IV. МОДЕЛИ СЕЙСМИЧНОСТИ

УДК 550.341+519

PROBABILITIES AND INFORMATION GAIN FOR EARTHQUAKE FORECASTING

D. Vere-Jones
Institute of Statistics and Operations Research,
Victoria University of Wellington, New Zealand

The paper discusses some issues arising out of attempts to calculate and score regular regional forecasts for earthquake probabilities. Given a conditional intensity model for earthquake occurrence, the model is first used to simulate occurrence patterns over the forecast interval. Then the simulations are used to estimate the required occurrence probabilities. A simple binomial score is suggested for monitoring and evaluating the performance of the probability forecasts. It is shown that an upper bound for the average score is provided by the information (or entropy) rate of the model. Similarly the improvement in the score over a standard model (constant rate Poisson with independent magnitudes) is bounded above by the entropy gain. Rates can be per unit time or per event. The performances of the ETAS and Stress Release models are described, and used to illustrate how the effectiveness of the forecasts depends on the type of model, the timing of the forecasts and the choice of forecast interval.

ПОВЫШЕНИЕ ВЕРОЯТНОСТИ И ИНФОРМАЦИИ В МОДЕЛЯХ ПРОГНОЗА ЗЕМЛЕТРЯСЕНИЙ

Д. Вере-Джонс
Институт статистики и исследования операций,
Университет Королевы Виктории, Веллингтон, Новая Зеландия

В работе обсуждаются вопросы, возникшие из попытки оценить вероятности землетрясений в региональном прогнозе. Сначала рассматривается задача численного моделирования последовательности землетрясений с помощью модели условной интенсивности с дискретизацией времени, определяемой прогнозом. Далее для оценивания качества прогноза используется простейший энтропийный показатель. Показано, что верхняя граница для среднего показателя качества прогноза определяется средней скоростью создания информации (энтропии) в исходной модели. Аналогично, относительное повышение качества по отношению к стандартной Пуассоновской модели ограничено сверху относительным изменением удельной энтропии. В заключение на двух статистических моделях показано, как эффективность прогноза зависит от типа модели, времени прогноза вперед и временно́й базы прогноза.

INTRODUCTION

Over the last few years, attitudes towards earthquake prediction have swung from a state of high optimism in the 1970's to one of considerable current caution. While there is an accumulation of evidence that major earthquakes are frequently preceded by a variety of unusual features, on widely varying time scales, the problem is that they do not do so consistently. Increasingly, emphasis has swung away from trying to discover exact predictors to the development of models from which the probabilities of future events can be estimated. In this sense, improvements in the identification of regions particularly prone to earthquake shaking may be seen as a form of earthquake prediction, albeit on a long time scale. On shorter time scales, stochastic models for earthquake occurrence provide low-grade predictions. Although the performance levels are still too low to be practically useful, this approach seems likely to be the vehicle through which more effective procedures are identified and developed. Indeed, it will be a vehicle hard to avoid if seismologists follow Aki [1], in seeing the essential role of the scientist in earthquake prediction as developing models for estimating the probabilities of future earthquakes.

Much in this spirit, the present paper looks at some aspects of the provision and evaluation of probability forecasts. It assumes that a model for the earthquake process is available in the form of a stationary, marked point process with explicitly specified conditional intensity function, giving current risk (in the sense of an intensity or probability per unit time) as a function of past observations (see [2] for applications and [3] for theoretical background). The conditional intensity is used to compute, via repeated simulations, occurrence probabilities for a prescribed sequence of forecasting periods. A simple binomial score is suggested for evaluating the forecasts. It is then shown that the expected score is bounded above by the entropy rate for the model. This implies that the properties of the model itself put bounds on the performance of any probability forecasting scheme based on that model. This result is closely related to the use of the average likelihood as a measure of the information available for forecasting, as suggested in Kagan and Knopoff [4]; here we try to develop this theme more systematically.

There is a close relation between the development of probability forecasts and the application of decision theory concepts to the issuing of alerts or other earthquake countermeasures. Several recent papers have discussed earthquake prediction from a decision point of view (see, for example, Molchan [5] or the summary in [2]). One important technical point which emerges from such discussions is that, in simple situations at least, the decision on whether or not to take a particular action is best based on values of the conditional intensity function. Indeed, it is this function which lies at the heart of all probability forecasting procedures for processes of a point-process character. In the present context it determines both the likelihood and the entropy rate for the model.

The material in the remainder of the paper is ordered as follows. In the next section we describe in more detail the class of models and probability forecasting procedures that we have in view. In §2 we examine scoring methods for such procedures and their relation to the entropy rate. In §3 we illustrate the use of the procedures with respect to two widely-used models, Ogata's "epidemic type aftershock sequence (ETAS)" model

(see, for example, [6] and [7]) and the "stress release model" (SRM) developed by the author and colleagues for application to historical earthquake data ([8–10]). The examples throw up a number of issues relating to the timing and duration of probability forecasts.

1. MODELS AND FORECASTS

We suppose that the aim of the forecasting exercise is to provide estimates, on a regular, regional basis, for the probability that an earthquake within a given magnitude band will occur within a selected region and a given forecast interval. We suppose also that the forecaster has available data on the times, magnitudes, and locations of past events, and a fully determined earthquake occurrence model. In particular the latter should determine probabilities for the occurrence of events, conditional on the observed past history. Note again that our aim is to illustrate how probability forecasts can be generated and how they perform, not to suggest that current models are particularly adequate for such a purpose. We shall return in the final section to a brief discussion of how the models might be modified and improved.

If we formulate the earthquake process in point process terms ([2,3]); the latter outlines the basic mathematical theory and can be referred to for definitions and terminology), each earthquake can be regarded as a point in time-location-magnitude space. Additional coordinates, such as the direction of slip along a fault plane, could be considered in principle, but will not be discussed here. For simplicity, we shall disregard location within the forecast region, and suppose that any influence from occurrence of earthquakes outside the region is incorporated into the past history.

We are then left with a marked point process in time-magnitude space, the points (events) being characterised by an origin time t_i and a magnitude M_i. The conditional intensity for such a process is given by

$$\lambda(t, M)dt\, dM = E[N(dt \times dM) \mid \mathcal{H}_t],$$

where the conditioning is taken with respect to the sigma-algebra $\mathcal{H}_t$ of events defined on the past of the process, extended if necessary to include information on relevant external variables. It will often be convenient to rewrite the conditional intensity in the form

$$\lambda(t, M)dt\, dM = \lambda(t)\, f(M|t)dt\, dM,$$

where

$$\lambda(t)dt = E[N(dt) \mid \mathcal{H}_t]$$

denotes the overall conditional intensity (for events of all magnitudes considered relevant; we assume it is finite), and

$$f(M|t) = E[N(dt \times dM) \mid \mathcal{H}_t] \,/\, E[N(dt) \mid \mathcal{H}_t]$$

is the magnitude distribution, conditional on the past history $\mathcal{H}_t$.

We now suppose given a sequence of magnitude ranges $\{(M_k, M_{k+1}), k = 0, 1, \ldots, K-1\}$, and forecasting intervals $(t_i, t_i + \delta_i)$. As the process evolves in time,

the forecasting task is to produce for each interval a a set of forecasts (one for each magnitude range) of the form "*There is probability* $p_{k,i}$ *that an event with magnitude in the range* (M_k, M_{k+1}) *will occur within the time interval* $(t_i, t_i+\delta_i)$ *under the constraint* $\sum p_{k,i} = 1$ *for each* i". We shall suppose for convenience that the forecast intervals are contiguous.

We shall generally assume that the interval δ is small enough for the probability of two or more events occurring within the interval to be small. This is not an essential assumption, and would be unrealistic within an aftershock sequence (for example), or if very small magnitudes were being considered. However it may serve to indicate the type of forecasting situation that we have in mind. Other types of forecast could certainly be considered - for the time to the next event of a given magnitude, for example, or for the expected number of events within a certain magnitude range - and could be produced equally easily by the procedures outlined below.

The simulation proceeds in the standard manner for point processes (see, for example, [11] or [12]). First, any parameters to the model are fitted from observations on the process up to the present time t, which we may take as the beginning of one of the required forecast intervals. Then, the history up to the present (t) is augmented by supposing that no additional events occur in the period $(t, t+x)$. Let the corresponding extension of the conditional intensity be denoted by $\lambda^*(t+x)$. Then $\lambda^*(t+x)$ acts as the hazard function for the time to the next event, which therefore has distribution function $1 - \exp[-\int_t^{t+x} \lambda^*(t+u)du]$. The simulations can then proceed as follows:

1. Simulate a time τ to the next occurring event using $\lambda^*(t+x)$. (Since its distribution is known as above, any of the standard methods, such as the inverse method or the thinning method, can be used here. The most convenient method will generally depend on the form of the conditional intensity function).

2. Simulate a magnitude M for the event at this time, using the appropriate conditional distribution $f(M|t+\tau-)$, with the history augmented by the assumption of no event in the interval $(t, t+\tau)$. Again, any standard method can be used.

3. Add the new event to the history, and return to step (1), replacing t by $t+\tau$.

4. Continue until the end of the time interval for forecasting, $(t, t+\delta)$, has been exceeded.

Each such simulation produces one possible scenario for the time interval $(t, t+\delta)$. From a large number - typically at least 1,000 - of such scenarios, the proportions $\hat{p}_k$ containing an event in the magnitude range (M_k, M_{k+1}) can be determined and used as estimates of the required forecast probabilities.

The whole process can be started again (updating the real history) at the beginning of the next forecast interval.

The only limitation on this procedure is the requirement of a model with a fully specified conditional intensity function. If conditioning on external variables is used, the model must be complete enough to allow the possible evolution of the external variables to be figured into the projected form of the conditional intensity function. The low level of understanding of the physical processes involved makes it very difficult to develop such extended models at the present time.

2. SCORING PROCEDURES AND INFORMATION GAIN

We suppose that the earthquake process can be modelled by a stationary, ergodic point process, and start by considering the simpler case of an unmarked point process, with condtional intensity function $\lambda(t)$. Assume that forecasts p_i for a sequence of intervals $(t_i, t_i + \delta_i)$ are available, produced by the preceding method or in any other way, together with the actual results X_i, recorded as a 1 if an event occurs in the forecast interval, and as a 0 otherwise. The results are analogous to those of a sequence of Bernoulli trials, with corresponding probabilities p_i. A natural score is then the likelihood for the sequence of trials, namely the *binomial score*

$$B = \sum [X_i \log p_i + (1 - X_i) \log(1 - p_i)].$$

The first term represents the contribution to the score from those intervals which contain an event, and the second term the contribution from the intervals which contain no event.

Suppose that the intervals cover a total period $T = \sum \delta_i$. Then B/T may be regarded as an estimate of the information per unit time generated by the process. For most purposes it is preferable to work with generalized entropies, or the information gain relative to a reference process. In the present case the natural reference process is a Poisson process with the same mean rate $\bar{\lambda} = E[\lambda(t)]$ as the original process (note that the right hand side of this equation is independent of t, because of the assumption of stationarity, and could equally well be replaced by $\lambda(0)$). Denoting by $\bar{p}_i = 1 - exp(-\bar{\lambda}\delta_i)$ the probability estimate for the i-th forecasting interval for this process, consider the averaged difference between the binomial scores, namely the quantity

$$\bar{\rho}_T = (B - \bar{B})/T = (1/T) \sum \left[X_i \log \frac{p_i}{\bar{p}_i} + (1 - X_i) \log \frac{1 - p_i}{1 - \bar{p}_i} \right].$$

We call this the *mean information gain per unit time* associated with the forecasting procedure. The ratio $p_i/\bar{p}_i$ is the *probability gain* for the i-th interval. It follows from Jensen's inequality that the expected value of the mean information gain is non-negative – on average, we must do better by incorporating relevant information from the past history.

Now recall that the *entropy rate* for a stationary, ergodic point process with (complete) conditional intensity $\lambda(t)$ is defined to be

$$H = -E[\lambda(0)\{\log \lambda(0) - 1\}]$$

(see [3], p. 569). A version of the law of large numbers (more precisely, McMillan's theorem) for point processes implies that this quantity can also be obtained as a long-term time average of the logarithm of the point-process likelihood - see [13] or [3],Theorem 13.5.IX. It is this idea which lies behind Kagan's use of the averaged loglikelihood as a measure of the information rate for the process (e.g. [4]). The corresponding generalized entropy rate, relative to the Poisson process with the same mean rate, is

$$E[\lambda(0) \log(\lambda(0) - \bar{\lambda} \log \bar{\lambda}] = E[\lambda(0) \log\{\lambda(0)/\bar{\lambda}\}] = \mathcal{I}.$$

We shall call $\mathcal{I}$ the *entropy gain per unit time.* Then we have the following result.

Proposition 1.

For all subdivisions of the interval $(t, t+T)$, the expected value of the mean information gain per unit time is bounded above by the entropy gain per unit time, $E(\bar{\rho}_T) \leq \mathcal{I}$.

This follows from results of Czizsar and Fritz on discrete approximations of entropies (see [14,15] or [3] Ex 13.5.3). In fact the result holds even for partitions generated by a sequence of stopping times for the histories $\mathcal{H}_t$. The essential point is that the partition used should be imbeddable into a sequence of partitions generating the Borel sets in $t, t+T)$, so that the σ-algebras of events defined on counts in the partitions should generate the full σ-algebra of events defined on the point process in this interval.

It is also true that $E(\bar{\rho}_T) \to \mathcal{I}$ as the partitions are refined (with $\max \delta_i \to 0$), and that $\bar{\rho}_T \to \mathcal{I}$ as $T \to \infty$. Again we refer to the references quoted above for details.

These results imply that the average score should approach its upper bound when the forecasting intervals become very small. This suggests that it should be advantageous to use small forecasting intervals. In practice, however, the use of small intervals may be unrealistic for several reasons. This suggests the following possible generalisation of the forecasting problem : *given any constraints of cost, etc, find a rule for choosing $\mathcal{H}_t$-predictable intervals (t_i, δ_i) which maximises the expected value of the mean information gain, subject to those constraints.* Since we have not specified the constraints, this remains somewhat vague, but we shall gain some hints as to its solution in examining the examples in Section 3.

Instead of dividing by the total time T to obtain the mean score, we may divide by the total number of events $N(T)$. This leads to an average score per event, and comparing to the Poisson we may form as above the average information gain per event. In this case the corresponding upper bound and limit is the entropy gain per event, namely the quantity $\mathcal{I}/\bar{\lambda}$ (see Daley and Vere-Jones, Proposition 13.5.X). Now when the δ_i are very small, only the first sum contributes significantly to the binomial score, so that the average score per event is approximately equal to the average of the logarithms of the probability gains $(p_i/\bar{p}_i)$ taken over the intervals containing events. In other words, the geometric mean of the probability gains for each event is approximately bounded above by $\exp(\mathcal{I}/\bar{\lambda})$. More precisely, we have the following corollary to the preceding result.

Proposition 2.

$$E\left\{N(T)^{-1} \sum_{1}^{N(T)} \log\left(\frac{p_i}{\bar{p}_i}\right)\right\} \approx E\left\{\frac{T}{N(T)}\bar{\rho}_T\right\} \leq \mathcal{I}/\bar{\lambda}.$$

Now let us return to the marked process case. For each magnitude interval (M_k, M_{k+1}) and forecasting interval $(t_i, t_i + \delta_i)$ we can define a binomial score as above, say $B(k)$. We can either study these scores separately, or combine them to form a total score S by summing over the magnitude classes. The mean information rate per unit time is then given by S/T.

To convert these to information gains we need an appropriate reference process. In the present context, such a process is the marked Poisson process with mean rate $\bar{\lambda} = E[\lambda(0)]$, and independently distributed magnitudes with common density function

$$f(M) = E[\lambda(0, M)]/E[\lambda(0)].$$

Setting

$$\bar{p}_i(k) = 1 - \exp(-\bar{\lambda} \int_{M_k}^{M_{k+1}} f(M)dM),$$

the average information gain per unit time for the process as a whole takes the form

$$\bar{\rho}_T = (1/T) \sum_k \left\{ \sum \left[X_i(k) \log \frac{p_i(k)}{\bar{p}_i(k)} + (1 - X_i(k)) \log \frac{1 - p_i(k)}{1 - \bar{p}_i(k)} \right] \right\}.$$

The entropy rate for the overall process is now given by

$$H = -E \left[\int \lambda(0, M) \log \lambda(0, M) dM - \int \lambda(0, M) dM \right] =$$

$$= -E[\lambda(0) \log \lambda(0) - \bar{\lambda}] + E[\lambda(0) \int f(M|0) \log f(M|0) dM]. \quad (1)$$

The first term in (1) is the entropy rate for the process without marks, i.e. the sequence of times only, while the second term defines an entropy rate for the magnitudes, given the times.

The corresponding generalized entropy rate, or expected information gain, relative to the marked Poisson process with independent magnitudes, is given by

$$\mathcal{G} = E[\lambda(0) \log(\lambda(0)/\bar{\lambda})] + E[\lambda(0) \int f(M|t) \log\{f(M|t)/f(M)\} dM,$$

where again the first term is the rate of information gain from the times, and the second term is the rate of information gain from the magnitudes, given the times. As in the earlier discussion, the results on discrete approximations to the entropy imply that $\mathcal{G}$ forms an upper bound to the expected value of the mean information rate per unit time, as well as a limit when the time and magnitude intervals become very small.

If we restrict attention to the point process formed by events falling into a single (say, the k-th) magnitude class, its $\mathcal{H}_t$-intensity is

$$\lambda_k(t) = \int_{M_k}^{M_{k+1}} \lambda(t, M) dM = \lambda(t) \int_{M_k}^{M_{k+1}} f(M|t) dM.$$

Note that this conditional intensity may depend on contributions from events outside the its intensity class, so that it is not the same (in general) as the conditional intensity that would be obtained by taking this process of events in isolation. The reference point process for these events is the constant rate Poisson process with rate

$$\bar{\lambda}^k = E[\lambda_k(0)] = \bar{\lambda} \int_{M_k}^{M_{k+1}} f(M) dM.$$

From these quantities we can form the expected information gain per unit time, say $\mathcal{G}_k$, for events in magnitude class k, just as we did for the simple point process considered at the beginning of the section. Note that

$$\sum \mathcal{G}_k \leq \mathcal{G},$$

the difference representing the loss of information due to working with discrete magnitude classes. In principle, the upper bounds derived above can be calculated from distributional properties of the process, but explicit analytical expressions are rarely available, and it will usually be necessary to resort to numerical methods (eg a Monte-Carlo simulation to estimate the expectations). A quicker and more convenient, if approximate, method is to revert to Kagan's original suggestion and estimate the entropy rate from the mean log-likelihood. To see the rationale for this, recall that, for a simple point process, the log-likelihood $\log L(0,T)$ can be written in the form

$$\int_0^T \log \lambda(t) dN(t) - \int_0^T \lambda(t) dt = \int_0^T \log \lambda(t) \, [dN(t) - \lambda(t) dt] + \int_0^T \lambda(t) [\log \lambda(t) - 1].$$

Taking expectations, the first term vanishes (it is a martingale, with increments having conditional expectations zero), while from stationarity the second reduces to $TE[\lambda(0)(\log \lambda(0) - 1)]$. Dividing by T makes the application of the law of large numbers (subject to standard regularity conditions) quite tranparent. A similar argument applies to the marked point process case.

3. APPLICATIONS

3.1. Stress release model

We shall consider the simple form of stress release model used by Zheng and Vere-Jones ([9,10]) to describe a single region with a single magnitude frequency law. It is governed by a Markov process $X(t)$ assumed to represent the total (ambient) stress in a given seismic region. This is assumed to increase gradually (assumed linearly) as a result of external tectonic forces, and to decrease by jumps each time an earthquake occurs, the size of the jump being related to the magnitude of the earthquake. The key parameters are the family of jump distributions $j(y|x)$ for the size of the decrease, given that the jump occurs when $X(t) = x$, and the risk function $\psi(x)$, where $\psi(x)dt$ gives the probability of a jump occurring in $(t, t+dt)$ when $X(t) = x$).

The density $f(x,t)$ of the distribution of $X(t)$ satisfies the forward equation

$$\frac{\partial f}{\partial t} + \rho \frac{\partial f}{\partial x} = -\psi(x) f(x,t) + \int_x^\infty f(y,t) \psi(y) j(y-x|y) dy.$$

A careful discussion of the behaviour of the process is given by Zheng [16], who also gives references to earlier contributions. We shall make the simplifying assumptions that $j(y|x)$ is independent of x (the empirical evidence suggests at least that the dependence is not strong), and that it has a finite second moment. The process $X(t)$ then has a stationary ergodic version with stationary distribution $f(x)$ satisfying the equation

$$\rho f(x) = \int_x^\infty \psi(y) f(y) S_J(y-x) dy, \tag{2}$$

where $S_J(x)$ is the survivor function of the jump distribution $j(x)$ (see Vere-Jones (1988)).

Integrating (7) over $(-\infty.\infty)$ yields directly

$$\bar{\lambda} = E[\psi(X(0))] = \int_{-\infty}^{\infty} \psi(x) f(x) dx = \rho / E(J),$$

a result which can also be derived by combining the law of large numbers with the observaton that the difference $\rho t - S(t)$ must remain bounded in expectation as $t \to \infty$. Similarly, multiplying through (2) by x and simplifying yields

$$E[\lambda(t) \log(\lambda(t)] = E[X\psi(X)] = \bar{\lambda} E(X) + \frac{\rho}{2} \frac{E(J^2)}{[E(J)]^2}.$$

It is surprising that these results are independent of the particular functional forms of both $\psi(x)$ and $j(x)$, although the expression for the entropy rate does depend indirectly on these though the value of $X(t)$. Unfortunately, however, we have not been able to find a closed form expression for $E(X)$ even in the simple case when $\psi(x)$ has exponential form.

In practice, the stress level $X(t)$ is not known, but has to be inferred as far as possible from the observations on the times and sizes of events. This means that it is not obvious that $X(t)$ belongs to the σ-algebra $\mathcal{H}_t$. It is one of the many remarkable features of this process that in fact $X(t)$ can be reconstructed from the observations. To see how this is possible, consider the special case we have generally used in applications, where $\psi(x) = \exp(\alpha + \beta x)$. We can then write the conditional intensity in the form

$$\lambda(t) = a + b(t - \gamma S(t)),$$

where $S(t)$ is the sum of the jumps over the time interval $(0, t)$, a subsumes the value of $X(0)$ as well as α, and $\gamma = \rho^{-1}$. The parameters a, b, γ can be consistently estimated by standard likelihood techniques (see [9]), that is as functions of the observations, and from these the past values of $X(t) = (t - \gamma S(t))/\gamma$ can be reconstructed. Hence we can treat $X(t)$ as an element of $\mathcal{H}_t$ and use the result derived above for the mean entropy.

To illustrate how the forecasting techniques might work in practice, we have taken the parameter values for the stress-release model fitted to the N. China data in [9], and simulated a 1,000 year record for use as a base data set. The magnitudes were simulated by randomly sampling from a Gutenberg–Richter relation truncated above at $M = 9$

$$\theta \exp(\theta(M_0 - M)/[1 - \exp(\theta(M_0 - 9))], \quad M_0 < M < 9,$$

and converting these to stress releases using the approximate formula $S_i = 10^{0.75(M_i - M_0)}$. Using the same parameter values, namely

$$a = -2.455, \ b = 0.0113, \ \gamma = 0.151, \ M_0 = 6, \ \theta = 1.8,$$

forecasts were then produced for 5-yearly intervals. Table 1 shows a typical subset of periods, with asociated probability estimates, binomial scores, and information gains over the Poisson process with the same rate, and independent magnitudes. The mean information gains for the different magnitude classes, and for the process as a whole, are also listed. The probability curves are shown in Figure 1.

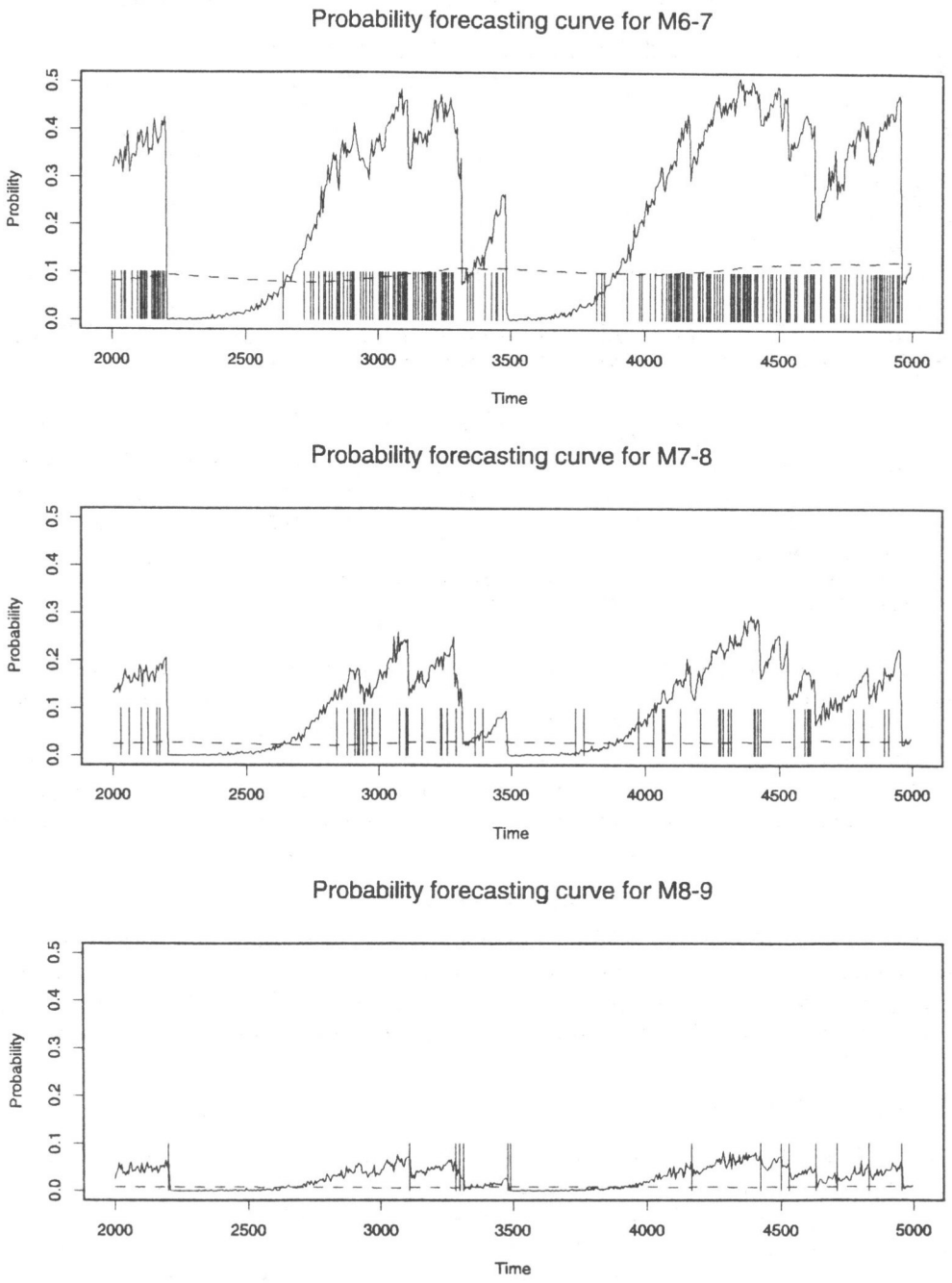

Fig.1. Probability forecasts for SRM model

Continuous lines represent probabilities estimated from the model, dashed lines represent probabilities estimated from background rate

Table 1: Scores and Information Gains for Stress Release Model

A: Sample scores for individual forecasts height

Magnitude	N	X	p	B	B^*	G
$6 \le M < 7$	2	1	0.61	-0.49	-1.29	0.80
$7 \le M < 8$	0	0	0.20	-0.23	-0.09	-0.14
$8 \le M$	0	0	0.04	-0.04	-0.02	-0.02
total	-	-	-	-0.76	-1.40	0.64
$6 \le M < 7$	2	1	0.62	-0.48	-1.28	0.80
$7 \le M < 8$	1	1	0.23	-1.47	-2.48	1.01
$8 \le M$	0	0	0.04	-0.04	-0.02	-0.02
total	-	-	-	-1.99	-3.74	1.75

N – the number and X the presence of events in the particular magnitude and time intervals, B, B^* – the binomial scores for the SRM and simple Poisson models, G – the information gain

B: Average scores and information gains

Magnitude	N_S	$\bar{B}_S$	$\bar{G}_S$	N_F	$\bar{B}_F$	$\bar{G}_F$	N	$\bar{B}$	$\bar{G}$	$\bar{G}_E$
$6 \le M < 7$	185	-1.06	0.39	415	-0.23	0.030	600	0.48	-0.14	0.45
$7 \le M < 8$	51	-1.92	0.79	549	-0.014	-.038	600	-0.26	0.033	0.39
$8 \le M < 9$	15	-3.26	0.66	585	-.028	-.007	600	-0.11	.009	0.37

N_S – the number of success intervals, $\bar{B}_S$ the average score and $\bar{G}_S$ the average gain in such intervals, $N_F, \bar{B}_F, \bar{G}_F$ – the corresponding quantities for the failure intervals, $\bar{G}, \bar{G}_E$ – the average gain per time interval and per success

The information gains for individual intervals range from something below zero to around 2, with a mean of around 0.4 in all three magnitude intervals. The corresponding upper bound, deduced from the mean log-likelihood ratio per event or from the theoretical results above, is 0.48, suggesting that we are losing a little predictive power by blocking into 5-year intervals, but not too much. The geometric mean of the probability gains is around 1.7, consistent with fluctuations of the risk enhancement factor (ratio of the conditional intensities) from below 1 to around 5.

3.2. ETAS model

The ETAS model is intended for use on much smaller time scales, and aims to capture the increase in risk due to aftershocks following an initial event. The conditional intensity at any time t is represented as the sum of a background intensity μ and contributions from each of the preceding events. These contributions are assumed to have the form

$$Ke^{a(M_i - M_0)}(t + c - t_i)^{-p},$$

where M_0 is a threshold magnitude (lower bound to the events considered) and K, a, c, p are parameters, and M_i, t_i are the magnitudes and times of the earlier (before time t) events. The parameter K is a measure of the strength of the clustering, a determines the sensitivity of the size of the cluster to the initiating magnitude, c is a scale parameter and p is a shape parameter for the power-law decay in intensity after the initiating event. The total intensity at time t is given by the sum

$$\lambda(t) = \mu + K \sum_{i:t_i<t} e^{a(M - M_i)}(t + c - t_i)^{-p}.$$

Further background and applications are described in Ogata's papers, for example [6,7].

As in the previous example, we simulated a base catalogue of some 1000 events, using parameter values

$$\mu = 0.030, K = 0.0286, a = 1.31, c = 0.0074, p = 1.246,$$

(time measured in days) obtained from fitting the model to five years' of data with local magnitudes $M \geq 3$ from the Wellington region. Again we assumed independently generated magnitudes. Here the interest is in short-term forecasting, so we have applied the forecasts at 2-day intervals. Results analogous to those of the preceding example are set out in Table 2 and Figure 2.

Table 2: Scores and Information Gains for ETAS Model

A: Sample scores for individual forecasts

Magnitude	N	X	p	B	B^*	G
$2.5 \leq M < 3$	0	0	0.20	-0.22	-0.39	0.17
$3 \leq M < 4$	0	0	0.11	-0.12	-0.12	-0.00
$4 \leq M < 6$	0	0	0.01	-0.01	-0.01	-0.00
total	-	-	-	-0.35	-0.52	0.17
$2.5 \leq M < 3$	12	1	0.16	-1.86	-1.12	-0.73
$3 \leq M < 4$	4	1	0.09	-2.42	-2.18	-0.23
$4 \leq M < 6$	1	1	0.01	-5.16	-5.06	-0.05
total	-	-	-	-9.39	-8.37	-1.02
$2.5 \leq M < 3$	1	1	0.34	-1.07	-1.11	0.03
$3 \leq M < 4$	0	0	0.49	-0.68	-0.12	-0.56
$4 \leq M < 6$	0	0	0.02	-0.02	-0.01	-0.01
total	-	-	-	-1.76	-1.24	-0.53

Interpretation as for Table 1A

B: Average scores and information gains

Magnitude	N_S	$\bar{B}_S$	$\bar{G}_S$	N_F	$\bar{B}_F$	$\bar{G}_F$	N	$\bar{B}$	$\bar{G}$	$\bar{G}_E$
$6 \leq M < 7$	410	-1.98	-0.83	1590	-0.14	0.24	2000	-0.52	0.02	0.10
$7 \leq M < 8$	148	-2.57	0.10	1852	-0.075	0.043	2000	-0.26	0.01	0.10
$8 \leq M < 9$	15	-5.13	-0.26	1988	-0.001	0.001	2000	-0.04	-0.00	-0.00

Interpretation as for Table 1B

It is immediately apparent that the forecasting performance here is very poor, although it does improve for the 1-day intervals. Commonly the score for periods containing events is actually worse than for the Poisson. The reason for this is not difficult to understand. Most of the forecast periods start at a time of relative quiescence, not during an aftershock sequence. If the interval contains an event, it will be scored at the rate of the background seismicity μ which is below the mean rate $\bar{\lambda}$ on account of the clustering. Its contribution to the information gain will therefore be negative. The Poisson model sufferes a compensating loss for the intervals not containing events, so that the overall information gain is still positive, though small. Overall, the mean information gain per event interval is around 0.1, which compares poorly with the mean log-likelihood ratio per event of 1.61.

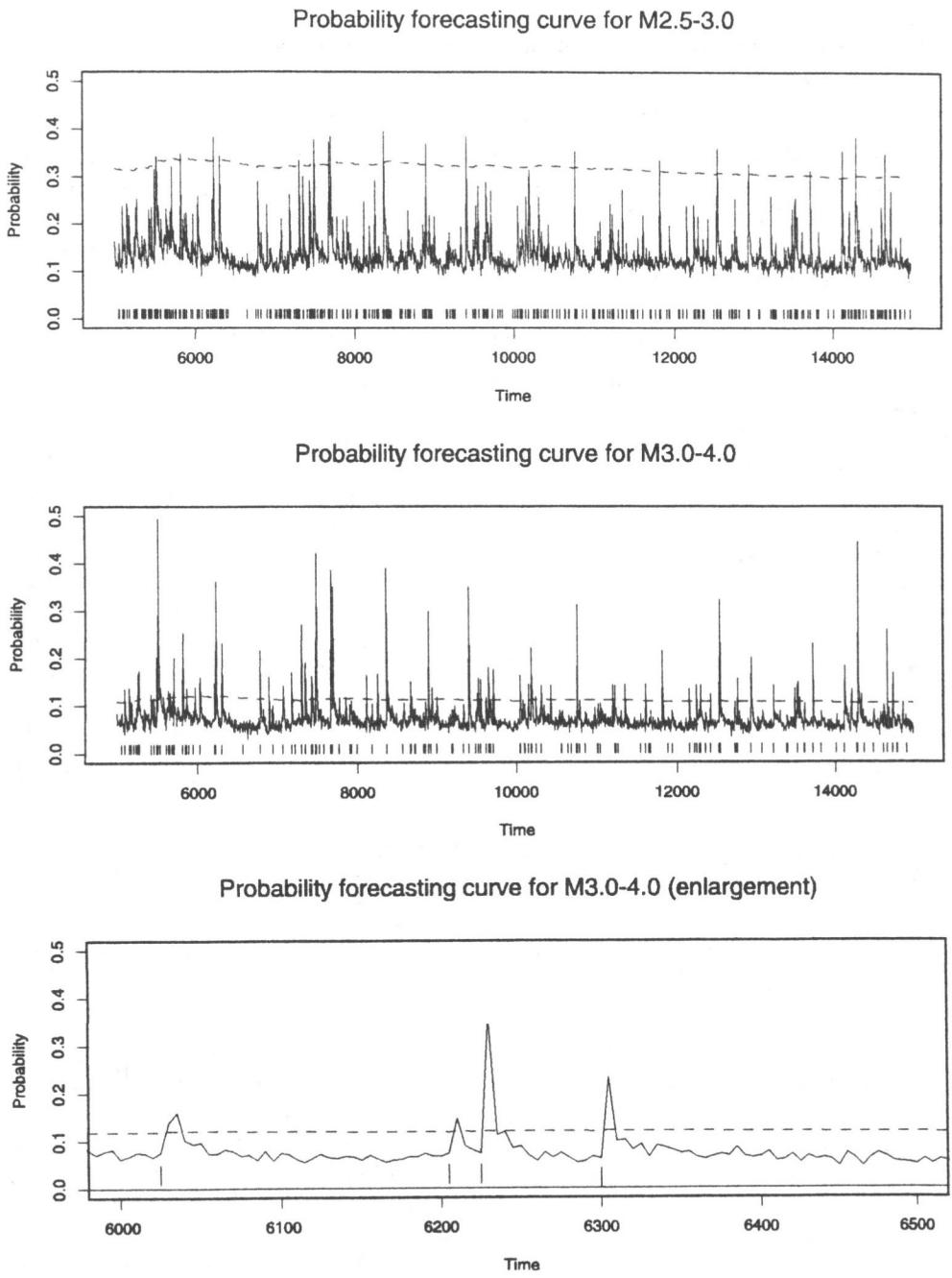

Fig.2. Probability forecasts for ETAS model

Continuous lines represent probabilities from model, dashed lines represent probabilities estimated from background rate

CONCLUDING REMARKS

Evidently the forecasting strategy used in the last example is failing to extract the predictive information that is potentially available in the model. In retrospect, this is hardly surprising, since the ETAS model is designed to fit aftershocks, whereas we used it to forecast events in general 2-day intervals. One way to use the model more effectively might be to issue a forecast immediately following any event, or more particularly any event above magnitude (say) 4. This would hardly be construed as a breakthrough for earthquake prediction, but it would significantly improve the scores.

The two examples illustrate, in somewhat exaggerated form, some of the difficulties of setting up a systematic forecasting scheme, even in the presence of a well-specified and effective model. One type of difficulty is represented by short-term clustering present in the ETAS model. It shows that a forecasting scheme cannot hope to be effective unless the decision rules for calling or terminating a forecast are flexible enough to allow the use of incoming information on the evolution of the process. In this example, in order to utilize the predictive power of the model, it is essential to issue forecasts as soon as possible after the occurrence of an event. No scheme based on predetermined intervals will be effective unless the intervals are impracticably small. From such an example we can conclude that effective forecasting procedures have to have an updating character. The forecast may have the form "the probability of an event in the next δ days is p", but it may have to be superceded before that time period is up. In particular, once a forecast event has occurred, the performance of the forecasting procedure can *never* be improved by waiting out the full period; it will always be advantageous (if not always practicable) to revise the forecast immediately after the event occurs. More generally, it may be desirable to update the forecast in response to substantial changes in the conditional intensity.

A different type of difficulty, which might be described as the *ripple effect* arises when the conditional intensity, though remaining small in absolute terms, is known to fluctuate in a predictable way. If, for example, major events were five times more likely to occur on Sunday mornings than on any other day, but still only likely to occur once every twenty years, could this information be usefully incorporated into the forecasts? Obviously it might be useful for some types of decision-making or risk reduction measures, but it does not lend itself to incorporation in the type of forecasting scheme we have been considering, where one would tend to work with longer intervals inside which the short-term features were lost.

As suggested earlier, one part of the problem may lie in the rather rigid structure of the forecasting scheme we have examined which is at fault. In particular the use of forecast intervals of predetermined length imposes a restriction which is not really inherent in the problem. In [8], it was argued that the conditional intensity itself, in the form of a probability per unit time, should be used directly as the subject of the forecast. A forecast might then take the form "The current daily risk for an earthquake in the magnitude range 6-6.5 is about 1 in 1000, which is about 3 times the long-term average risk; on Sunday mornings it increases to 15 times the average risk." One is not then tied to intervals of predetermined length. Such a forecast could remain in force until it was superceded by a later one. The times at which the forecasts were modified would form a sequence of stopping times, which could be determined sequentially in such a way as to minimize some cost function. Again one might anticipate that the optimal decision rule would hinge on values of the conditional intensity.

The other aspect of interest is the extent to which the performance of such forecasting schemes could be improved by refining or extending the models. The two examples we have considered are at opposite extremes of the time scale, and neither by itself offers a basis for practical forecasts. The stress release model, in the simple form used here, produces probability gains of at most one order of magnitude. The ETAS model allows one to forecast aftershocks, but little else. Neither has any great predictive power; they were chosen to illustrate principles. On the other hand, there are at present some half-dozen models, or procedures, each of which offers the prospect of gains of a similar order of magnitude. These would certainly include foreshocks, M8-style activation, precursory quiescence, precursory swarms, accelerated moment release and possibly others. They apply at different time-scales, have been tested in different environments, all have limitations and uncertainties associated with them. Nevertheless it is not a great step to a situation where some one or more of these is improved to the extent that it offers probability gains of the order of 20-50, rather than 2-5 as at present. Even now it is possible that by effective combinations of procedures gains of this order of magnitude could be achieved. Three issues then seem to stand out as being crucial for further progress:

1. Attempts need to be made to embed into conditional intensity models those procedures which at present are purely empirically based. The effort to find the model is nearly always insightful, and once found it offers a sounder basis for decision-making, and allows a greater range of decisions to be considered, than a purely empirical procedure.

2. The problem has to be considered of combining forecasts based on different models and sources of information. It seems altogether too sanguine to expect that a solution to the prediction problem will be found, which will answer all practical issues simultaneously. The problem of combining forecasts is one faced by most forecasters (cf the brief discussion of this point in [2]), and it inevitably involves an element of expert judgement. Formally, this can be accomplished through weighting the forecasts within a generalised Bayesian scheme. In a real context, however, a single weighted forecast may have less value than other forms of collective discussion, since it is difficult to capture in any single formula all the possible aspects and interactions that may need to be considered. Even here, however, quantitative probability estimates make an excellent starting point and a check for further procedures.

3. The problem of transferring the information needs careful attention. There are many different possible users - government agencies, utilities such as gas and electricity, local government, civil defense, operators of critical facilities, design and planning groups, insurance, etc. It is not at all clear what levels of probability gain, on what type of time scale, would be of most use to different users; in addition there may be little appreciation within a given user organisation of how to use a probability estimate. Indeed, in any given context, it might take quite a deep study to determine how best and at what levels such information could be used. Until such studies are undertaken, the scientist has to operate in a kind of vacuum, not knowing exactly what estimates are really important for the users.

Acknowledgments. The procedures for implementing the probability forecasts for the stress release and ETAS models were developed by A. Tokeley, J. Liu, and J. Zhuang. Their help and insights are gratefully acknowledged. The routines are available within an S-Plus statistical seismology library, based principally around conditional intensity models, being developed by David Harte (David.Harte@vuw.ac.nz) from whom further information can be obtained. This work forms part of a NZ-China scientific collaboration project between the Institute of Statistics and Operations Research, Victoria University of Wellington, the New Zealand Institute of Geological and Nuclear Sciences, and the Centre for Analysis and Prediction, State Seismological Bureau, China. Financial assistance from the Asia 2000 Foundation of New Zealand and the New Zealand Foundation for Research, Science, and Technology is gratefully acknowledged.

REFERENCES

1. *Aki K.* Ideal probabilistic earthquake prediction // Tectonophysics. 1989. Vol.169. P.197-198.
2. *Vere-Jones D.* Forecasting earthquakes and earthquake risk // Internat. J. Forecasting. 1995. Vol.11. P.503-538.
3. *Daley D.J., Vere-Jones D.* An introduction to the theory of point processes. N-Y.: Springer-Verlag, 1988. 702 p.
4. *Kagan Y.Y., Knopoff L.* Earthquake risk prediction as a stochastic process // PEPI. 1977. Vol.14. P.97-108.
5. *Молчан Г.М.* Оптимальные стратегии в прогнозе землетрясений // Современные методы интерпретации сейсмологических данных. (Вычисл. сейсмология. Вып. 24). М.: Наука, 1991. С.3-19.
6. *Ogata Y.* Statistical models for earthquake occurrence and residual analysis for point rocesses // J. Amer. Statist. Assoc. 1998. Vol.83. P.9-27.
7. *Ogata Y.* Detection of precursory quiescence before major earthquakes through a statistical model // J. Geophys. Res. 1992. Vol.97. P.19845-19871.
8. *Vere-Jones D.* Earthquake prediction: a statistician's view // J. Phys. of the Earth. 1978. N.26. P.129-146.
9. *Vere-Jones D., Zheng X.* Applications of stress release models to earthquakes from North China // PAGEOF. 1991. Vol.135. P.559-576.
10. *Zheng X., Vere-Jones D.* Further applications of stress release models to historical earthquake data // Tectonophysics. 1994. Vol.229. P.101-121.
11. *Ogata Y.* On Lewis' simulation method for point processes // IEEE Trans. Inform. Theory. IT-30. 1981. P.23-31.
12. *Wang A.L., Vere-Jones D., Zheng X.* Simulation and estimation procedures for stress-release models // Stochastic. Process. Appl. (M.J.Beckmann, M.N.Gopalan, R.Subramanian, eds.). Lecture notes in economics and mathematical systems, 370. N-Y.: Springer-Verlag. 1991. P.11-27.
13. *Papangelou F.* On the entropy rate of stationary point processes and its discrete approximation // Z. Wahrsch. Verw. Gebiete. 1978. Vol.44. P.191-211.
14. *Cziszar I.* On generalized entropy // Stud. Sci. Math. Hungar. 1969. N.4. P.401-419.
15. *Fritz J.* An approach to the entropy of point processes // Period. Math. Hungar. 1973. Vol.3. P.73-83.
16. *Zheng X.* Ergodic theorems for stress release processes // Stochastic. Process. Appl. 1991. Vol.37. P.239-258.