

Predictability and the Assessment of Probability Forecasts

Predicibilità e Assegnazione di Previsioni Probabilistiche

David Vere-Jones

Victoria University and Statistical Research Associates

Wellington, New Zealand

email: David.Vere-Jones@mcs.vuw.ac.nz

Abstract: The development of medium-term (days to years) probability forecasts is becoming an increasingly important and realistic task in connection with many geophysical and other risks. In this talk I address two closely related issues which arise in connection with the development and assessment of probability forecasts for earthquakes, although similar issues arise in many other contexts where the events to be forecast can be formulated in point process terms. The first issue concerns the inherent predictability of a given stochastic model. This turns out to be closely associated with entropy, or entropy rate, of the process, with the constant rate Poisson process representing the maximum entropy model in the point process context, and other processes being assessed by their performance relative to that model. The assessment of probability forecasts can then be thought of as an attempt to estimate their entropy rate relative to a Poisson process. This leads to a natural score statistic, the 'entropy score', which can be thought of as the average 'probability gain' for the forecasts, equivalently as a mean likelihood per event. This scoring method can be compared with more traditional scores based on reduction of the forecasts to a 2 x 2 table of successes and failures relative to the actual occurrence or non-occurrence of forecast events.

Keywords: point processes, earthquake, entropy, Poisson process, probability gain.

1. Introduction

The aim of this paper is to informally describe the mathematical and statistical background to some issues which have become increasingly important in the field of earthquake prediction.

Earthquake prediction is a term used with considerable caution by present-day seismologists. Several decades ago, hopes were unduly raised by a combination of observations and emerging theories that seemed to suggest that genuine and reliable precursors to large earthquakes were to hand. Alas. The observations proved hard to repeat, the reliability of the proposed precursors proved to be low if not negligible, and the theories were superseded or abandoned. The profession of seismology was shamed, money for research on earthquake prediction dried up, and prediction became, and still largely remains, a word better to be avoided.

Such high quality data demands a corresponding quality in the statistical modelling and analysis. From being an aspect of the theory which, to quote Jeffreys (1939) 'is quickly forgotten or altogether disregarded in physical theories,' the probabilistic (or 'error') term in models for earthquake occurrence became a principle subject of study. It is

recognized that our knowledge of the underlying physics, at least at the level of predicting individual events, is still very limited, and that any realistic attempts at forecasting must be associated with error statements or some explicit modelling of the uncertainties. From being a fringe topic, probability forecasting has become a new focus of attention in seismology, with new initiatives, including internationally supported analysis centres, set up to receive and evaluate proposed forecasting schemes. If 'earthquake prediction' is out, long-term, medium-term and even short term probability forecasting is coming in.

Probability forecasts have not proved easy to sell to the general public or to management or to politicians. They require some kind of commitment to long-term strategies whose immediate effectiveness may not be apparent. There are major issues concerning the best way to present probability statements about future events, and their real usefulness in decision-making. But I do not want to touch on these issues here. Rather I want to consider the mathematical issues of developing and assessing probability forecasts for earthquakes and other phenomena which can usefully be represented as 'point processes' - events which occur in time and space, which may have other attributes of importance such as size and orientation and even duration, but for which the interoccurrence times are large compared to the durations, so that each event can be associated with a single point on the time axis.

2. Predictability of point process models

In a practical situation, forecasts will need to take into account not only the time of the event, but also its magnitude, location, and possibly other features. We shall treat the magnitude and location as marks, so that in principle the model we have in view is a *marked point process*, $\{t_n, x_n, M_n\}$. From a practical point of view again, forecasts are made for discrete time intervals (day or month or year, etc) and finite divisions of space and magnitude. Thus the forecasts will take the form of a grid of probability statements $\{\hat{p}_{k,x,m}\}$ governing the occurrence of an event in each cell (k, x, m) of the grid (Gerstenberger, Wiemer, Jones and Reasenber, 2005).

Underlying the performance of the model, however, are properties of the process in continuous time, and conceptual analysis is better done for this continuous model than for a discrete model with arbitrarily chosen divisions of the underlying variables. When the conceptual model is in place, the results for any discretized forecasting scheme can be worked out in principle, using a combination of numerical and simulation techniques.

The key to forecasting the behaviour of a point process is its *conditional intensity* $\lambda^*(t)$, giving, in loose terms, the immediate risk of an event occurring in $(t, t + dt)$ given the past history of the process up to time t :

$$\lambda^*(t)dt = E[dN(t) \mid \mathcal{H}_t], \quad (1)$$

where $\mathcal{H}_t$ is the *history* (σ -algebra) of the process up to time t .

In a renewal process, $\lambda^*(t)$ is just the hazard function of the distribution of inter-event times, evaluated at the elapsed time since the last-occurring event:

$$\lambda^*(t) = h(t - t_{last}) = \frac{g(t - t_{last})}{1 - G(t - t_{last})}$$

where g and G are the density and distribution functions, respectively, of the inter-occurrence time distribution.

In more general point processes, $\lambda^*(t)$ can still be interpreted as the hazard function, but conditioned not only on the elapsed time $t - t_{last}$ but also on knowledge of the previous history up until t_{last} .

This interpretation makes it relatively straightforward to design simulation routines for point processes when the conditional intensity is known explicitly as a function of the past history. These simulation schemes can then form the basis of forward predictions: at any time point t the process can be simulated forward in time, based on the history up to time t , and the probabilities of future events or combinations of events assessed as averages over a large number of such simulations. Examples and further discussion can be found, for example, in section 7.5 of Daley and Vere-Jones (2003), and in the point process section of SSlib (Brownrigg and Harte, 2005).

If the point process is marked (ie an MPP), with marks κ (space and magnitude in the earthquake context) then the conditional intensity can be conveniently written in product form as

$$\lambda^*(t, \kappa) = \lambda_g^*(t) f^*(\kappa|t) \quad (2)$$

where $f^*(\kappa|t)$ is the conditional distribution of the marks, given the past history $\mathcal{H}_t$, and $\lambda_g^*(t)$ is the conditional intensity of the *ground process* of the occurrence times by themselves, irrespective of the mark (but note that the past history now incorporates information on the marks as well as the occurrence times of past events). The simulation routines required for simulating future events then have to be extended to include marks (generated by the conditional distribution $f^*(\kappa|t)$) as well as occurrence times, but this extends the numerical computations involved in the routines rather than their inherent difficulty.

The conditional intensity function is not only a useful characteristic of the point process, but also identifies it uniquely. It follows that if any useful meaning can be attached to the concept of the *predictability* of a point process, it should be expressible in terms of the conditional intensity. Let us consider for simplicity the case of a stationary (in time) simple or marked point process, with histories available over the infinite past $(-\infty, t)$.

Recall that the *entropy* of a discrete probability distribution $\Pi = \{p_k\}$ is defined by the quantity

$$H(\Pi) = - \sum p_k \log p_k. \quad (3)$$

The entropy of a point process (or MPP) over a finite time interval $(t, t + \Delta)$ is a more complicated affair, for two reasons. The first is because it has to take into account, not just the number of events in the interval, but also their locations within the interval. The second is that since the locations are specified exactly within a continuum, the entropy as defined above becomes infinite, and meaningful results have to be obtained by taking conditional entropies. Here the term $\log p_k$ above is replaced by $\log(p_k/p_k^o)$ where p_k^o is a convenient *reference probability distribution*, which in this context is most conveniently taken to be that of a stationary Poisson process with the same average rate, say m as the original process. In the case of an infinitesimal interval $(t, t + dt)$, either no events occur in the interval, or just one does, with (conditional) probability $\lambda^*(t)dt$. Taking out the conditioning we are left with the expression

$$E \left[\lambda^*(t) \log \frac{\lambda^*(t)}{m} \right] dt \equiv G dt \quad (4)$$

for the increment in the relative entropy. From stationarity, the coefficient of dt is independent of t , and defines the *information gain per unit time* G of the given process relative

to the Poisson process with the same rate m . Alternatively, we may divide by the mean rate $m = E[\lambda^*(t)]$ to obtain an *information gain per event*, say $G_e = G/m$.

It is this quantity G , or its equivalent G_e , that we suggest as best representing the predictability of the process. It is essentially a normalized version of the Kullback-Leibler distance between the distributions of the given point process and the nearest matching Poisson process, both restricted to a finite interval. The predictability of a Poisson process is then zero, which reflects its role as the point process with maximum entropy subject to a given rate.

Suppose we now think back to the problem of stating the probability forecasts in a discrete-time setting. The analogous quantity to the conditional intensity is the conditional probability p_n^* of an event occurring in the time interval $(n, n + 1)$, given the history of the outcomes up to n . The expected value $E[p_n^* \log p_n^*]$ again describes an entropy rate, in this case per time step. Again it is better to take as the argument of the logarithm the ratio to a reference probability, say of iid trials with the same expected probability of success $\bar{p}$, say. The resulting information gain $E[p_n^* \log(p_n^*/\bar{p})]$ provides information about the relative predictability of the two models over a single time step in the discrete scheme. Dividing by the length of this time step gives us again an information gain per unit time. It is then quite easy to show that the information gain for the discrete forecasting scheme cannot exceed that of the continuous scheme (for one reason, its forecasts are based on less of the past history than the continuous scheme). Thus the information gain for the underlying continuous time model not only provides a numerical characteristic of the model, but also an upper bound on the effectiveness of forecasts that can be made from any approximating discrete model.

A more formal treatment of these ideas is in Daley and Vere-Jones (2004), with examples and illustrations in Harte and Vere-Jones (2005). One particular example from these papers shows the information gains for a range of renewal process models with gamma-distributed interevent times. The information gain for a renewal process can be written in the form

$$G = m \left[1 - \log m + \int_0^\infty f(y) \log f(y) dy \right], \quad (5)$$

where m is the mean rate and $f(y)$ the density of the interevent distribution. Keeping the mean rate constant, the highest gains are for the models that are closest to deterministic, or else have highly skewed (J-shaped) distributions, leading to very tight clusters (many short intervals; occasional very large intervals) and major differences from a Poisson process.

For a MPP the information gain has to be averaged over the gains for different values of the marks. The corresponding expression is

$$G = E \left[\lambda^*(t, \kappa) \log \frac{\lambda^*(t, \kappa)}{m_g f_0(\kappa)} \right] = E \left[\lambda_g^*(t) \log \frac{\lambda_g^*(t)}{m_g} \right] + m_g E \left[\log \frac{f^*(\kappa|t)}{f_0(\kappa)} \right], \quad (6)$$

where m_g is the ground rate and $f_0(\kappa)$ is the stationary mark distribution. The reference probability here is taken to be that associated with a compound Poisson process with the same ground rate as the original process, and mark distribution equal to the stationary mark distribution of the original process.

We conclude by noting that the above discussion assumes that the model is *complete* in the sense that it specifies not only the probability structure of the point process, but also of any auxiliary quantities (precursors) which may be used in constructing the forecasts.

In the case of a *partial model*, where the structure is known conditionally on the observation of the auxiliary quantities but their distribution remains unknown, the long-term performance of forecasts also remains unknown, since the long-term behaviour of the precursors is not specified. For a complete model, expectations can be taken across the joint distributions of both the auxiliary quantities and the observed points of the process.

3. The assessment of probability forecasts

Early studies on the assessment of prediction schemes were based on decision rules which resulted either in 'failure' (the predicted result did not occur) or 'success' (it did occur). A probability forecasting scheme as such does not embody any decision rules; it merely states the probability of any postulated future event, given the past. To bring such schemes into play, an arbitrary decision rule - such as forecast that an event will occur if its forecast probability exceeds some given threshold - has to be imposed. This introduces an undesirable element of arbitrariness into the assessment procedure. Recently, however, procedures have been developed for summarizing the information in a whole family of such schemes, as the threshold probability is increased from zero to one. In this section we shall describe two possible procedures for assessing probability forecasts for point processes. The first is based on the ideas of information gain introduced in the preceding section, and the second on procedures for summarizing information over a family of dichotomous schemes.

Before describing either approach let us first deal with one statistical issue. From the classical statistical point of view, the crucial issue is not so much assessing the forecast as assessing the model. Tests of the model should encompass tests based on the forecasts, since the latter reflect only one aspect of the performance of the model.

However, this point of view ignores the fact that the model may be fitted by methods which do not seek to optimize the forecasting performance of the model. It also ignores the basic point that when the model is uncertain (the one certain thing about earthquake modelling!) errors in the model specification, or in the parameters within a specified parametric form, need to be taken into account. In general it is safer to take a weighted mixture of forecasts from a range of models than to base the forecast exclusively on any one model, even if this last is optimal in some sense.

For the rest of the discussion, we consider only forecasts for a discrete-time set-up, and suppose that, by whatever procedure, the forecaster is able to produce a probability, at each time step, that a specified event will or will not occur within the forecast time interval. Extensions to marked point processes, when the forecaster allocates probabilities to a range of possible outcomes associated with particular marks, are possible, but through shortage of space we shall not deal with them here.

3.1 Probability gains and the entropy score

Let us again denote by p_n^* the forecast probability of the occurrence of an event at time step n , and by $\bar{p}$ some kind of reference probability for the same event, for example the long-run average probability. The *probability gain* $p_n^*/\bar{p}$ is a natural measure of the performance of a single forecast in the case that the predicted event actually occurs. Similarly $(1 - p_n^*)/(1 - \bar{p})$ is a natural measure of the performance when it does not occur. In a good forecasting scheme, both should be above unity most of the time.

Taking logarithms, and then taking averages over the 0 – 1 sequence of events X_n , $n = 1, 2, \dots, N$ we obtain the *entropy score*

$$\hat{G}_N = (1/N) \sum_1^N \{X_n \log[p_n^*/\bar{p}] + (1 - X_n) \log[(1 - p_n^*)/(1 - \bar{p})]\}, \quad (7)$$

which is nothing other than the mean log-likelihood ratio $(1/N) \log[L_N^*/L_N^0]$, where L_N^* is the likelihood of the forecast probabilities, considered as a chained sequence of dependent trials with prescribed conditional probabilities, and L_N^0 is the likelihood for the reference model. The expected value of this quantity reverts to the information gain G when the model used for forecasting is the true model. In general it can be interpreted as the difference in the Kullback-Leibler distances between the forecast model and the true model, and the reference model and the true model. In a stationary situation, we expect the sample value $\hat{G}_N$ to converge to its expected value; this leads to the possibility of using $\hat{G}_N - G$ as a statistic for testing the goodness of fit of the model used to generate G (Bebbington, 2005). The use of the entropy score is not restricted to situations where the probabilities are generated from a complete model for the point process. It can be just as useful when dealing with a partial model, or a purely ad hoc specification of the probabilities. By restricting the summation to special classes of events (for example the events successfully predicted, or those not predicted, or the events within a certain distance range of the observation station, etc, it can be used to provide helpful insights into the behaviour of the forecasts and of the process under study (cf the discussion of forecasts based on electrical signals in Zhuang *et al.* (2005)).

3.2 The R-score and the Molchan diagram

Continuing the discussion and notation from the previous section, we consider the commonest type of decision rule based on a probability forecast, namely to take action (meaning in most situations to predict the occurrence of the event) if and only if p^* lies above a certain threshold p_0 . The choice of threshold p_0 is somewhat arbitrary, but in principle should be based on a cost-benefit analysis of the costs of different outcomes. A common choice is to take $p_0 = \pi$, although this can lead to considerable overforecasting 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.

This 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 failures to predict,
- (c) number of successful forecasts of non-occurrence,
- (d) number of false alarms.

Let X_n denote the occurrence of the event of interest at time n , and Y_n the issuance of a forecast for its occurrence at time n . Then $a = \sum_1^N X_n Y_n$; $b = \sum_1^N X_n (1 - Y_n)$; $c = \sum_1^N (1 - X_n) (1 - Y_n)$ and $d = \sum_1^N (1 - X_n) Y_n$. Dividing the table entries by N , to convert the numbers into proportions, and assuming stationarity, the corresponding expected values can be written as

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

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 prob-

ability 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 approximately represented as $\nu = E[X(1 - Y)]/E[X]$.

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

$$R = \frac{ac - bd}{(a + b)(c + d)} = \frac{a}{a + b} - \frac{d}{c + d} = \frac{c}{c + d} - \frac{b}{a + b}. \quad (8)$$

The middle form, for example, can be interpreted as the proportion of events successfully forecast, less the proportion of false alarms out of the total number of non-events.

Using the above notations,

$$R = \frac{\pi}{\tau(1 - \tau)}[(1 - \nu) - \tau], \quad (9)$$

showing that the essential contribution to the *R - score* comes 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, $R = -1$.

(Molchan, 1990) suggested the following extension of the procedures derived from the 2×2 table, assuming that this is based on the number of probability forecasts over a nominated threshold. For every possible threshold p_0 , $0 < p_0 < 1$, plot the proportion of failures to predict

$$\nu(p_0) = \frac{c}{a + c}$$

against the fraction of time on forecast

$$\tau(p_0) = \frac{a + b}{a + b + c + d}$$

The resulting $\nu - \tau$ diagram provides a comprehensive summary of the behaviour of the probability forecasting scheme. 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 $(1, 0)$. In more general cases, the diagram typically consists of a convex (downwards) curve lying below the diagonals, and reducing to the two $(0, 1)$ segments of the axes in the case of perfect prediction.

In fact the $\nu - \tau$ plot is essentially a Q-Q (quantile-quantile) plot of two distribution functions against each other, the first of the distribution of the time on trial, and the second (in the opposite sense) of the proportion of failures to predict. Further analysis is given in Daley and Vere-Jones (2004), where it is shown that 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. The $\nu - \tau$ curve based on the true probabilities lies always below the line $\nu = 1 - \tau$, and is always convex downwards. The more predictable the process, the more rapid the initial change in slope, and the closer the graph approaches to the extreme case of perfect prediction. Moreover, the curve based on use of the true probabilities provides a lower envelope for the $\nu - \tau$ curves generated by other prediction schemes.

Finally, we consider the connection between the entropy score (information gain) and the Molchan diagram. Although in principle the former can be written down in terms of an integral involving the same quantities entering into the construction of Molchan diagram, there does not seem to be any simple way of relating the curve to the information gain. It is one numerical characteristic of the forecasting scheme, but there are many others, such as the area between the curve and the diagonal, which could also be used to characterize the predictability of the model. Again some further discussion is given in Harte and Vere-Jones (2005).

Acknowledgements

I am grateful to my colleagues David Harte and Mark Bebbington for discussions and assistance.

References

- Bebbington M. (2005) Information gains for stress release models, *Pure and Applied Geophysics*, 162, 2299–2319.
- Brownrigg R. and Harte D. (2005) Using r for statistical seismology, *R New*, 5, 1, 31–35, ISSN 1609–3631. URL: [cran.r-project.org/doc/Rnews/Rnews 2005-1.pdf](http://cran.r-project.org/doc/Rnews/Rnews%2005-1.pdf).
- Daley D. and Vere-Jones D. (2003) *An introduction to the Theory of Point Processes, (2nd Edition) Vol I: Elementary theory and methods*, Springer, New York.
- Daley D. and Vere-Jones D. (2004) Scoring probability forecasts for point processes: entropy score and information gain, *Journal of Applied Probability*, 42, 297–312.
- Gerstenberger M., Wiemer S., Jones L. and Reasenber P. (2005) Real-time forecasts of tomorrow's earthquakes for california, *Nature*, 435, 328–331.
- Harte D. and Vere-Jones D. (2005) The entropy score and its uses in earthquake forecasting, In Ben-Zion, Y, Zuniga, R., and Vere-Jones, D. (Eds), *Statistical Seismology; Pure and Applied Geophysics*, 162, 1229–1254.
- Jeffreys H. (1939) *Theory of Probability*, Cambridge, University Press.
- Kagan Y. and Knopoff L. (1977) Earthquake risk perediction as a stochastic process, *Physics of the Earth and Planetary Interiors*, 14, 97–108.
- Molchan G. (1990) Strategies in strong earthquakes prediction, *Physics of the Earth and Planetary Interiors*, 61, 84–98.
- Zhuang J., Vere-Jones D., Guan H., Ogata Y. and Ma L. (2005) Preliminary analysis of observations in the ultra-low frequency electric field in the beijing region, In Ben-Zion, Y, Zuniga, R., and Vere-Jones, D. (Eds), *Statistical Seismology; Pure and Applied Geophysics*, 162, 1367–1396.