Issues in the Development and Evaluation of Earthquake Forecasts

by

David Vere-Jones

Victoria University and Statistics
Research Associates,
Wellington, New Zealand

It is an honour to be invited to speak at the ERI, and I am very grateful to the organizers and my host Yosihiko Ogata for the invitation and for arranging my visit.

These are briefly the contents of my talk:

1. Reminiscences

2. The RELM framework

3. Assessment of probability forecasts

4. Summary and conclusions

REMINISCENCES

In fact this is the second lecture I have givern to ERI. The first was in 1976 and was entitled

"*Earthquake Prediction: a Statistician's View*"

It is somewhat embarrassing that this could easily have served as the title for today's talk. It prompts me to ask, what has changed between then and now?

From my angle, some things have changed a lot, and others not much. Some of the crucial developments have been:

(i) Increase in the qantity and quality of seismological data.

A huge development. Catalogue data, in particular, but other sources also.

Better and more extensive data calls for better and more extensive statistical analysis.

Indeed, the role of statistical analysis in the interpretation of seismological data has also increased greatly. It has been a major factor in

(ii) The rise of statistical seismology.

In 1976, there was just a scattering of scientists around the globe trying to develop statistical models for catalogue data.

Among them were Utsu, Aki, and Mogi in Japan; Gaisky in Russia; Yan Kagan and Leon Knopoff in California; Cinna Lomnitz in Mexico, and myself in New Zealand.

Now it has become an important subdiscipline, with many members, and an increasing range of topics and applications.

Today's meeting is an illustration of these features.

(iii) Development of good models.

Here there has been less progress.

The development of better models, embodying a better understanding of the physical processes which determine the probabilities of earthquake occurrence, is still the number one priority.

Earthquake occurrence is a complex process, and catalogue data by itself may never be enough to unravel the process fully, nor therefore to predict it more than very partially.

We may be reaching the limit of what may be achieved by caalogue analyses alone. But we are still very far from being able to incorporate other factrors into quantitative predictions.

(iv) The statistical mis-education of seismologists.

This is something else which still continues.

The problem is that Geophysics Departments are committed to other courses, and there is never room for a course on 'Statistical Seismology' or something similar.

From other courses or from chance remarks, students gain a wrong impression of the potential and importance of these topics in modern geophysics. Ultimately they struggle to find and use some advanced statistical software for their experimental data. They rarely get a good background for independent work in these areas.

(v) Influence of Frank Evison

I want to briefly pay tribute to Frank's work. He was a pioneer and persistent contributor to earthquake prediction.

It was Frank who encouraged my early interest in seismology. From the beginning, we both shared the view that earthquake predictions should be couched in terms of probabilities.

I borrowed many ideas from him, including several which appeared in my 1976 talk and will resurface today:

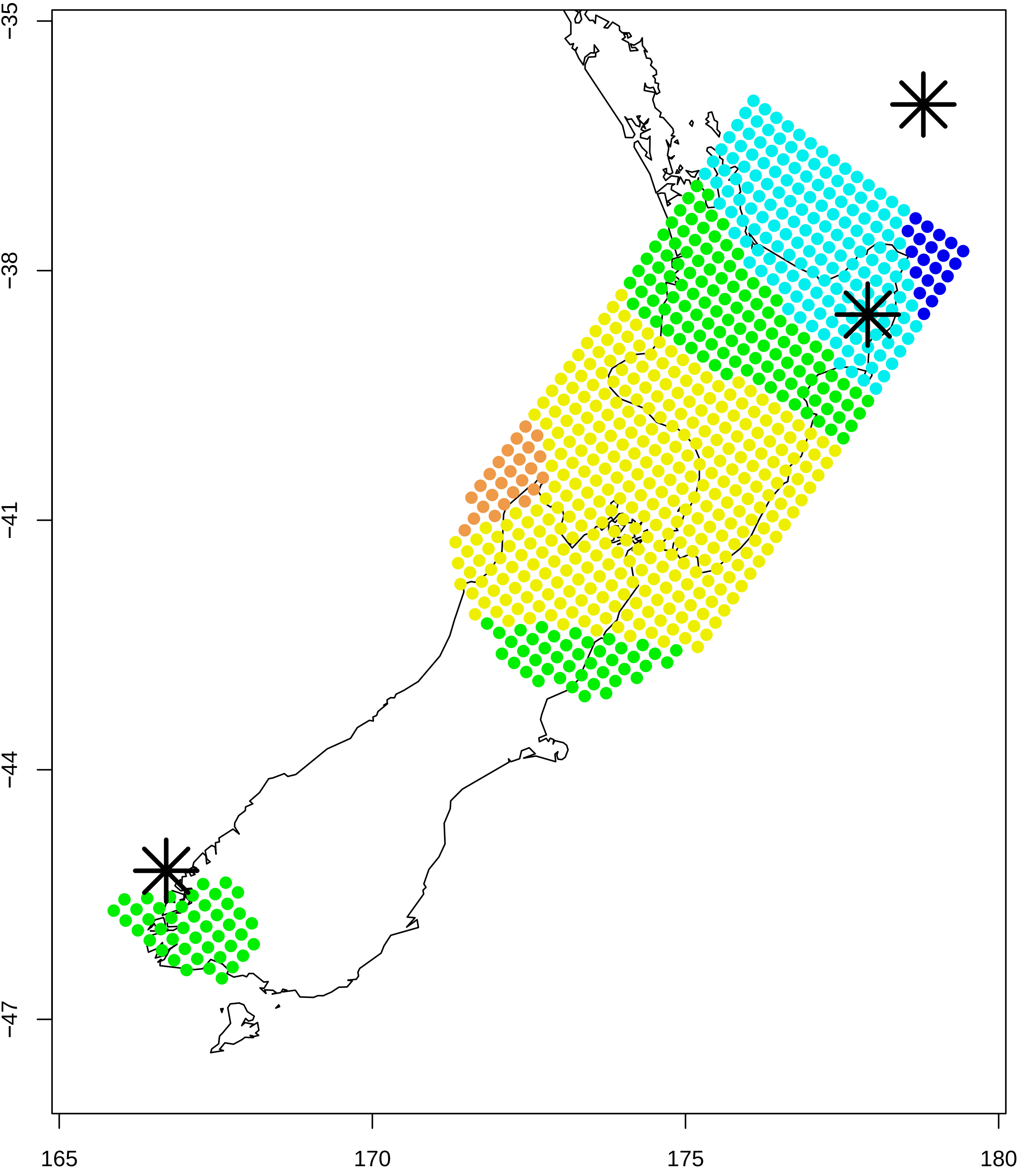
Synoptic forecasts

Risk enhancement factors (Continuous version of probability gain).

Recently he and David Rhoades spent many years developing and improving the precursory swarm and EEPAS models. I think it has become the most successful of the current medium-term forecasting models.

Synoptic Forecast: 01Jul1993−31Dec1993

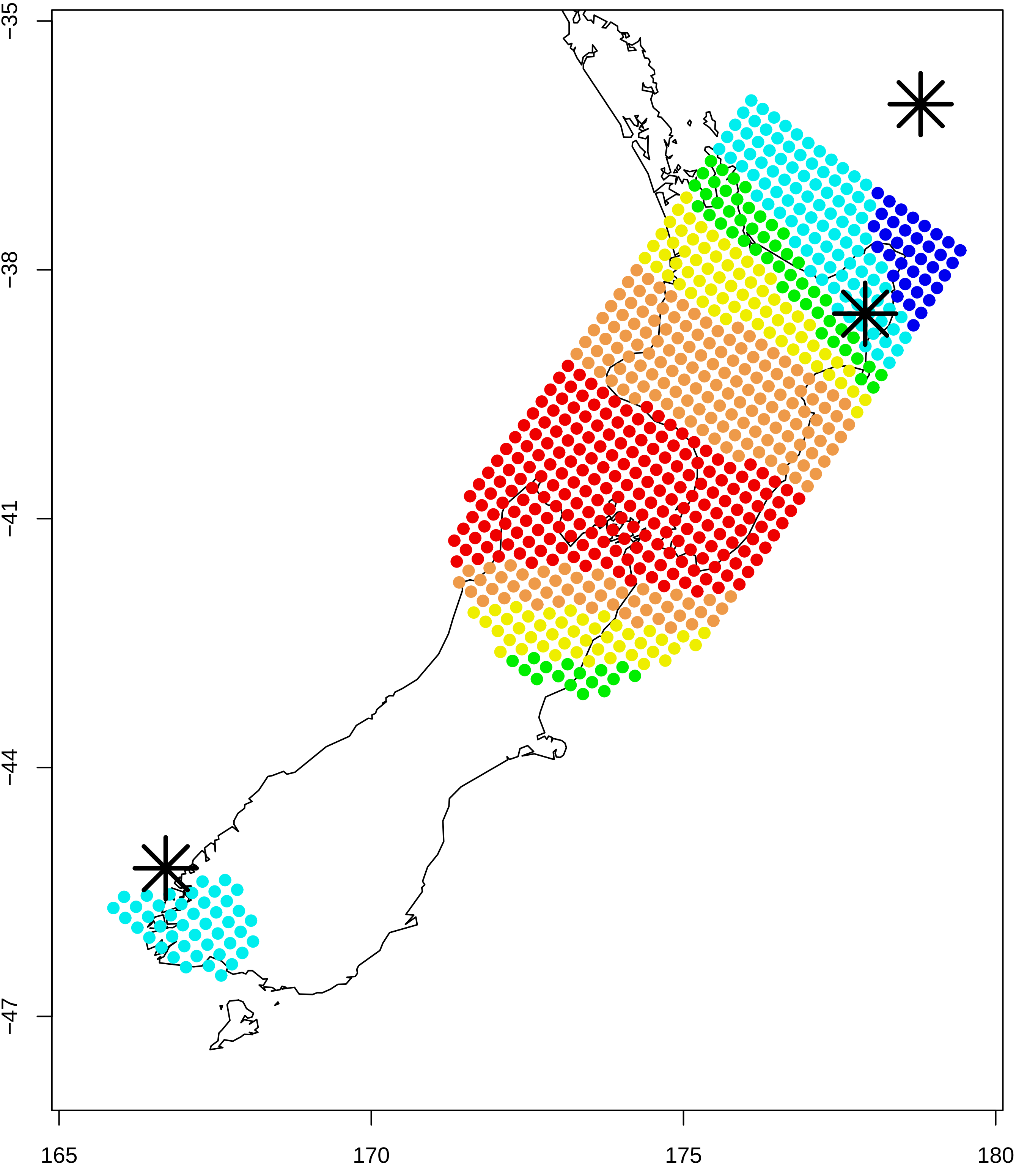
All/All $M_{min} = 4$ $M_0 = 6.8$ $M_1 = 6.3$ $R_1 = 167.3$ Lag = 1



$0 \le$ blue $<$ 2e−04 $\le$ cyan $<$ 4e−04 $\le$ green $<$ 5e−04 $\le$ yellow $<$ 6e−04 $\le$ tan $<$ 7e−04 $\le$ red $<$ 9e−04

Synoptic Forecast: 01Jul1993−31Dec1993

All/All $M_{min} = 4$ $M_0 = 6.8$ $M_1 = 6.3$ $R_1 = 167.3$ Lag = 1



0 ≤ blue < 2e−04 ≤ cyan < 4e−04 ≤ green < 5e−04 ≤ yellow < 6e−04 ≤ tan < 7e−04 ≤ red < 0.0015

TOWARDS FORECAST TESTING CENTRES

Let me return to the main theme of my talk. Rightly or wrongly, I felt my most useful contribution might be to provide an independent commentary on the RELM procedures.

My comments are based mainly on informal discussions with colleagues in Wellington, and the two recent papers in Seismological Letters by Daniel Schorlemmer and co-authors.

My aim has been to try and pick out those points where, if I was starting up a new centre, I think some further discussion might be helpful before the schemes were finalized.

The RELM framework

2.1 Benefits from adopting a common framework

I support the view that there are benefits in imposing a common framework on the format and assessments of earthquake probability forecasts. It is necessary to bring some order in what has tended to be a very disordered field.

From my own personal experience, despite having been a somewhat reluctant starter, I have found the discipline of being required to shape the forecasts to a particular framework quite helpful. Aspects you had hoped to brush under the carpet are forced out into the open. This is ultimately helpful.

The framework does not have to be perfect, nor does it have to be the only procedure used. But using a common framework casts a much sharper light on the differences as well as the relative merits of different schemes.

2.2. Framework overview

The forecasts are currently prepared for a rectangular grid system. Individual probabilities are required for disjoint cells. Each cell is defined by a

$$location \times magnitude \times time....$$

coordinate set, and has a fixed length in each dimension.

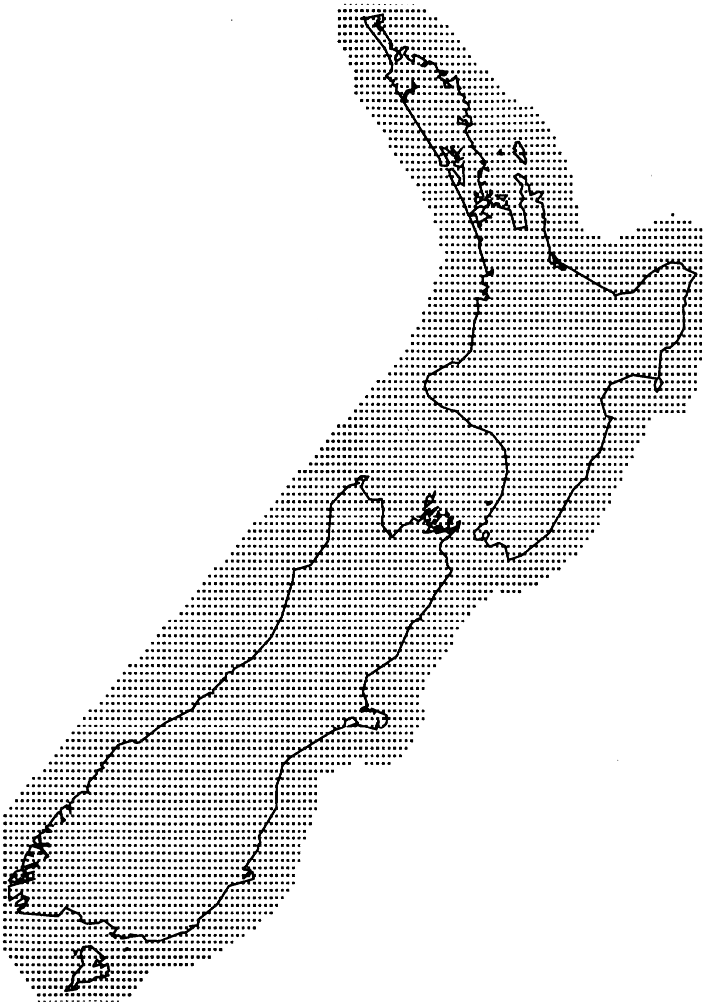
Submitters of a forecasting scheme have to provide programmes which calculate probability forecasts for as many cells in the grid as they can. Scoring is based on the performance of the forecasts over those cells for which forecasts are provided.

There are pros and cons in the choice between continuous models (ie based on densities) and cell-based discrete models, but I see no major advantages either way, nor difficulties in tranferring from one to the other.

2.3. The choice of space, magnitude and time scales

New Zealand, like California, has opted for a very fine spatial mesh, also for fine magnitude gradations. This is somewhat burdensome for the computations, but helps to keep the forecasts independent, given the forecast information.

The larger the cells, the greater the opportunities for unwanted complexities to creep in.



Along the time axis, however, the forecasts are not divided into cells. Instead, the forecaster is offered a limited choice of forecasting intervals, of very different lengths: one day, three months, one year, five years.

2.4 Why not shorter, disjoint time intervals?

The more I reflect on it, the more I wonder if these conventions might not be reconsidered.

From my view point, a forecaster who provides only 5-year forecasts should not mind if his forecasts are broken to shorter periods.

This can be done, if crudely, by supposing the event to be distributed uniformly at random over the long interval, so that equal weight is given to subintervals of equal length.

For example, a scheme giving probability 0.1 for a 1-year period can be treated as a scheme giving equal (and independent) probabilities

$$(0.1)^{1/365}$$

for successive 1-day intervals. The probabilities of at least one event over the 1-year period are the same.

The important point to the forecaster is that his forecasts are still assessed over the longer interval.

The advantage of using the shorter periods is that the forecasts can be used and compared more flexibly, in a wider range of ways.

In their accounts of the RELM procedures, the authors emphasize the qualitative difference between forecasts aimed at 1-day intervals and forecasts aimed at 1-year or 5-year intervals.

To my mind the differences lie not in the nominal forecast period, but in the nature of the phenomena being studied. 1-day forecasts attempt to capture sudden changes in risk. 1-year forecasts attempt to capture changes which occur slowly over longer time periods.

By dividing both types of forecasts into 1-day intervals, the forecasts can be compared and assessed over time periods of any desired length, 1-day or 1-month or 1-year.

2.5 What about long-range (gap) forecasts?

I mean by this a forecast made (say) now, for a magnitude 6 event in Wellington in the first 6 months of 2010.

Unless I have misunderstood the formulation, there seems to be no place for such forecasts in the present scheme. Such long-range forecasts are useful in meteorology, why not here?

Such possibilities arise naturally when a longer forecast is broken into shorter pieces.

Breaking up the forecasts into periods of shorter length, but allowing for varying time-gaps, might yield more precise and more flexible forecasts than the present system of forecasts for a restricted range of fixed periods.

The question of the range is also relevant to the question of when forecasts might be refreshed. For assessment purposes, you would not want to refresh a time-lapse forecast before its time had elapsed. But after that, why not?

The ultimate goal might be something more like a synoptic forecast over both time and space, frequently updated.

2.6. Probabilities v Poisson rates

A special feature of the RELM procedures is that, although probability forecasts are in view, forecasters are requested to frame their forecasts in terms of expected rates per cell.

This stipulation represents only a slight logical gloss.

The Poisson approximation to p_0 is

$$e^{-\mu} \approx 1 - \mu \approx p_0$$

,

$$\mu = p_1 + 2p_2 + 3p_3 + ...$$

which is very close for small cells and small probabilities. Moreover probabilities are very easily converted into rates and vice versa.

A more serious gloss, I think, is in the further assumption that, given the information on which the forecasts are based, the events in different cells, whether described by rates or probabilities, can be treated as independent.

Consider, for example, a 5-year model which suggested that every time a magnitude 6 event occurs in the Tokyo cell, a similar event will occur in the Kamakura cell within the next 6 months, i.e. within the same 5-year time slot.

If the probability of either event individually is 1/10, and they are treated as independent, the probability of both occurring in the same 5-year period together is 1/100, whereas the modeller would like to give this joint occurrence essentially the same probability (1/10) as the Tokyo event.

This model cannot be fairly scored within the independence scheme.

Whenever, as in this example, joint probabilities are provided, they should be used. They can be converted to rates for joint occurrences if desired.

The difficulties lie not in the scoring, but in devising models capable of producing such joint probabilities. And, ultimately, that is what is needed.

2.7. Aftershocks and other complexities.

Aftershocks represent the characteristic form of earthquake dependence. and raise in an extreme form the kind of difficulty described in the previous example.

In principle, I feel that forecasts should be for real scenarios, aftershocks included.

In the present framework, forecasters are allowed to dodge the issue by stipulating, for example, that their forecasts are for ‘independent events’ only.

In order to score such forecasts, the authors of the scheme are forced into elaborate and somewhat arguable procedures:

(a) calculate the probabilities that an observed earthquake is either truly independent, or is an aftershock,

(b) test the forecasts, not against the actual occurrences, but against a whole family of simulations which randomly describe the earthquakes which occur as either ‘independent’, or ‘aftershock’, in accordance with the estimated independence probabilities.

From my viewpoint, the onus should be on the forecaster to find a way of (say) forecasting the probability of a magnitude 6 event, which allows for the possibilities that that event might be independent, or might be an aftershock of some previous event.

If the forecasters want to use independence probabilities to help them on their way, that is fine, but the types of event they end up forecasting must be clearly identifiable. This is not the case for 'independent events'.

Ultimately, forecasts should aim to give the joint probabiliti es for a whole scenario, aftershocks and errors included.

3. Assessment of the probability forecasts

From the classical statistical point of view, assessing probability forecasts based on a model is assessing only one aspect of the model's performance. The crucial issue is assessing the model itself.

However, in the earthquake context at least, it is considered essential to directly assess the forecasts. There may be aspects of the model selection or model fitting which have been glossed over or remain hidden. Hence considerable emphasis is placed on careful tests of the forecasts themselves.

Early studies often assumed that the schemes take the form of decision rules which result either in ' failure' (some predicted result did not occur) or 'success' (it did occur).

A probability forecasting scheme does not of itself embody any decision rules, so some alternative form of assessment is needed.

Many schemes can be suggested for this purpose, but personally, I am quite happy with the likelihood scores suggested for RELM. However, I would introduce them somewhat differently, as below.

3.1 Probability gains and likelihood scores.

Consider forecasts for a single cell, and denote the forecast probability that an event occurs in the n-th time step of length Δ by $p_n^*(\Delta)$, and the corresponding reference (long-run) probability by $\bar{p}(\Delta)$.

The *probability gain* $p_n^*(\Delta)/\bar{p}(\Delta)$ is a natural measure of the performance of the forecast when the predicted event actually occurs. Similarly $(1 - p_n^*(\Delta))/(1 - \bar{p}(\Delta))$ 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, $n = 1, 2, \ldots, N$ we obtain the *entropy score*

$$\begin{aligned}\widehat{G}_N(\Delta) &= (1/N)\sum_1^N \Big\{ X_n \log \frac{p_n^*}{\bar{p}(\Delta)} \\ &+ (1 - X_n) \log \frac{1 - p_n^*(\Delta)}{1 - \bar{p}(\Delta)} \Big\}.\end{aligned}$$

It is this quantity which is used as a measure of the forecast performance. It is just the mean log-likelihood ratio

$$(1/N)\log[L_N^*/L_N^0].$$

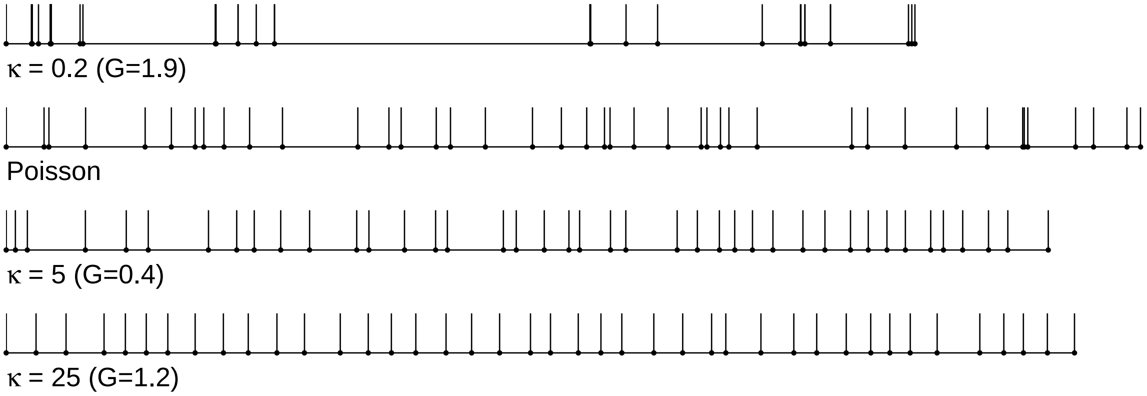
Its expected value and hence also its long-run value in a stationary scheme, is a constant, which we may call the *information gain* G_Δ.

Choosing the model with the highest log-likelihood ratio is the same as choosing the model that produces the best average log probability gains.

G_Δ may also be interpreted as a measure of the 'predictability' of the model generating the data, or of the distance (Kullback Leibler distance) of the model being tested from the reference model.

For example, the diagrams show the value of G_Δ for a family of renewal models, sometimes used to describe recurrence times of large events on a single fault.

All models were assumed to have the same average rate of occurrence.


κ = 0.2 (G=1.9)
Poisson
κ = 5 (G=0.4)
κ = 25 (G=1.2)

The highest gains are for the models that are closest to deterministic, or else have highly skewed (J-shaped) distributions.

3.2 Testing the models: pairwise comparisons

The RELM procedures focus on two types of tests. The main test is based on the likelihood scores as described above, and compares the performance of two models.

Because we are concerned with ratios, there is no difficulty in swapping from one reference model to another, or to comparing models among themselves on a pairwise basis. Thus we can compile a table of pairwise comparisons.

The choice of model used as reference is unimportant here; the pairwise comparisons will be unaffected..

Selecting the best from a class of models is more difficult, because the best model with respect to one reference model may not be best with respect to some other reference model.

Hence there is current discussion about the best reference model to use in developing an overall ranking.

Falling back on absolute probabilities does not avoid this problem, because it amounts to making a comparison with a model that allots equal probabilities to each cell.

3.3 Testing the models: consistency tests

The authors also devote considerable effort to a further question: is the model producing the forecasts compatible with the events it purports to forecast?

This is nothing other than the question of goodness of fit. How well does the model fit the somewhat selected class of data for which forecasts are being prepared?

The authors suggest using the numerical value of the loglikelihood (or its ratio against some fixed alternative) as the test statistic.

A supplementary test is also proposed in which the feature tested is the total number of events in a range of interest.

In any such case the test proceeds by computing (analytically or by extensive simulations) the distribution of values of the proposed characteristic under the assumption that the model is true.

If the value from the real data is out in the tails of this distribution (i.e. very uncommon under the model) then the model is rejected. If it is in the area of common values, then the model may be accepted. Choosing how far out in the tails the real data must be before the model is ‘rejected’ corresponds to choosing the significance level.

To pacify those scientists whose model is rejected because the likelihood is too high, they suggest using a further test (that based on total numbers) to confirm that the model is unreasonable.

The advantage of the likelihood tests is that they can be applied to any model for which the likelihood can be computed. However, I know little about the statistical properties of such tests.

3.4. Who wants probability forecasts, anyway?

This is not quite a humorous issue. There are genuine grounds for doubt as to whether probability forecasts for earthquake occurrence can be practically useful.

For the scientist, I think it is a matter of scientific integrity that every forecast should be accompanied by some attempt to quantify its uncertainty.

The only satisfactory basis for quantifying uncertainty is probability theory. Hence quantifying the uncertainty means moving to a probability framework.

Where this is not done, then I think the scientist's job is incomplete.

But for communicating such information to user groups and the general public, probability forecasts as such may not be very effective. Where possible, people prefer definite answers.

Probability forecasts can be formulated in betting terms, or incorporated into decision rules, cost-benefit analyses, and so on. These may provide a better basis for public announcements.

Summary and conclusions

The main points of my discussion are summarized below.

- There are substantial benefits to scientists concerned with earthquake prediction in having to produce forecasts within a well-specified framework, and having them evaluated within that framework. If the framework outline, at least, can be shared internationally, that could be a further benefit, facilitating exchange of ideas and techniques.

- The RELM framework seems to me a sound starting point, although I have reservations about some aspects, as set out below:

(i) There may be merits in breaking down long-term forecasts into a sequence of forecasts for shorter periods. The performance of different types of models can still be compared over intervals of any desired lengths.

(ii) Long-range forecasts (forecasts with gaps) could be given a place within the framework.

(iii) Replacing probabilities by Poisson rates is not a severe approximation, but the independence assumptions accompanying this approximation are of some concern, and merit revue.

(iv) The present approach of 'doctoring the data', to deal with problems arising from aftershocks and catalogue errors, is also questionable. Ultimately, the aim should be to For example, a scheme giving probability 0.1 for a 1-year period can be treated as a scheme giving equal (amd independent) probabilities

$$(0.1)^{1/365}$$

for successive 1-day intervals. The probabilities of at least one event over the 1-year period are the same. allow for such features in producing the forecasts. The forecasts should be directly assessible in terms of the actual data.

(v) More information would be helpful on the power of the likelihood tests used in assessing a model's consistency with the data.

- To my mind, the biggest problem still lies in developing better models.

  I hope that better assessment of existing models, through setting up testing centres such as RELM, may help to focus our ideas on just this issue.

Some further references

Bebbington, M.S. (2004) Information Gains for Stress Release Models.

Daley, D.J. and Vere-Jones, D. (1988). *Introduction to the Theory of Point Processes*; (2003) *2nd edition, Vol 1 2003; Vol 2 2007* Springer, New York etc.

Molchan, G.M. (1990). Strategies in strong earthquake prediction. *Phys. Earth Plan. Int.*, **61,** 84–98.

Molchan,G.M. and Kagan Y.Y (1992) Earthquake prediction and its optimization. Jl Geophysi. Res. **106** 4823-4838.

Shi, Y., Liu, J., Zhang, G. (2001) An evaluation of Chinese annual earthquake predictions , 1990-1998. J. Appl Prob. **38A** 222-231.

Vere-Jones, D. and Daley D.J. (2004) Scoring probability forecasts for point processes: entropy score and information gain. *Journal of Applied Probability* **42A**, 297–312.