Pure appl. geophys. 162 (2005) 1229–1253
0033 – 4553/05/071229–25
DOI 10.1007/s00024-004-2667-2

© Birkhäuser Verlag, Basel, 2005

Pure and Applied Geophysics

The Entropy Score and its Uses in Earthquake Forecasting

DAVID HARTE[1] and DAVID VERE-JONES[2]

Abstract—Suppose a forecasting scheme associates a probability p^* with some observed outcome. The entropy score given to this forecast is then $-\log p^*$. This article provides a review of the background to this scoring method, its main properties, and its relationships to concepts such as likelihood, probability gain, and Molchan's ν-τ diagram. It is shown that, in terms of this score, an intrinsic characterization can be given for the predictability of a given statistical forecasting model. Uses of the score are illustrated by applications to the stress release and ETAS models, electrical signals, and M8.

Key words: Entropy score, probability forecasts, earthquake forecasts.

1. Introduction

There are many ways of scoring earthquake forecasts. Most are based on treating the forecast as a decision to assert the occurrence or non-occurrence of a certain type of event within a given period. The problem is then reduced to studying the proportions of successes, failures to predict, and false alarms. Relatively few methods make use of the full information in a probability forecast.

In this paper we outline an approach to scoring probability forecasts based on the *entropy score*. Suppose a trial can have several outcomes, indexed by the letter j, and that the forecast probability of outcome j is p_j^*. Then if outcome j occurs, we take the score to be $-\log p_j^*$.

The origin of the ideas here are hard to track down. The link between entropy, likelihood and one-step prediction errors motivated much of Akaike's original work on AIC, although the approach outlined here does not, to the extent that we are aware, figure explicitly in his papers. Information scores were used even earlier by HARTLEY (1928). Within seismology, the basic ideas were foreshadowed in early work by Yan Kagan (see, KAGAN and KNOPOFF, 1977), in which the average likelihood per event was used as a measure of the predictive power of the model. Although this appears as a

[1] Statistics Research Associates Ltd., P.O. Box 12-649, Wellington New Zealand.
E-mail: david@statsresearch.co.nz

[2] Victoria University of Wellington, P.O. Box 600, Wellington New Zealand.
E-mail: david.vere-jones@mcs.vuw.ac.nz

natural method of scoring probability forecasts, with important links to ideas such as probability gain and likelihood, until recently it does not seem to have been widely used either in weather forecasting or in earthquake forecasting. In recent years, however overlapping ideas have appeared in quite a range of contexts: see, for example, IMOTO (2000), KAGAN and JACKSON (2000), KLEEMAN (2002), and KLEEMAN *et al.* (2002). A discussion of the theoretical background was given by VERE-JONES (1998), who introduced the term *information gain* to stand for the expected value of the difference between the entropy scores of a model under test and a reference model. Theoretical aspects have been further reviewed in DALEY and VERE-JONES (2003, Ch. 7; 2004).

The purpose of the present paper is to provide an accessible review of the method and its applications to earthquake forecasting. Properties of the entropy score, including its links to other scoring methods such as probability gain and Molchan's ν-τ diagram (MOLCHAN, 1990, 1991), are developed in Sections 2 and 3, respectively. In particular, Section 4 describes an intrinsic bound on the performance of any forecasting scheme based on a fully specified point process model. Then Section 4 illustrates some uses of the entropy score, using as examples the renewal, stress release and ETAS models, forecasts based on electrical signals, and forecasts based on M8. The examples illustrate the use of the information gain as an intrinsic measure of model predictability, of the mean entropy score (mean log-likelihood) in tests of model consistency or goodness of fit, and of the average log probability gain in comparing the effectiveness of predictions for different classes of events.

2. *The Entropy Score and its Properties*

2.1 Probability Gain and Information Gain

The entropy score $-\log p_j^*$ can be interpreted as the *unpredictability* of outcome j. Its expected value (assuming the p_j^* are the true probabilities) is the *mean unpredictability* or *entropy* of the distribution $\{p_j^*\}$:

$$\mathscr{E} = -\sum p_j^* \log p_j^*. \tag{1}$$

The performance of the $\{p_j^*\}$ is usually related to that of a background scheme $\{\pi_j\}$, such as an independence model based on long-run probabilities. Then for each outcome we record the difference in the entropy scores, namely $\log(p_j^*/\pi_j)$ (note the change of sign). This quantity is the logarithm of the *probability gain*. Its expectation is

$$I^* = \sum p_j^* \log \frac{p_j^*}{\pi_j}, \tag{2}$$

and it measures the increase in predictability which comes from using the distribution $\{p_j^*\}$ (still assumed true) in place of $\{\pi_j\}$.

In information theory, this quantity is called the *relative entropy*, or the *Kullback-Leibler distance* $\mathscr{D}(p_j^*, \pi_j)$ between the distributions $\{p_j^*\}$ and $\{\pi_j\}$. Holder's inequality implies that it is always non-negative. On average, when making probability forecasts, we cannot do better than use the true probabilities.

If the true probabilities are not known, and neither of the proposed forecasting models is the true one, the expected probability gain can still be written as the difference between two Kullback-Leibler distances. Denoting the true probabilities by p_j^t, we have

$$\begin{aligned} E^t\left[\log\frac{p_j^*}{\pi_j}\right] &= \sum p_j^t \log\frac{p_j^*}{\pi_j} \\ &= \sum p_j^t \log\frac{p_j^t}{\pi_j} - \sum p_j^t \log\frac{p_j^t}{p_j^*} \\ &= \mathscr{D}(p_j^t, \pi_j) - \mathscr{D}(p_j^t, p_j^*). \end{aligned}$$

The expected gain here measures how much closer to the true model is the forecasting model than the background model. It is still a measure of the improvement in predictability which results from using $\{p_j^*\}$ in place of $\{\pi_j\}$. However, it could be negative if our choice of $\{p_j^*\}$ was a poor one.

2.2 Expected Information Gain

Usually, the prediction probabilities $\{p_j^*\}$ are computed from a conditional distribution, given a past history $\mathscr{H}_t$. Then the probabilities $\{p_j^*\}$ and the information gain I^* are themselves random, as they depend on the random past. In this case interest centres on the expected value

$$G = E(I^*).$$

When the whole process is stationary, this quantity is independent of t; we call it the *expected information gain per trial*, or just the *information gain* when the time unit is clear. It is an inherent characteristic of the process generating the trials, a numerical measure of its intrinsic predictability.

In the stationary case, the family of probability forecasts can be characterized by a stationary distribution of forecast probabilities, say $F(p_1^*, \ldots, p_J^*)$. This is the long-run distribution function of the set of forecast probabilities $\{p_1^*, \ldots, p_J^*\}$ over the hyperplane $\sum p_j^* = 1$. In this notation,

$$G = \int_{\sum p_j^* = 1} \left[\sum p_j^* \log\frac{p_j^*}{\pi_j}\right] F(dp_1^* \times \ldots \times dp_J^*). \tag{3}$$

In the simplest case, when there are only two possible outcomes — either an event occurs or it does not — the distribution F can be replaced by the distribution of the probability p^* that the event occurs. Assuming that the distribution of p^* has a density ϕ, we can then write

$$G = \int_0^1 \left[p^* \log \frac{p^*}{\pi} + (1 - p^*) \log \frac{1 - p^*}{1 - \pi} \right] \phi(p^*) dp^*. \tag{4}$$

2.3 Link to the Likelihood Ratio and Use in Testing Goodness of Fit

The mean log-likelihood ratio for N trials with outcomes $(j_1, \ldots, j_N)$ can be written in the form

$$\hat{G}_N \equiv \frac{1}{N} \log[L_1/L_0] = \frac{1}{N} \sum_{r=1}^{N} \log \frac{p^*_{j_r}}{\pi_{j_r}}. \tag{5}$$

It is just the average of the log-probability gains for the 1-step forecasts based on p_j^* and π_j. The formula remains valid even for dependent trials, provided the probabilities are written down sequentially, each one depending potentially on the past history up to and including the time of the preceding trial.

Under appropriate conditions the mean log-likelihood can be regarded as an estimate of the expected information gain per trial G (KAGAN and KNOPOFF, 1977). Indeed, under appropriate conditions, G is the expected value of the mean log-likelihood ratio, and we anticipate that as $N \to \infty$, $\hat{G}_N \to G$. For stationary processes, this is just an application of the ergodic theorem (equivalence of time and ensemble averages) to the logarithm of the probability gain.

The link between forecast performance, as measured by G, and the log-likelihood, provides an additional motivation for using likelihood methods. Loosely speaking, choosing the model that maximizes the likelihood ratio amounts to choosing the model with the best average 1-step prediction performance. In practice, with small data sets in particular, allowance has to be made for the numbers of unknown parameters that have been estimated from the data. This type of argument underlies the use of the *Akaike Information Criterion*, which we write here in the incremental form

$$\Delta\text{AIC} = 2 \log[L_1/L_0] - 2(k_1 - k_0), \tag{6}$$

where k_1, k_0 are the numbers of estimated parameters in the forecasting and background models, respectively. That model is chosen which maximizes the value of ΔAIC. In such cases,

$$\frac{\Delta\text{AIC}}{2N} = \hat{G}_N - \frac{k_1 - k_0}{N}$$

may be a better indicator of the improvement in predictability than $\hat{G}_N$ itself.

The mean log-likelihood (Eq. 5) can be used not only to estimate the expected information gain G itself, but also as a score which can form the basis of a crude goodness of fit test for the given model on the given data set. Such a test can be based on the departure of the mean loglikelihood (Eq. 5) from its expected value G under the model. Once the model is fully specified, G can either be calculated analytically (as in some of the examples discussed in Section 4) or obtained from repeated simulations. Provided the individual simulations are for a similar time period, the set of simulated score values then provides a population distribution to which the score for the real data can be compared. If the score from the real data lies near the extremes of the simulated scores, there is a prima facie case that the real data differ significantly from typical data sets produced by the model. Of course many other numerical characteristics of the data could be studied from a similar point of view, and we would not claim that this procedure has optimal qualities as a goodness-of-fit test; rather, it is a relatively convenient and accessible tool for this purpose. Further discussion is given in Section 4.2.

2.4 Point Process Models: Upper Bounds for the Information Gain

The analogue in continuous time to the expected information gain per trial is the *expected information gain per unit time*. The continuous time models that we shall discuss will be marked point process models defined by their *conditional intensities* $\lambda(t, M)$, where t is time, and M is a mark (magnitude and location, for example). The conditional intensity is an instantaneous rate, as in a Poisson process, but modified by a family of *histories* $\mathscr{H}_t$ recording information on past events and other relevant precursory information. In terms of the random counting measure N, recording the numbers of events in given time-mark sets, we can write

$$\lambda(t, M)dt\; dM = E[N(dt \times dM)]. \tag{7}$$

In the simplest case of a stationary process in time only, where we ignore the marks, the expected information gain per unit time can be expressed in terms of the conditional intensity at an arbitrary time $t = 0$, as

$$G = E\left[\lambda(0) \log \frac{\lambda(0)}{\mu} - 1\right], \tag{8}$$

where $\mu dt = E[N(dt)]$. If marks as well as time-points are observed, it is convenient to rewrite the marked conditional intensity in the form

$$\lambda(t, M) = \lambda_g(t) f(M|t),$$

where $\lambda_g(t)$ is the conditional intensity of the *ground process* (locations of time points irrespective of their associated marks), and $f(M|t)$ is the conditional probability density function for the distribution of marks, conditional on the past history and on the occurrence of an event at t. In this case the expected information gain per unit time can be represented as the sum of two terms, the first giving the gain from the predictions of the time points, and the second the gain from the predictions of the marks:

$$G = E\left[\lambda_g(0)\log\frac{\lambda_g(0)}{\mu_g} - 1\right] + E\left[\lambda_g(0)\int f(M|t)\log\frac{f(M|t)}{\overline{f}(M)}dM\right], \tag{9}$$

where $\overline{f}(M)$ is a reference density for the marks.

When comparing information gains from different data sets, with different average rates of occurrence, it is often convenient to rescale the information gains by taking the unit of time to be the mean time interval between events, or equivalently by dividing the above expressions for G by $E[\lambda_g(0)]$. The rescaled values can then be interpreted as the expected information gain per event. Further information on point process models and information gains can be found in DALEY and VERE-JONES (2003), in particular Chapters 6 and 7, and DALEY and VERE-JONES (2004).

Since forecasting schemes are likely to be restricted, by practical necessity, to forecasts made at regularly spaced or at least separated time points, and therefore come under the heading of discrete-time models, the value of pursuing continuous-time point-process models may be questioned. Several answers can be given to this question. The underlying process itself takes place in continuous time, and is hence more easily understood in terms of continuous models. In addition, as we shall see below, the information gain for the underlying continuous-time model provides an upper bound for the information gain of the corresponding discrete processes. Finally, the continuous time models are free of any specific assumptions concerning the forecasting intervals which are built into a discrete time model, but rather provide a common reference point to which the efficiency of various discrete forecasting schemes can be compared.

To compare the efficiency of forecasting schemes for the same data, but with different time intervals between forecasts, it is necessary to first convert the information gains per trial into information gains per unit time, by dividing by the mean interval length between forecasts. Suppose that the different schemes are based on successively refined partitions of the observation period, so that the length of the maximum interval between forecasts approaches zero. Assume also that the underlying process is stationary, with mean rate of occurrence of events μ, and that the background model is taken as a constant rate Poisson model with the same mean rate (intensity) μ. Then the following results hold for simple (unmarked) processes.

Upper Bound Theorem:

Under the above conditions:

(i) *the finer the partition, the larger the expected information gain per unit time;*

(ii) *an upper bound on the expected information gain per unit time for all discrete schemes is given by the point process expected information gain per unit time, namely*

$$G = E\left[\lambda(0)\log\frac{\lambda(0)}{\mu} - 1\right],$$

where $\mu dt = E[N(dt)]$; *and*

(iii) *the results remain true even for partitions of random length, provided the lengths can be determined from the histories* $\mathcal{H}_t$.

Similar results also apply for marked processes (cf., VERE-JONES, 1998). Thus, no matter what forecasting intervals are used, the expected information gain from any programme of regular forecasts is bounded above by the expected information gain for the continuous time model.

Calculations of the information gain for some standard models, including the renewal, ETAS and stress-release models, are outlined in Sections 4.1 and 4.2.

2.5 Complete and Partial Models

In the application of these ideas to earthquake forecasting, an important distinction has to be made between examples such as those discussed in the previous subsection, where all information needed for producing the forecasts stems from the history of the process itself, and models where some predictive information is taken from external variables which lie outside the scope of the model. We shall refer to the former as *complete models*, and the latter as *partial models*. In particular, partial models arise whenever the forecasts are based on *precursors* which lie outside the given model. To complete a partial model, such as a regression model onto some external precursory variables, the model must be extended to include the external variables as well as the forecast variables. If desired, the forecasts could then be extended to joint forecasts for both precursors and main events, thus opening the way to more effective simulation and forecasting schemes.

Even in the case of a partial model, when the true G is inaccessible (although it can be estimated from observations), the information gain and entropy score can provide useful tools for studying and comparing the performance of different forecasting schemes. We shall illustrate this in Section 4.3 by reference to electric signals and M8 studies.

3. Overview of Alternative Scoring Methods

There are, of course, many possible ways in which probability and other forecasts can be scored. The most commonly used scoring techniques are based on 0–1 situations in a discrete time context, so that a future event is either predicted or not predicted, and the prediction procedure is evaluated in terms of successes, false alarms, and failures to predict. Probability forecasts do not fall directly into this category, for no unequivocal prediction is made as to whether some particular event will or will not occur. However, they can be be mapped into such a 0–1 procedure by making the decision to issue a prediction dependent on whether or not the forecast probability exceeds a given threshold. Our main aim in the present section is to relate the entropy score and associated information gain to various scores arising from the analysis of such 0–1 schemes.

It may be worth to emphasize that making a prediction is in itself a decision, namely the decision to issue the prediction, to whatever audience is in view, and with whatever costs or benefits that decision may bring. Typically, probability forecasts are used as input to decision-making procedures, underlying calculations of expected costs or benefits. So it is here. Just as forecasts of long-term probabilities of earthquake occurrence are used to guide decisions on zoning schemes, site analyses, earthquake insurance etc., so, too, probability forecasts can be used to assess the expected costs and benefits of issuing shorter-term predictions, under various assumptions regarding the costs involved. MOLCHAN (1991), and MOLCHAN and KAGAN (1992), among others, have discussed earthquake prediction from this point of view.

From a decision point of view, a score is just a surrogate for a proper cost-benefit analysis. Tuning a forecasting procedure to optimize a particular type of score is no guarantee that the optimizing procedure will perform well in a particular decision-making context where costs and benefits have been quantified. Instead, use of a score such as the the entropy score offers a quick guide to performance, with a range of uses that would lie outside the scope of any particular cost-benefit analysis.

3.1 Scoring a 2×2 *Table*

We consider the simplest and most common case where the forecast covers just the occurrence or non-occurrence of an event, i.e., there are just two outcomes which we shall label by $j = 1$ and $j = 0$, respectively. Let the corresponding forecast probabilities be denoted by p^* and $1 - p^*$. Suppose also that the decision rule is to take action (i.e., predict the occurrence of the event) if and only if p^* lies above a certain threshold p_0.

This decision framework sets up the basis for classifying the results in a 2×2 table, with entries, reading clockwise from the top left corner:

$a =$ number of successful forecasts of occurrence,
$b =$ number of false alarms,
$c =$ number of successful forecasts of non-occurrence, and
$d =$ number of failures to predict.

The most common score derived from such a table is the *Hanssen-Kuiper skill score*, or *R-score*, namely

$$R = \frac{ac - bd}{(a+b)(c+d)} = \frac{a}{a+b} - \frac{d}{c+d} = \frac{c}{c+d} - \frac{b}{a+b}. \qquad (10)$$

The middle form, for example, can be interpreted as the proportion of successful forecasts less the proportion of non-forecasts resulting in failures to predict.

The choice of threshold p_0 is somewhat arbitrary. A common choice, which optimizes the value of R under certain conditions, is to take $p_0 = \pi$, the long-run probability of occurrence. In principle, the choice should be made on a cost-benefit analysis of the different possible outcomes. A different balance of the costs and benefits will mean a different choice of threshold, in general. The choice $p_0 = \pi$, for example, leads to considerable over forecasting when π is small, and in such a situation should be adopted only when the cost of a failure to predict is substantially higher than the cost of a false alarm.

Insight into the behaviour of the table, and of the R-score, can be obtained by replacing the entries in the 2×2 table by their expected values (long-run proportions) under the assumption that the whole process is stationary and ergodic. Denote the actual outcome (occurrence or non-occurrence of the event) by X, and by Y the issuance or otherwise of the prediction, so that the event occurs if and only if $X = 1$, and is predicted to occur if and only if $Y = 1$. Both X and Y can be treated as random variables within some more extensive sequence of such variables. Generally, Y will be *predictable*, that is, determined by information available up to but not including the outcome of the trial on test. When the forecast is based on the forecast probability p^* exceeding the threshold p_0, then Y will have the form $Y = I_{p^* > p_0}$.

The probabilities (long-term proportions) corresponding to the numbers a to d in the 2×2 table are then given by

$$\rho_a = E[XY],\ \rho_b = E[(1-X)Y],\ \rho_c = E[X(1-Y)],\ \rho_d = E[(1-X)(1-Y)].$$

The long-run proportion of the ‘time on trial’, namely the proportion of time periods for which an event is forecast, is then the probability $\tau = E[Y]$, while the long-run probability of occurrence of an event is $\pi = E[X]$. Another important characteristic is the proportion of occurrences in which there is a failure to predict, which can be represented as $\nu = E[X(1-Y)]/E[X]$. Using these notations,

$$R = \frac{\pi}{\tau(1-\tau)}[(1-\nu) - \tau], \tag{11}$$

showing that the essential contribution to the R-score derives from the difference between the proportion of successfully predicted events and the proportion of time on trial.

In the case of random forecasts, X and Y are independent so $E(XY)/E(X) = E(Y)$, that is, $1 - \nu = \tau$. Thus the long-run R-score should be zero. In the case of perfect prediction, $Y = X$, so $E[XY] = E[X^2] = E[X]$, and $\nu = 0,\ \pi = \tau$, so $R = 1$. Similarly in the case of perfect anti-prediction, $R = -1$.

3.2 Molchan's ν-τ Diagram

MOLCHAN (1990) suggested the following extension of the procedures derived from the 2 × 2 table. For every possible threshold p_0, plot the proportion of failures to predict

$$\nu = \frac{d}{a+d}$$

against the fraction of time on forecast

$$\tau = \frac{a+b}{a+b+c+d}.$$

The resulting ν-τ diagram provides a comprehensive summary of the observed performance of a probability forecasting scheme for 0–1 outcomes. In the case of purely random forecasts, as we have seen, $\nu = 1 - \tau$, and the diagram consists of the diagonal joining the points $(0, 1)$ and $(1, 0)$.

To describe the behaviour of the diagram in more general cases, let us denote the forecast probability that the event occurs by $\tilde{p}$, reserving p^* for the corresponding true probability. We suppose that in both cases the probabilities are functions of the history up to but not including the outcome of the current trial, and that the whole procedure is embedded within a stationary, ergodic (but not necessarily independent) sequence of trials. Thus for example $\tilde{p}$ might be obtained from calculations based on a wrong model for the process, as will be virtually inevitable when the true model is unknown.

In the stationary regime the pair $(\tilde{p}, p^*)$ will have some long-run joint distribution, averaged over all possible histories. For ease of notation, assume that this distribution has a density $f(\tilde{p}, p^*)$ (this will normally be the case unless $\tilde{p}$ and p^* are functionally related, as in the case $\tilde{p} \equiv p^*$, when the joint distribution becomes degenerate). Then we can represent the expected values in the 2 × 2 table in terms of this joint density. For example

$$\rho_a = E[XI_{\tilde{p}>p_0}] = \int_{\tilde{p}=p_0}^{1} \int_{p^*=0}^{1} p^* f(\tilde{p}, p^*)\ d\tilde{p}\ dp^*, \tag{12}$$

with similar representations for the other entries. For the expected proportion of time on trial we can write

$$\tau = E(Y) = \Pr[\tilde{p} > p_0] = 1 - \tilde{H}(p_0), \tag{13}$$

where $\tilde{H}()$ is the stationary distribution of $\tilde{p}$. Similarly, for the long-run proportion of failures to predict, we can write

$$\nu = \rho_a/\pi = J(p_0), \tag{14}$$

where the distribution $J()$ is defined by

$$J(x) = \frac{\int_{\tilde{p}=0}^{x} \int_{p^*=0}^{1} p^* f(\tilde{p}, p^*)\ d\tilde{p}\ dp^*}{\int_{\tilde{p}=0}^{1} \int_{p^*=0}^{1} p^* f(\tilde{p}, p^*)\ d\tilde{p}\ dp^*} = \Pr(\tilde{p} \le x | X = 1). \tag{15}$$

We see that the v-τ plot is essentially a *Q-Q* (quantile-quantile) plot of these two distribution functions against each other, reducing to the diagonal when the two distributions are equal, as occurs when X, Y are independent. In general, the slope of the plot at the point corresponding to the threshold p_0 is given by

$$\frac{dv}{d\tau} = \frac{dv}{dp_0} \Big/ \frac{d\tau}{dp_0} = -\frac{\int_{p^*=0}^{1} p^* f(p_0, p^*)\ dp^*}{\int_0^1 f(p_0, p^*)\ dp^*} = -\frac{1}{\pi} \Pr(X = 1 | \tilde{p} = p_0). \tag{16}$$

The slope is everywhere negative, and increases in magnitude with p_0 (and hence decreases with $\tau = 1 - \tilde{H}(p_0)$) provided $E[X = 1 | \tilde{p} = p_0]$ increases with p_0. In other words, the curve will be convex at every point where an increase in the predicted probability that the event will occur is associated with an increase in the actual probability of its occurrence.

If the predicted probabilities are equal to the true probabilities, i.e. $\tilde{p} = p^*$, then writing $\phi(p)$ for the stationary density of p^*, we have

$$\frac{dv}{d\tau} = \frac{dv}{dp_0} \Big/ \frac{d\tau}{dp_0} = -\frac{p_0 \phi(p_0)}{\pi \phi(p_0)} = -\frac{p_0}{\pi},$$

which is certainly monotonic in p_0. It is equal to unity (i.e., the curve has tangent parallel to the diagonal) when $p_0 = \pi$, with a slope steeper than unity for larger values of p_0 (smaller values of τ), and a slope less than unity for smaller values of p_0 (larger values of τ). Thus the v-τ curve based on the true probabilities lies always below the line $v = 1 - \tau$, and is always convex downwards. The more predictable the process, the more rapid the initial change in slope, and the closer the graph approaches the extreme case of perfect prediction, when it reduces to the left vertical and bottom horizontal sides of the unit square.

Moreover, the curve based on use of the true probabilities provides a lower envelope for the v-τ curves generated by other prediction schemes. To see this, consider a fixed proportion of time on trial, say τ, corresponding to thresholds $\tilde{p}_0$ and p_0^* for the prediction schemes based on some wrong model and the true model, respectively. If we compare

$$\tilde{a} = E[XI_{\tilde{p} > \tilde{p}_0}] = E[p^* I_{\tilde{p} > \tilde{p}_0}]$$

with

$$a^* = E[XI_{p^* > p_0^*}] = E[p^* I_{p^* > p_0^*}],$$

the latter concentrates the available probability mass into high values of p^*, increasing $1 - v$ as far as possible, whereas the former distributes the available probability mass into a set which will include a greater proportion of smaller values of p^*, giving a smaller value of $1 - v$. Thus the curve for a wrong model always lies above the curve for the true model. Similarly it lies below the curve giving the worst possible performance, namely that for which $\tilde{p} = 1 - p^*$, which is just the reflection of the optimal curve in the diagonal.

Finally, we consider the connection between the entropy score and the v-τ diagram. The information gain (expected entropy score) can be written in the notation of the preceding discussion as

$$G = \int \int \left[p^* \log \frac{\tilde{p}}{\pi} + (1 - p^*) \log \frac{1 - \tilde{p}}{1 - \pi} \right] f(\tilde{p}, p^*) \; d\tilde{p} \; dp^*. \tag{17}$$

Unfortunately, there seems to be no simple way of relating the quantities appearing in this integral to features of the v-τ curve; the curve depends on the bivariate density $f(\tilde{p}, p^*)$, but in general this is not uniquely defined by the Molchan curve, so that we cannot go back from the curve to the quantities defining G.

An exception is the curve generated by forecasts based on the true model. In this case we can write

$$G = \int \left[p \log \frac{p}{\pi} + (1 - p) \log \frac{1 - p}{1 - \pi} \right] \phi(p) \; dp,$$

and make use of the relations

$$\frac{dv}{d\tau} = -\frac{p}{\pi}; \; d\tau = \phi(p) dp$$

to obtain G in the form

$$G = \int \left[-\pi \frac{dv}{d\tau} \log \left(-\frac{dv}{d\tau} \right) + \left(1 + \pi \frac{dv}{d\tau} \right) \log \frac{1 + \pi \frac{dv}{d\tau}}{1 - \pi} \right] d\tau, \tag{18}$$

which can be determined from a knowledge of π and the form of the Molchan diagram. Even here, however, there seems to be no obvious geometrical interpretation of G in terms of the Molchan diagram. It is one numerical characteristic of the curve, but there are many others, such as the maximum distance between the curve and the diagonal, which could also be used to characterize the predictability of the model.

It is worth noting that, like G, the Molchan diagram admits an extension to continuous time models. As the forecasting intervals decrease, the probability of success becomes of the form $\lambda^*(t)dt$ (for the true model) or $\tilde{\lambda}(t)dt$ for the model under test. From such considerations, it is not difficult to show that the Molchan diagram converges towards the Q-Q plot for two distributions of the test conditional intensity: the first, corresponding to $\tilde{H}$ in the discrete time case, is just the stationary distribution of $\tilde{\lambda}(t)$ at an arbitrary time t; the second, corresponding to J in the discrete case, is the stationary distribution of $\tilde{\lambda}(t)$ at a time t when an event occurs. These can also be treated as the stationary distributions of $\tilde{\lambda}(t)$ and of $\tilde{\lambda}(t)\lambda^*(t)/m$, where m is the mean rate of occurrence of points.

Any score can be transformed into a "skill score" by taking the ratio

$$\Sigma = \frac{\text{actual score} - \text{baseline score}}{\text{perfect score} - \text{baseline score}}.$$

In the case of the entropy score, the perfect score has value 0 (= log 1) and so the "entropy skill score" is given by

$$\Sigma_H = 1 - \frac{\sum \log p^*_{X_n}}{\sum \log \pi^*_{X_n}}$$

where the $\{X_n\}$ are the observed outcomes.

4. *Uses of the Entropy Score*

In this section we give examples to show how the properties summarized above can be brought to bear in applications.

4.1 Characterizing Model Predictability: Renewal and ETAS Models

We have emphasized at several points that the expected information gain G is a number characterizing the inherent predictability of the model. To gain feeling for the degree of predictability inherent in a given value of G it may be helpful to see how G varies over a range of members of a well-known family. Following DALEY and VERE-JONES (2004), we show how G varies for a family of renewal processes in which the interevent-distribution is of gamma $\Gamma(\alpha, \lambda)$ form, with constant mean length 1, but varying shape parameter. In seismology, the renewal model is often used as the model of choice for the occurrence times of characteristic earthquakes along a given fault. The gamma distribution can be used for this purpose, although longer-tailed distributions such as the log-normal, are commonly preferred, as they more readily accommodate occasional very large values among the interval lengths; see, for example, OGATA (1999), and RHOADES and VAN DISSEN (2003).

The information gain for a stationary renewal process is related to the entropy of the underlying interevent distribution. For a renewal process with interevent distribution $F(\cdot)$, tail (or survivor function) $\overline{F}(\cdot)$, density $f(\cdot)$ and mean $1/\mu = \int_0^\infty \overline{F}(y)\mathrm{d}y$, the conditional intensity is $\lambda^*(t) = f(B_t)/\overline{F}(B_t)$, where B_t denotes a (stationary) backward recurrence time random variable, here the time since the last occurring event; B_t has density $\mu\overline{F}(\cdot)$. After some manipulations, we find

$$G = \mu\left[1 - \log\mu + \int_0^\infty f(y)\log f(y)\mathrm{d}y\right], \tag{19}$$

and the algebraic evaluation of G in any particular case reduces to the computation of this last integral. Equivalently, the information gain per event is obtained by omitting the multiplier μ.

Consider the lifetime distribution has the gamma density function

$$f_\kappa(x; v) = \frac{v^\kappa x^{\kappa-1} e^{-vx}}{\Gamma(\kappa)} \qquad (0 < v < \infty, 0 < \kappa < \infty,\ 0 < x < \infty). \tag{20}$$

In this case the mean interval length $1/\mu = v/\kappa$, while for the entropy function we find

$$\int_0^\infty f_\kappa(x, v)\log f_\kappa(x, v)\ \mathrm{d}x = -\kappa - \log\Gamma(\kappa) + \log v + (\kappa - 1)\psi(\kappa), \tag{21}$$

where $\psi(z) = (d/du)\log\Gamma(u) = \Gamma'(u)/\Gamma(u)$.

Setting the mean equal to 1, allowing the shape parameter κ to vary, and numerically evaluating the above expressions leads to Table 1 and Figure 1; algebraic details are given in Daley and Vere-Jones (2004). The corresponding Molchan diagram is plotted in Figure 2. It will be noted that the gain increases not only as the process becomes deterministic, but also as the process becomes highly clustered ($\kappa < 1$). The reason in the latter case is that there is an extremely high risk for values of the interval length close to zero, with a much lower risk for longer intervals. The probability gain in this case can be immense for small intervals, and large also over long intervals, resulting in a significant average gain. Thus the gain in predictability is associated with the extreme range of interval lengths, and in particular with the very high short-term risk.

A similar feature is seen with the ETAS model, which captures well the increase in risk due to events (aftershocks) immediately following an initial event, but is less effective as a basis for longer-term forecasts at regular intervals. In this model the

Table 1

Information gains distributed intervals (mean = 1)

κ	Gain
0.1	5.72608
0.2	3.28402
0.5	0.21624
1.0	0.00000
5.0	0.45585
10.0	0.76653
50.0	1.54377

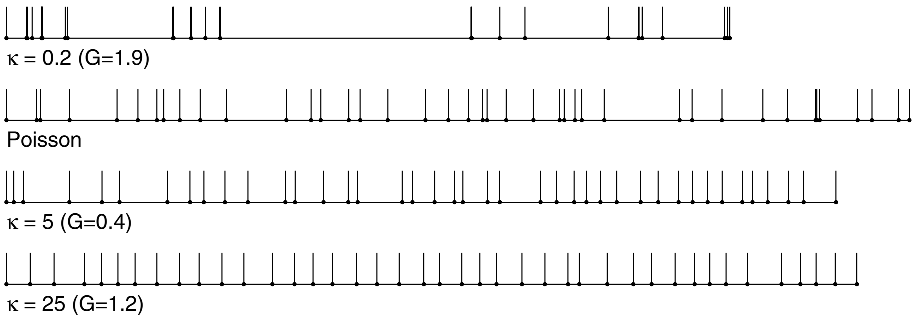

Figure 1
Parts of realizations of four different gamma-distributed renewal processes.

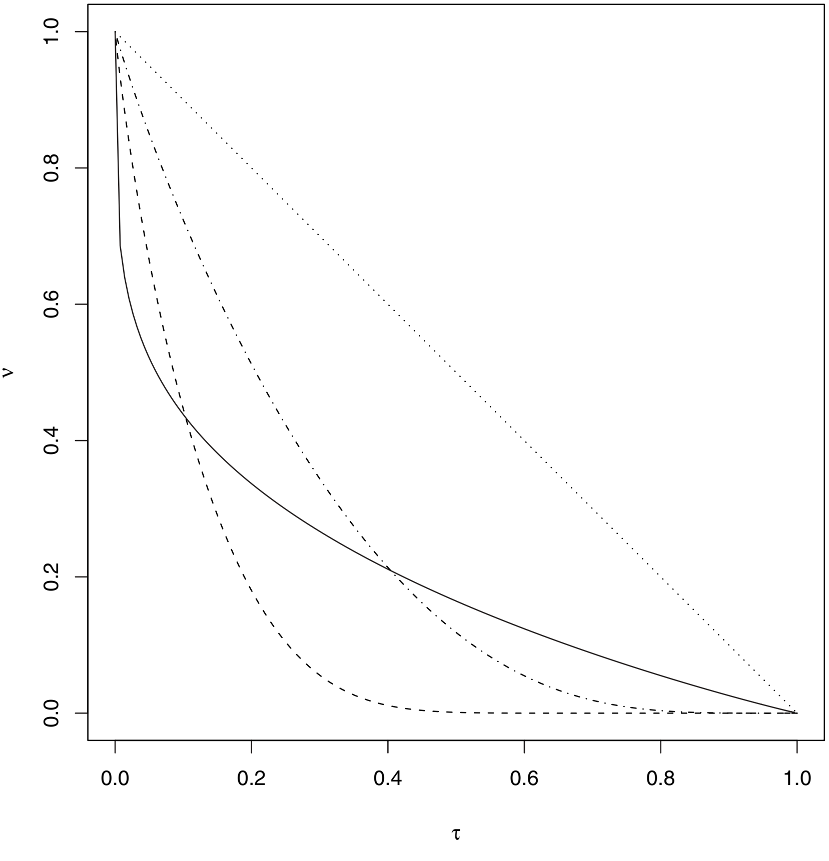

Figure 2
Molchan diagrams for the renewal models of Figure 1. The solid line is for $\kappa = 0.2$, the dotted line is for $\kappa = 1$, the dash-dot line is for $\kappa = 5$, and the dashed line is for $\kappa = 25$.

conditional intensity at any time t is represented as the sum of a background intensity μ and contributions, of Omori-law type, from each of the preceding events. The total intensity at time t is given by the sum

$$\lambda(t) = \mu + K \sum_{i:t_i<t} e^{a(M_i - M_0)}(t + c - t_i)^{-p}.$$

Here M_0 is a threshold magnitude (lower bound to the events considered), K, a, c, p are parameters, and (t_i, M_i) are the times and magnitudes of the earlier events before time t. 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 magnitudes are assumed to be identically distributed, usually according to the standard (exponential) form of the Gutenberg-Richter Law. Further background and applications are given, for example, in OGATA (1988, 1992).

Although the information gain characterizing the model is high, its performance as a basis for probability forecasts over longer intervals can be surprisingly low. This was illustrated in VERE-JONES (1998), from which Tables 2 and 3 are extracted. In this study a base catalogue of some 1000 events was simulated 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.

Table 2 illustrates how the information gain can be determined empirically from the scores from successive finite-length forecasting intervals. Forecasts were made

Table 2

Scores and information gains for the ETAS model (after VERE-JONES, 1998). The three sets of results in the table represent scores for three forecast intervals. N is the number of events that occurred in the interval, X is the binary score, p is the forecast probability of success, B and B are the entropy scores for the ETAS and Poisson models, and G is the log-probability gain*

Mag. Class	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	−1.12	0.04
$3 \leq M < 4$	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	−1.25	−0.52

Table 3

Average scores and information gains for the ETAS model (after VERE-JONES, *1998).* N_S and N_F *denote numbers of successes and failures,* $\bar{G}_S$ *and* $\bar{G}_F$ *denote the corresponding average scores;* N, $\bar{G}$ *and* $\bar{G}_E$ *denote the total number of trials, the average score per unit time, and the average score per event, respectively*

Mag. Class	N_S	$\bar{G}_S$	N_F	$\bar{G}_F$	N	$\bar{G}$	$\bar{G}_E$
$6 \le M < 7$	410	−0.83	1590	0.24	2000	0.02	0.10
$7 \le M < 8$	148	0.10	1852	0.043	2000	0.01	0.10
$8 \le M < 9$	15	−0.26	1988	0.001	2000	−0.00	−0.00

every two days, and the results represented in binary form (events of a given class did or did not occur in any given interval). The log probability gain was calculated for each outcome. Then the average of the log probability gains was computed for the entire sequence. In this example the outcomes were marked (either no events, or events divided into magnitude classes), and the gains computed and averaged for each magnitude class separately, then summed to produce an overall gain. Resulting scores and gains for three typical forecasting periods are shown in Table 2. Averages for the entire sequence are shown in Table 3.

It will be seen that the expected information gain for 2-day intervals is only positive. The explanation is that, with this model, considerable information is lost by not tailoring the forecasting intervals to the occurrence of events. In the second row of Table 2, for example, a cluster of events occurred entirely within the forecasting interval, but no advantage was gained from the model's ability to forecast aftershocks. In such intervals the model may even perform worse than the Poisson model, for the background rate μ is necessarily lower than the overall rate used by the Poisson model, so that if an event occurs after some gap, it is better predicted by the Poisson than the ETAS model. As shown recently by HELMSTETTER and SORNETTE (2003), however, the ETAS model has significant predictive power if small events are viewed as potential foreshocks and used to predict larger events. However in such a case the forecasting interval has to be short and initiated by an event.

4.2 Comparison of Mean Log-likelihood and Information Gain: Stress Release Model

The stress-release model (e.g., ZHENG and VERE-JONES, 1991) is another model for which detailed investigations of the information gain have been carried out. Here the conditional intensity has the form

$$\lambda(t) = \psi(X(t)) = \exp[\beta(X(0) + \rho t - S(t) - x_0)], \tag{22}$$

where $\psi(x) = \exp[\beta(x - x_0)]$ determines the hazard rate when the local stress level, which is treated as a scalar, reaches the value x. The evolution of the local stress level is assumed to be governed by an equation of the form

$$X(t) = X(0) + \rho t - S(t),$$

where $X(0)$ is the initial value, ρ is a constant loading rate from external tectonic forces, and $S(t)$ is the accumulated stress release from earthquakes within the region over the period $(0, t)$. That is, $S(t) = \sum_{t_i<t} S_i$, where t_i and S_i are the origin time and the stress release associated with the ith earthquake. The S_i themselves are independently determined from the Gutenberg-Richter relation or some alternative frequency-magnitude law. Unlike the ETAS model, this model does not attempt to forecast aftershocks, but models rather the build-up of risk over longer time periods. There is an increase in activity as the time since the last major event increases, much as in the accelerated moment release model, followed by relative quiescence after the next major event. In applications to typical historical data, such as Zheng and Vere-Jones (1994), the information gains are of the order of 0.1–0.2 per event, consequently in these contexts the predictive power is modest.

A discussion of information gains for this model is contained in Bebbington (2004). In particular, Bebbington derives an analytical expression for G as a function of the model parameters and the magnitude-frequency distribution, and examines the influence of the form of the frequency magnitude distribution on G when the other parameters are held fixed. Some examples of these calculations are shown in Tables 4 and 5. Two interpretations of the stress drop are used in Table 4: in the first row it is taken to be proportional to the Benioff strain (using $X \propto 10^{0.75M}$), indicated by $\eta = 0.75$ in the first column, and in the second row it is taken to be proportional to the seismic moment release ($X \propto 10^{1.5M}$), corresponding to $\eta = 1.5$. The likelihood estimates of the parameters x_0, β, and ρ in the conditional intensity (Eq. 22) are summarized in Table 4. They do not depend on the form of the frequency-magnitude distribution, because in estimating the parameters the observed magnitudes are treated as given quantities.

The corresponding information gains are shown in Table 5. Here the form of the magnitude distribution does make a difference. Even within the restricted options shown in the table, the form of the frequency-magnitude distribution can affect the information gain G by up to an order of magnitude. At first sight this may seem surprising, because in the ETAS model the magnitude of each event is presumed to be selected independently of past events according to a fixed distribution. The dependence arises because the magnitudes of past events do affect current values of the conditional intensity, so that typical patterns of the conditional intensity, on

Table 4

Fitted SRM parameters from North China data following the notation of Eq. (22)

η	$\widehat{\beta}$	$\widehat{\rho}$	$\widehat{x_0}$	$\ln L$
0.75	0.010	1.176	246.2	−195.87
1.5	0.000134	47.3	18193	−196.68

Table 5

Calculated entropy gains for North China. "Empirical" means resampling from the observed set of magnitudes; "Degenerate" means fixed magnitude value; "Truncated G-R" means the usual Gutenberg-Richter distribution (exponential magnitudes or power-law moments) with an upper truncation point; and "Tapered Pareto" means a distribution for the moments which is basically of power-law form, but changes to an exponential form for very high values of the moment

Stress Distribution, $J(\mathrm{d}x)$	η	E[S]	$\bar{\lambda}$	G	$G/\bar{\lambda}$
Empirical	0.75	8.81	0.1335	0.0120	0.0902
Empirical	1.5	330.5	0.1434	0.0213	0.1486
Degenerate, $M = 7.26$	0.75	8.81	0.1335	0.0029	0.0220
Degenerate, $M = 7.68$	1.5	330.5	0.1434	0.0016	0.0111
Truncated G-R, $\theta = 1.22$, $m_{\max} = 9.0$	0.75	8.83	0.1331	0.0160	0.1202
Truncated G-R, $\theta = 1.36$, $m_{\max} = 9.0$	1.5	352.6	0.1344	0.0453	0.3368
Tapered Pareto, $\alpha = 0.72$, $\gamma = 8.6$	0.75	8.83	0.1331	0.0187	0.1407
Tapered Pareto, $\alpha = 0.36$, $\gamma = 8.5$	1.5	352.2	0.1343	0.0285	0.2126

which G depends, can be quite different for different magnitude distributions, even when the other parameters in the model are held fixed. Even stronger effects of this kind are implicit in the study by LU and VERE-JONES (2001) of Ben-Zion's numerical models for a major fault system (BEN-ZION, 1996). Here both the original data and the fitted stress-release models can exhibit anything from Poisson/Gutenberg-Richter behaviour, to highly periodic, characteristic earthquake behaviour.

The values of G shown in Table 5 can be compared with the mean likelihood ratios $\hat{G} = T^{-1}\log(L/L_0)$ obtained directly from the data. For the case $\eta = 0.75$, this yields the estimate $\hat{G} = 0.00775$. The likelihood estimate is based on the observed stress-drops estimated, albeit crudely, from the historical data. The estimated $\hat{G}$ is therefore probably most meaningfully related to the calculated values of G based on the empirical distribution of magnitudes.

We see that the estimated values of $\hat{G}$ are of the right order of magnitude, although somewhat smaller than that in Table 5 for the empirical distribution. The question then is whether this difference should be considered as evidence of a significant departure of the observed data from the sort of data to be expected from the model. This is exactly the question, raised in Section 2.3, of the possibility of using the sample mean likelihood as the basis for a test of goodness of fit.

A test of this kind can be developed, and again based on simulations, provided the procedures used in obtaining the observed value of $\hat{G}$ are repeated in obtaining corresponding values from the simulated data. In the present example, this means first simulating many replicas of the earthquake data, for the same length of time as the N. China data, using the fitted stress release model for the N. China data, together with the empirical frequency-magnitude distribution, as the basis for the simulations. Then the stress release and Poisson models are refitted to each set of simulated data, and the resulting log-likelihood ratios recorded and collected

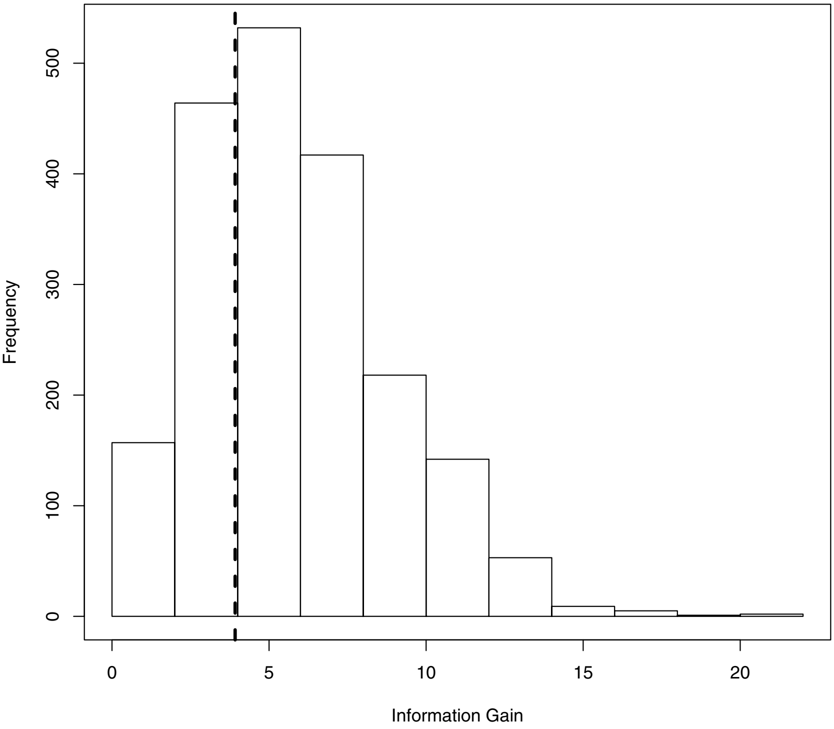

Figure 3
Histogram of simulated log-likelihood ratios, with the value from the real data shown by the dashed vertical line.

together into a histogram. The resulting distribution of the simulated log-likelihood ratios then forms a basis for comparison with the value from the real data. In the present case, such a procedure leads to the results illustrated in Figure 3, where the location of the value from the real data is shown by a vertical line, and the histogram refers to values from the simulated data. No suggestion of major discrepancies emerges here. Relative to the general spread of values, the sample value is close both to the center of the distribution of simulated values, and to the exact values calculated for the model.

This procedure is important also because it links into the problem of testing the performance of a given model on a given data set, as envisaged with the RELM model testing framework. Indeed, the simulation procedure just outlined appears similar to those recommended for the RELM procedures (see SCHORLEMMER *et al.*, 2004).

In this context there are two aspects to consider. First, the log-likelihood ratios against a standard null model can be computed for each of the competing models, and used as a rough yardstick of performance. If the test data is different to the data from which the model parameters have been estimated, bias due to the differing numbers of parameters between models should not arise, so that the models can be ranked according to the values of their gains against the null model. The null model here does not play a major role, but can be any convenient reference model, such as the Poisson model with uniform probabilities and the standard Gutenberg-Richter distribution of magnitudes. It will be common to all the competing models, and will drop out of any direct comparison between models. In effect, its role is to bring the values of the scores into a convenient numerical range.

The second aspect is that of checking the internal consistency of each model with the given data set, that is, of checking its goodness of fit. Here a comparison of the log-likelihood ratio from the real data, with values from simulations from the fitted model, as considered with the N. China data above, should provide such a check, although it should be noted that at this stage little is known about the power of such a test in these situations.

4.3 Some Uses with a Partial Model: Electric Signals and M8 Models

There are many ways that the entropy score, or log probability gain, can be used to help assess the forecasting performance of a model, even in situations where the model is only a partial model, and therefore the option of simulating from the model is not available. We briefly illustrate two such uses, from models using electric signals and M8 signals as input data.

In recent studies OGATA and ZHUANG (2001) and ZHUANG *et al.* (2002, 2005) have examined Chinese data on fluctuations in the electric ground potential as a possible precursor for earthquakes in the vicinity of Beijing. It is not our purpose here to discuss the validity of the models or the data on which they are based, but simply to illustrate how the entropy score can be used to compare the performance of forecasts on different classes of events.

In these studies the signals from four stations around Beijing were used as explanatory variables, external to the model, in a generalized regression model for the occurrence of events with magnitude $M \geq 4$. Signal values were computed on a daily basis, and used to predict the number of such events in the following day, this number being treated as a Poisson variable with its parameter conditional on the observed signal value.

The events were not further classified in the regression, but differed in their sizes and in their distances from the measuring stations. In such circumstances one might anticipate that larger or closer events would be predicted more effectively than smaller or more distant ones. One way of looking at such possibilities is to calculate the probability gains, or their logarithms, for different classes of events. The logarithm of the probability gain is just the difference in the entropy scores of the model being used and a reference model. Its average over a certain class of events then forms a relative measure of the effectiveness of the model in predicting events in that class. Such an average cannot be taken as measuring the overall performance of the model on events of that class, as an overall measure should also take into account the performance of the model during intervals when no target events occurred.

The most suitable average to use in such comparisons, because of its link to log-likelihoods, is the average of the difference in entropy scores, corresponding to the geometric mean of the probability gains. Thus we compute

$$\bar{H} = \frac{1}{M}\sum_{i=1}^{M} \log\left(\frac{p(t_i)}{\bar{p}}\right).$$

The geometric mean of the probability gains themselves is then

$$\tilde{p} = e^{\overline{H}}.$$

As an illustration, Table 6 shows the mean probability gains $\tilde{p}$ computed in this way for various classes of interest from a bivariate regression on signals from two of the recording stations near Beijing. The target events in this case referred to events with magnitude $M \geq 4$ falling within a 300 km radius circle around Beijing. The first row shows the overall performance and the mean information gain per day.

The latter may seem small, but since there are 5478 days in all only 84 of which contained an event, the mean information gain per event is 0.31, nearly double that achieved by the stress-release model on a very different time scale. The remaining rows indicate the mean probability gains for days on which events of a particular type occurred. Thus, the second row gives the average gain for the days containing no events. This average is slightly negative, indicating that on such days the model forecast is, on average, slightly higher than the forecast from the Poisson model. The next row gives the average over days containing an event in the range $4 \leq M \leq 4.4$ but no larger event, and so on. The improvement in the performance of the forecasting model for larger events is very apparent.

Table 6

Probability gains for the electric signals model (after ZHUANG *et al., 2002). H is the entropy score for the fitted model, H_0 is that for the Poisson model, M is the number of relevant event days, $\bar{H}$ is the log probability gain per event-day and $\bar{p}$ is the geometric mean of the probability gains. The scores arise from the prediction performance of a regression on the signals from Qingxian and Sanhe stations over a time period of 5477 days, from 1982 to 1998. The first line ("All") gives the overall performance, the second ("Non") the performance over days on which no event occurred within the test region, while the remaining lines summarize the performances over days on which an event occurred in the specified magnitude range*

Class	H	H_0	$H - H_0$	M	$\bar{H}$	$\bar{p}$
All	−395.915	−434.352	38.437	5478	0.0070	1.0070
Non	−81.242	−79.891	−1.352	5394	−0.0003	0.9997
4.0-4.4	−184.877	−198.330	13.453	47	0.2862	1.3313
4.5-4.9	−95.422	−109.714	14.292	26	0.5497	1.7327
5.0-5.4	−25.747	−33.758	8.011	8	1.0013	2.7218
≥5.5	−8.626	−12.660	4.034	3	1.3447	3.8369

Although the values of H and $\overline{p}$ for a given class cannot be taken as an overall assessment of the performance of the model in predicting events within the given class, a rough estimate of the overall performance can be made as follows. Let us consider the class of all events with magnitudes above 5, corresponding to the last two rows of Table 6. A rough model for such events might be obtained by taking the forecast probabilities for "All" events in the table and multiplying these by the probability that the forecast event was greater than 5, as estimated from the Gutenberg-Richter law. Since the same factor would also be used for modifying the corresponding Poisson model, it drops out of the initial term for the log-probability gain (involving $\sum \log[p_i/\bar{p}]$) corresponding to the days on which an event in the given class was observed. Thus the contribution to the entropy score from those days remains as given in the last two lines of the table, viz $8.011 + 4.034 = 12.045$. The contribution to the scores from the remaining days is reduced by the fact that $1 - p_i$ for the model probabilities, and $1 - \overline{p}$ for the Poisson probabilities, are both close to 1. Their contribution to the score, although still negative, is less than that for row 2 in the table, and works out to approximately -0.40. Thus a rough estimate for the likelihood ratio of the model for events with $M \geq 5$ is about 11.6, yielding a mean gain/event of nearly 1, which is still considerably higher than the value $38.437/84 = 0.46$ obtained for the overall model from the values in row 1 of Table 6.

Considerations of a similar kind arose in the other application mentioned, where the external explanatory variable was obtained from the six series output by the M8 algorithm developed by Keilis-Borok, Kossobokov and others (see HARTE *et al.* 2003, 2004 for details). The main difficulty in this application did not arise from the use of the entropy score, which played a similar role in regard to assessing the performance of the model as in the previous application, but from the problem of selecting an appropriate null model in a situation where the data displayed strong inhomogeneities in both time and space. A further difficulty was that the forecasts were produced for a network of overlapping spatial circles, leading to the additional issue of how the forecasts should be combined to produce a final averaged version. We refer to HARTE *et al.* (2004) for further discussion of these questions.

Acknowledgements

This paper was produced with the help of New Zealand Marsden Fund grant 00–SRA-002-MIS, and the 2002 Programme of the Institute of Mathematics and its Applications, University of Minnesota, Minneapolis. We are grateful to numerous colleagues for comments, references and suggestions, and in particular to Mark Bebbington, Masajiro Imoto, and Yan Kagan.

REFERENCES

BEBBINGTON, M.S. (2004), *Information Gains for Stress Release Models*, submitted to Scandinavian J. Statist.

BEN-ZION, Y. (1996), *Stress, Slip, and Earthquakes in Models of Complex Single-fault Systems Incorporating Brittle and Creep Deformations,* J. Geophys. Res. *101*, 5677–5706.

DALEY, D.J. and VERE-JONES, D., *An Introduction to the Theory of Point Processes, Vol. 1, 2nd ed.* (Springer, New York 2003).

DALEY, D.J. and VERE-JONES, D. (2004), *Scoring Probability Forecasts for Point Processes: The Entropy Score and Information Gain*, J. Appl. Proba. *41A*, 297–312.

HARTE, D., LI, D., VREEDE, M., and VERE-JONES, D. (2003), *Quantifying the M8 Prediction Algorithm: Reduction to a Single Critical Variable and Stability Results*, New Zealand J. Geol. Geophys. *46*, 141–152.

HARTE, D., LI, D.-F., VERE-JONES, D., VREEDE, M., and WANG, Q. (2004), *Quantifying the M8 Prediction Algorithm: Model, Forecast, and Evaluation*, submitted to the New Zealand J. Geol. Geophys., under revision.

HARTLEY, R.V.L (1928), *Transmission of Information*, Bell System Tech. J. *7*, 535–563.

HELMSTETTER, A. and SORNETTE, D. (2003), *Predictability in the ETAS Model of Interacting Triggered Seismicity*, J. Geophys. Res. *108*, 2482, doi:10.1029/2003JB002485.

IMOTO, M. (2000), *A Quality Factor of Earthquake Probability Models in Terms of Mean Information Gain*, Zisin *53*, 79–81.

KAGAN, Y.Y. and KNOPOFF, L. (1977), *Earthquake Risk Prediction as a Stochastic Process*, Phys. Earth Plan. Int. *14*, 97–108.

KAGAN, Y.Y. and JACKSON, D. (2000), *Probabilistic Forecasting of Earthquakes*, Geophys. J. Int. *143*, 438–453.

KLEEMAN, R., MAJDA, A.J., and TIMOFEYEV, I. (2002), *Quantifying Predictability in a Model with Statistical Features of the Atmosphere*, Proc. Nation. Acad. Sci. *99*, 15291–15296.

KLEEMAN R. (2002), *Measuring Dynamical Prediction Utility Using Relative Entropy*, J. Atmos. Sci. *59*, 2057–2072.

LU, C. and VERE-JONES, D. (2001), *Statistical Analysis of Synthetic Earthquake Catalogs Generated by Models with Various Levels of Fault Zone Disorder*, J. Geophys. Res. *106*, 11,115–11,125,

MOLCHAN, G.M. (1990), *Strategies in Strong Earthquake Prediction*, Phys. Earth Plan. Int. *61*, 84–98.

MOLCHAN, G.M. (1991), *Structure of Optimal Strategies of Earthquake Prediction*, Tectonophysics *193*, 267–276.

MOLCHAN, G.M. and KAGAN, Y.Y. (1992), *Earthquake Prediction and its Optimization*, J. Geophys. Res. *97*, 4823–4838.

OGATA, Y. (1988), *Statistical Models for Earthquake Occurrence and Residual Analysis for Point Processes*, J. Amer. Statist. Ass. *83*, 9–27.

OGATA, Y. (1992), *Detection of Precursory Quiescence Before Major Earthquakes Through a Statistical Model*, J. Geophys. Res. *97*, 19,845–19,871.

OGATA, Y. (1999), *Estimating the Hazard of Rupture Using Uncertain Occurrence Times of Paleoearthquakes*, J. Geophys. Res. *104*, 17,995–18,014.

OGATA Y. and ZHUANG, J. (2001), *Statistical Examination of Anomalies for the Precursor to Earthquakes, and the Multi-element Prediction Formula: Hazard Rate Changes of Strong Earthquakes ($M \geq 4.0$) around Beijing Area Based on the Ultra-low Frequency Electric Observation (1982–1997)*, Report of the Coordinating Committee for Earthquake Prediction *66*, 562–570 (Geophysical Survey Institute, Tsukuba, Japan).

RHOADES, D. and VAN DISSEN, R.J. (2003), *Estimates of the Time Varying Hazard of Rupture of the Alpine Fault, New Zealand, Allowing for Uncertainties*, New Zealand J. Geol. Geophys. *46*, 479–488.

SCHORLEMMER, D., JACKSON, D.D., and GERSTENBERGER, M. (2004), *Earthquake Likelihood Model Testing*, to be submitted to BSSA.

VERE-JONES, D. (1998), *Probabilities and Information Gain for Earthquake Forecasting*, Comput. Seismol. *30*, 248–263.

ZHENG, X. and VERE-JONES, D. (1991), *Applications of Stress Release Models to Earthquakes from North China*, Pure Appl. Geophys. *135*, 559–576.

ZHENG, X. and VERE-JONES, D. (1994), *Further Applications of the Stress Release Model to Historical Earthquake Data*, Tectonophysics *229*, 101–121.

ZHUANG, J., OGATA, Y., VERE-JONES, D., MA, L., and GUAN, H. (2002), *Statistical Confirmation of a Relationship Between Excitation of the Low-frequency Electric Field and Magnitude $M \geq 4$ Earthquakes in a 300 km Radius Region around Beijing*, Research Memorandum *847*, Institute of Statistical Mathematics, Tokyo.

ZHUANG, J., VERE-JONES, D., GUAN, H., OGATA, Y., and MA, L. (2005), *Preliminary Analysis of Precursory Information in the Observations on the Ultra-low Frequency Electric Field in the Beijing Region*, Pure Appl. Geophys. *162*, 617.

(Received December 24, 2003, accepted August 5, 2004)

To access this journal online:
http://www.birkhauser.ch