory should be the sum of the expected numbers for each zone, and similarly that the total number over a certain period should be the sum of expected numbers per year : this answers a question in the "cahier des charges".

We show how to compute the confidence interval associated to the expected number of earthquakes, in a zone, for a given year, or for a whole period, or for a whole territory. Using the data communicated to us by the CEA, we show that these real data fall significantly outside the confidence interval : this indicates that there are strong doubts about the correctness of the model.

I. Introduction

Probabilistic methods in seismology are rather recent, and their conceptual frame is not well established, as the number of discussions about such matters clearly shows.

Two key assumptions are usually made, but sometimes they are not explicit. They are :

1. The process is stationary. This means that the probability to have an earthquake, in a given zone, does not depend on the past. It is the same every year (usually, the probabilities are computed for a year, but it would be the same for other intervals of time). This probability is not modified if an earthquake occurred a year before or, conversely, if that zone has not seen an earthquake for quite a long time.

This leads of course to simple formulas. But an earthquake is the result of tensions and compressions inside the Earth. When this energy has been freed (and this may be done through several consecutive earthquakes, upon several days), then it takes some time to accumulate again this energy. So, a model which does not reflect these links with energy is certainly quite superficial.

2. Two distinct zones are assumed to be independent, in the sense that if some earthquake occurs in one of them, it does not change the probability to have an earthquake in the other.

This may be true if the zones are large enough, but not so if the zones are small. Note that we take into account only the earthquakes which are "born" in each zone, not the possibility for a captor in a zone to detect an earthquake in another zone.

Under these two assumptions, the "cahier des charges" mentions a question about the total number of expected earthquakes, and the way to compute it, using the law prescribed for each zone. We answer this question here.

II. Computing the number of expected earthquakes in a zone

There are two possible presentations. In order to clarify the matters, and make all assumptions explicit, we indicate both.

A. Poisson process

The usual presentation assumes that the number of earthquakes, above a certain magnitude, follows a Poisson law, that is :

$$P\left\{N(t+\tau)-N(t)=k\right\}=\frac{e^{-\lambda\tau}\left(\lambda\tau\right)^{k}}{k!} \tag{1}$$

where $N(t)$ is the number of occurrences before time t, so $N(t+\tau)-N(t)$ is the number of occurrences between time t and time $t+\tau$. Usually, the time τ is taken to be one year, so formula (1), with $\tau=1$, indicates the number of occurrences of earth-

quakes, above a fixed magnitude, during a year. The coefficient λ is fixed from experimental data. We note that the right-hand side of (1) does not depend on t; this is our "stationary" assumption : the probability to have a certain number of earthquakes in a year is the same, no matter when this year starts.

In order to simplify our notation, let us denote by X the random variable indicating the total number of earthquakes, above a certain level, in a certain zone. Formally, $X = X(m,z)$ depends on the magnitude and on the zone. Then this random variable follows a Poisson law :

$$P\{X = k\} = \frac{e^{-\lambda}\lambda^k}{k!},\ k = 0,1,... \tag{2}$$

where λ is a coefficient, empirically fixed, which depends both on the magnitude and on the zone : $\lambda = \lambda(m,z)$.

Simple computations on the Poisson Law show that the expectation $E(X)$ is :

$$E(X) = \lambda \tag{3}$$

and the variance $V(X)$ is :

$$V(X) = \lambda \tag{4}$$

So, the parameter λ has a simple meaning : it corresponds to the expected value (average number) of the number of earthquakes each year. The specific properties of the Poisson Law give, by (4), that the dispersion is $\sigma(X) = \sqrt{\lambda}$.

Presented this way, the law looks rather artificial : why should X follow a Poisson Law, rather than any other law ? In order to understand the meaning of this choice, we turn to the second presentation, which is much easier to understand.

B. Good days, bad days

Let us build a very simple model : in any year, in any zone, we have good days and bad days. Bad days are those during which some earthquake occurs in that zone (above a certain magnitude), good days are those during which nothing happens. Let us denote by p the probability that a day should be a bad day. Formally, $p = p(m,z)$ depends on the chosen magnitude and on the zone.

In this description, we do not consider the possibility to have several earthquakes the same day (that's just a bad day), and the probability p does not depend on the past. All days are independent : you may have 5 bad days in a recent past, the probability to have a bad day tomorrow is the same p.

Then the probability to have k earthquakes (above a certain level, in a certain zone) is given by a binomial law :

$$P\{Y=k\}=\binom{N}{k}p^k(1-p)^{N-k} \tag{5}$$

with $N=365$.

Then the expected value is (by easy computations on the binomial law) :

$$E(Y)=Np \tag{6}$$

and the variance :

$$V(Y)=Np(1-p) \tag{7}$$

But the Binomial Law $B(N,p)$ may be approximated by the Poisson Law with parameter $\lambda=Np$ if N is large (here $N=365$) and p small, which is the case here. Also, we see that the variance in (7) satisfies :

$$V(Y)=Np(1-p)\approx Np=\lambda$$

which is the variance in (4).

So we see that our model, using Poisson Law, is completely equivalent, both in theory and in practice, with a model using bad days, good days, with parameter :

$$p=\frac{\lambda}{365} \tag{8}$$

The latter model is much easier to understand, conceptually speaking, but the former is easier to use for computations.

III. Taking several years into account

The above formulas are valid for any given year. Let us now see how we can take several years into account : let n be any number of years.

Let X_j be the random variable which indicates the number of earthquakes above a certain magnitude, for year $j=1,\ldots,n$, in a given zone. It follows a Poisson Law, with the same parameter for all years :

$$P\left\{X_j=k\right\}=\frac{e^{-\lambda}\lambda^k}{k!},\ j=1,\ldots,n,\ k=0,1,\ldots \tag{9}$$

The total number of earthquakes in that zone, during n years, is the sum $X_1+\cdots+X_n$. Since the r.v. X_j are assumed to be independent, and have the same law, the sum $X_1+\cdots+X_n$ follows a Poisson Law with parameter $n\lambda$.

So, the answer to the question raised in the "cahier des charges" is that the computation proposed there is correct : under the previous assumptions (Poisson process, independ-

ence of years), the expected number of earthquakes during n years is n times the number of earthquakes expected each year.

IV. Taking several zones into account

Let us now assume that we have several zones Z_i, $i = 1,...,I$. Each zone is characterized by a different parameter λ_i (the predicted number of earthquakes is not the same in each zone) and a different history : the number of years of observation is not the same everywhere.

Let $X_{i,j}$ be the random variable which indicates the number of earthquakes, in zone i, in year j (above a certain level). Then the total number of earthquakes, in the total area, during all years, is the sum :

$$S = \sum_{i=1}^{I} \sum_{j=1}^{n_i} X_{i,j} \tag{10}$$

where n_i is the length of observation (number of years) for the $i-$th zone.

Since all variables $X_{i,j}$ are assumed to be independent (and this means specifically that an earthquake detected in a zone is not counted elsewhere as well), with parameters λ_i, the sum S follows a Poisson Law with parameter :

$$\lambda = \sum_{i=1}^{I} n_i \lambda_i \tag{11}$$

and so the total expected number of earthquakes, in the whole territory, is the sum of the number of earthquakes expected for each zone.

V. Confidence intervals

In order to compare our model with reality, we have to build confidence intervals. Indeed, just say that the expected value is close or not to the observed value is of little significance. If we take someone at random in France, you do not expect him or her to be close, in height, to the average height of the population. We have to present some information about the dispersion of the results, that is about the variance.

Fix some (small) number $\varepsilon > 0$, for instance $\varepsilon = 0.025$. For a given $\lambda > 0$, choose $k_0 \geq 0$ such that :

$$e^{-\lambda} \sum_{k=0}^{k_0} \frac{\lambda^k}{k!} \leq \varepsilon \tag{12}$$

and $k_1 \geq 0$ such that :

$$e^{-\lambda} \sum_{k=k_1}^{+\infty} \frac{\lambda^k}{k!} \leq \varepsilon \tag{13}$$

Then :

$$e^{-\lambda}\sum_{k=k_0+1}^{k_1-1}\frac{\lambda^k}{k!}\geq 1-2\varepsilon \tag{14}$$

which means that :

$$P\left\{k_0+1\leq X\leq k_1-1\right\}\geq 1-2\varepsilon \tag{15}$$

The interval $\left[k_0+1,k_1-1\right]$ is therefore a confidence interval within $1-2\varepsilon$: in our case, with $\varepsilon=0.025$, this gives 95 %. The computation of k_0 and k_1 is made from equations (12) and (13) respectively, using numerical methods. For instance, for $\lambda=79.2$ (value given in Table 1, cahier des charges), we get :

$$e^{-\lambda}\sum_{k=0}^{54}\frac{\lambda^k}{k!}=0.0017 \tag{16}$$

and :

$$e^{-\lambda}\sum_{k=98}^{+\infty}\frac{\lambda^k}{k!}=0.0226 \tag{17}$$

So, the 95% confidence interval is 55 - 97, and the expected value is 79.2. The question is not whether the real value is close or not to 79, but whether it falls or not in the interval 55 - 97. If it does not, we are entitled to consider that there is something wrong in the model.

VI. Numerical computations for the confidence interval

The formulas above can be simplified, or, more exactly, they may become "automatic".

Using the Gamma function :

$$\Gamma(\mathrm{z})=\int_0^{+\infty}t^{z-1}e^{-t}dt \tag{18}$$

and the incomplete Gamma function :

$$\Gamma(\mathrm{a},x)=\int_x^{+\infty}t^{a-1}e^{-t}dt \tag{19}$$

the quantity

$$s_n(\lambda)=e^{-\lambda}\sum_{k=0}^{n}\frac{\lambda^k}{k!} \tag{20}$$

may be written as :

$$s_n(\lambda) = \frac{\Gamma(n+1,\lambda)}{\Gamma(n+1)} \tag{21}$$

which may simplify the numerical computations, since both the Gamma function and the Incomplete Gamma function are easily available in symbolic computation systems (see references [1] and [2] for further properties of these functions).

VII. Using real data

Let us now use the data communicated by the CEA (table 12 of the "cahier des charges"). For a threshold of 1 cm/s2, the MEDD model predicts 253 events. By the previous computation, the 95% confidence interval is 222-284. But the recorded number is 208, which is well outside the interval.

On the other way, the probability that the total number is at most 210 is only 0.003 (using again the parameters of the model).

The same holds with the threshold 5 cm/s2 : the predicted number is 105, which gives a confidence interval of 84-125. The recorded number is 41, well outside this interval.

This indicates that there is a strong difference between predicted and observed data and indicates that, very likely, the model is not correct. Further investigation about the possible explanations will be made in subsequent reports.

We note here that the "attenuation" process has never been used. We just work with the number of earthquakes in a zone, during a year. How these earthquakes have been detected (by a nearby captor or a captor situated far from the source) is not taken into account at this stage.

References

Abramowitz, M. and Stegun, I. A. (Eds.). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing.* New York: Dover, p. 260, 1972.
Arfken, G. "The Incomplete Gamma Function and Related Functions." §10.5 in *Mathematical Methods for Physicists, 3rd ed.* Orlando, FL: Academic Press, pp. 565-572, 1985.