Shaking Down Earthquake Predictions

Department of Statistics
University of California, Davis
25 May 2006

Philip B. Stark
Department of Statistics
University of California, Berkeley
www.stat.berkeley.edu/∼stark

joint work with David A. Freedman, Brad Luen

Outline

Earthquake phenomenology; precursors; stochastic models

Forecasts: What is the chance of an earthquake?

Coin tosses:

- Equally likely outcomes
- Frequency theory
- Subjective theory
- Probability models

Weather predictions

The USGS forecast for the SF Bay Area

Evaluating predictions: null hypotheses, common tests

It's easy to predict earthquakes!

Earthquake Phenomenology

Clustering in space:

- About 90% of large events in "ring of fire" (circum-Pacific belt, plate margins)
- Most earthquakes are on pre-existing faults
- Depths 0–700 km; most are shallow; most large quakes are shallow

Clustering in time:

- Foreshocks, aftershocks, swarms

Globally, on the order of 1 magnitude 8 earthquake per year.

Locally, recurrence times for big events $O(100\text{ y})$.

Big quakes deadly and expensive.

Much funding and glory in promise of prediction.

Would be nice if prediction worked.

Claimed precursors:

- foreshocks, patterns
- electromagnetics in ground and air; resistivity
- cloud formations
- infrared
- well water composition, temperature and level
- geodetics
- animal behavior

Some stochastic models for seismicity:

- Poisson (spatially heterogeneous; temporally homogeneous; marked?)
- Gamma renewal processes
- Weibull, lognormal, normal, double exponential, ...
- ETAS
- Brownian passage time

Coin Tosses. What does $P(\text{heads}) = 1/2$ mean?

- Equally likely outcomes: Nature indifferent; principle of insufficient reason
- Frequency theory: long-term limiting relative frequency
- Subjective theory: strength of belief
- Probability models: property of math model; testable predictions

Math coins $\neq$ real coins.

Weather predictions: look at sets of assignments. Scoring rules.

Littlewood (1953):

Mathematics (by which I shall mean pure mathematics) has no grip on the real world; if probability is to deal with the real world it must contain elements outside mathematics; the *meaning* of 'probability' must relate to the real world, and there must be one or more 'primitive' propositions about the real world, from which we can then proceed deductively (i.e. mathematically). We will suppose (as we may by lumping several primitive propositions together) that there is just one primitive proposition, the 'probability axiom,' and we will call it A for short. Although it has got to be *true*, A is by the nature of the case incapable of deductive proof, for the sufficient reason that it is about the real world ....

There are 2 schools. One, which I will call mathematical, stays inside mathematics, with results that I shall consider later. We will begin with the other school, which I will call philosophical. This attacks directly the 'real' probability problem; what are the axiom A and the meaning of 'probability' to be, and how can we justify A? It will be instructive to consider the attempt called the 'frequency theory'. It is natural to believe that if (with the natural reservations) an act like throwing a die is repeated n times the proportion of 6's will, *with certainty*, tend to a limit, p say, as $n \to \infty$. (Attempts are made to sublimate the limit into some Pickwickian sense—'limit' in inverted commas. But either you *mean* the ordinary limit, or else you have the problem of explaining how 'limit' behaves, and you are no further. You do not make an illegitimate

conception legitimate by putting it into inverted commas.) If we take this proposition as 'A' we can at least settle off-hand the other problem, of the *meaning* of probability; we define its measure for the event in question to be the number p. But for the rest this A takes us nowhere. Suppose we throw 1000 times and wish to know what to expect. Is 1000 large enough for the convergence to have got under way, and how far? A does not say. We have, then, to add to it something about the rate of convergence. Now an A cannot assert a *certainty* about a particular number n of throws, such as 'the proportion of 6's will *certainly* be within $p \pm \epsilon$ for large enough n (the largeness depending on ϵ)'. It can only say 'the proportion will lie between $p \pm \epsilon$ *with at least such and such probability (depending on* ϵ *and* n_0*) whenever* $n > n_0$'. The vicious circle is apparent. We have not merely failed to *justify* a workable A; we have failed even to *state* one which would work if its truth were granted. It is generally agreed that the frequency theory won't work. But whatever the theory it is clear that the vicious circle is very deep-seated: certainty being impossible, whatever A is made to state can be stated only in terms of 'probability'.

USGS 1999 Forecast

$$P(\text{M}\geq 6.7 \text{ event by } 2030) = 0.7 \pm 0.1$$

What does this mean?

Where does the number come from?

Two big stages.

Stage 1

1. Determine regional constraints on aggregate fault motions from geodetic measurements.

2. Map faults and fault segments; identify segments with slip rates $\geq$1 mm/y. Estimate the slip on each fault segment principally from paleoseismic data, occasionally augmented by geodetic and other data. Determine (by expert opinion) for each segment a 'slip factor,' the extent to which long-term slip on the segment is accommodated aseismically. Represent uncertainty in fault segment lengths, widths, and slip factors as independent Gaussian random variables with mean 0. Draw a set of fault segment dimensions and slip factors at random from that probability distribution.

3. Identify (by expert opinion) ways in which segments of each fault can rupture separately and together. Each combination of segments is a 'seismic source.'

4. Determine (by expert opinion) the extent to which long-term fault slip is accommodated by rupture of each combination of segments for each fault.

5. Choose at random (with probabilities of 0.2, 0.2, and 0.6) 1 of 3 generic relationships between fault area and moment release to

characterize magnitudes of events that each combination of fault segments supports. Represent the uncertainty in the generic relationship as Gaussian with zero mean and standard deviation 0.12, independent of fault area.

6. Using the chosen relationship and the assumed probability distribution for its parameters, determine a mean event magnitude for each seismic source by Monte Carlo.

7. Combine seismic sources along each fault 'to honor their relative likelihood as specified by the expert groups;' adjust relative frequencies of events on each source so that every fault segment matches its estimated geologic slip rate. Discard combinations of sources that violate a regional slip constraint.

8. Repeat until 2,000 regional models meet the slip constraint. Treat the 2,000 models as equally likely for estimating magnitudes, rates, and uncertainties.

9. Estimate the background rate of seismicity: Use an (unspecified) Bayesian procedure to categorize historical events from three catalogs either as associated or not associated with the seven fault systems. Fit generic Gutenberg-Richter magnitude-frequency relation $N(M) = 10^{a-bM}$ to the events deemed not to be associated with

the seven fault systems. Model background seismicity as a marked Poisson process. Extrapolate the Poisson model to $M \geq 6.7$, which gives a probability of 0.09 of at least one event.

Stage 1: Generate 2,000 models; estimate long-term seismicity rates as a function of magnitude for each seismic source.

Stage 2:

1. Fit 3 types of stochastic models for earthquake recurrence—Poisson, Brownian passage time (*Ellsworth et al.*, 1998), and ' time-predictable'—to the long-term seismicity rates estimated in stage 1.

2. Combine stochastic models to estimate the probability of a large earthquake.

Poisson and Brownian passage time models used to estimate the probability an earthquake will rupture each fault segment.

Some parameters fitted to data; some were set more arbitrarily. Aperiodicity (standard deviation of recurrence time, divided by expected recurrence time) set to three different values, 0.3, 0.5, and 0.7. Method needs estimated date of last rupture of each segment.

Model redistribution of stress by large earthquakes; predictions made w/ & w/o adjustments for stress redistribution.

Predictions for segments combined into predictions for each fault using expert opinion about the relative likelihoods of different rupture sources.

'Time-predictable model' (stress from tectonic loading needs to reach the level at which the segment ruptured in the previous event for the segment to initiate a new event) used to estimate the probability that an earthquake will originate on each fault segment. Estimating the state of stress before the last event requires date of the last event and slip during the last event. Those data are available only for the 1906 earthquake on the San Andreas Fault and the 1868 earthquake on the southern segment of the Hayward Fault. Time-predictable model could not be used for many Bay Area fault segments.

Need to know loading of the fault over time; relies on viscoelastic models of regional geological structure. Stress drops and loading rates modeled probabilistically; the form of the probability models not given. Loading of San Andreas fault by the 1989 Loma Prieta earthquake and the loading of Hayward fault by the 1906 earthquake were modeled.

The probabilities estimated using time-predictable model were converted into forecasts using expert opinion for relative likelihoods that an event initiating on one segment will stop or will propagate to other segments.

The outputs of the three types of stochastic models for each segment weighted using opinions of a panel of 15 experts. When results from the time-predictable model were not available, the weights on its output were 0.

So, what does it mean?

I have no idea. It's just a number.

None of the standard interpretations of probability applies.

Method has aspects of Fisher's fiducial inference, frequency theory, probability models, subjective probability.

Frequencies equated to probabilities; outcomes assumed to be equally likely; subjective probabilities used in ways that violate Bayes' Rule.

Calibrated using data that are not commensurable–global, or extrapolated across magnitude ranges using 'empirical' scaling laws.

Models upon models; ad hoc ad nauseum.

Inconsistent and virtually opaque.

Better to spend resources on preparedness, education, outreach.

Testing predictions

Some predictions hold "by chance."

Can't conclude a method has merit just because some predictions come true.

How to evaluate? Ideas from hypothesis testing.

Chance model for successful predictions: Does method succeed 'beyond chance?'

Null hypotheses for testing predictions:

- Poisson seismicity, historical rates; predictions fixed
- Poisson seismicity after 'declustering,' historical rates; predictions fixed
- Locations from catalogs, times uniform; predictions fixed
- Locations from catalogs, times permuted; predictions fixed

Methodological examples

Jackson, 1996

Tests deterministic predictions using a probability distribution for the number of successful predictions, derived from a null hypothesis that specifies chance each prediction succeeds.

Does not say how to find these probabilities, although says that usually the null hypothesis is that seismicity follows a Poisson process with rates equal to the historical rates.

Assumes that successes are independent.

Advocates estimating the P-value by simulating the distribution of the sum of independent Bernoulli variables.

Console, 2001

Rejects the null hypothesis if more events occur during alarms than are expected on the assumption that seismicity has a homogeneous Poisson distribution with true rate equal to the observed rate.

No discussion of significance level or power.

Shi, Liu & Zhang, 2001

Evaluated official Chinese earthquake predictions for magnitude 5 and above, 1990–1998.

Divided study region into 3,743 small cells in space, and years of time.

In a given cell in a given year, either an earthquake is predicted to occur, or—if no—that's a prediction that there will be no event in that cell during that year.

Test statistic is R-score:

$$R = \frac{\#\text{ cells in which earthquakes are successfully predicted}}{\#\text{ cells in which earthquakes occur}} - \frac{\#\text{ cells with false alarms}}{\#\text{ aseismic cells}}, \quad (1)$$

Compare the R-score of the actual predictions on the declustered catalog with the R-score of 3 sets of random predictions:

1. Condition on the number of cells in which earthquakes are predicted to occur. Choose that many cells at random without replacement from the 3,743 cells, with the same chance of selecting each cell; predict that earthquakes of magnitude 5 or above will occur in those randomly-selected cells.

2. For the jth cell, toss a p_j-coin, where p_j is proportional to the historical rate of seismicity in that cell. If the jth coin lands heads, predict that an earthquake of magnitude 5 or above will occur in the jth cell. Toss coins independently for all cells, $j = 1, \ldots, 3,743$. Yields a random number

of predictions, with predictions more likely in cells where more events occurred in the past.

3. Condition on the number of cells in which earthquakes are predicted to occur. Choose that many cells at random without replacement from the 3,743 cells. Select the jth cell with probability p_j, with p_j set as in (2). Predict that earthquakes of magnitude 5 or above will occur in the selected cells.

None of 3 methods depends on the observed seismicity during the study period, 1990–1998.

Claims of successful predictions

Varotsos, Alexopoulos and Nomicos (VAN)

Literature debate: v. 23 of *Geophysical Research Letters*, 1996.

Participants did not even agree about the number of earthquakes that were predicted successfully, much less whether the number of successes was surprising. Participants disagreed about whether the predictions were too vague to be considered predictions, whether some aspects of the predictions were adjusted *post hoc*, what the null hypothesis should be, and what tests were appropriate.

Wyss and Burford, 1987

Predicted $M_L = 4.6$ earthquake of 31 May 1986 near Stone Canyon, California, ≈1y before it occurred.

Examined the rates of earthquakes on different sections of the San Andreas fault. Identified 2 fault sections in which the rate dropped compared with the rates in neighboring sections.

Say "the probability [of the prediction] to have come true by chance is $< 5\%$."

That's the chance that an earthquake would occur in the alarm region, if earthquakes occurred at random, independently, uniformly in space and time, with rate equal to the historic rate in the study area over the previous decade. Thus, null hypothesis is that seismicity follows a homogeneous Poisson process with rate equal to the historical rate; clustering is not taken into account.

Kossobokov et al., 1999

Claim to have predicted four of the five magnitude 8 and larger earthquakes that occurred in the circum-Pacific region between 1992 and 1997. "[t]he statistical significance of the achieved results is beyond 99%."

Predictions based on pattern recognition.

Calculate statistical significance by assuming that earthquakes follow a Poisson process: homogeneous in time, heterogeneous in space. Intensity estimated from historical data. Condition on number of events in the study area, so locations and times are iid across events, the epicenters and times are independent of each other, the temporal density of earthquake times is uniform, and the spatial distribution of epicenters is given by the historical distribution between 1992 and 1997.

Calculation does not take clustering into account, and conditions on the predictions. Treat successes as independent with probability p equal to the normalized measure of the union of the alarms. Measure is product of the uniform measure on time and counting measure on space, using historical distribution of epicenters in the study volume.

Analogy to weather prediction:

- Predictions depend on weather history
- "If rain today, predict rain tomorrow" works well
- Most schemes lose dependence of predictions on history
- Don't account for clustering

"If earthquake today, predict earthquake tomorrow" works well, too.

year	M_τ	events	succ	succ w/o	max sim	P-value (est)	v
2004	5.5	445	95	30	28	< 0.001	3.9×10^{-4}
2004	5.8	207	24	7	10	0.041	1.8×10^{-4}
2000-2004	5.5	2013	320	85	48	< 0.001	3.6×10^{-4}
2000-2004	5.8	996	114	29	19	< 0.001	1.8×10^{-4}

Simulations using Harvard CMT catalog. Col. 4: Events with magnitude at least M_τ that are within 21 days following and within 50 km of the epicenter of an event with magnitude M_τ or greater. Col. 5: Events within 21 days following and within 50 km of the epicenter of an event whose magnitude is at least M_τ but no greater than that of the event in question. Events that follow within 21 days of a larger event are not counted. Col. 6, 'max sim,' is the largest number of successful predictions in 1,000 random permutations of the times of the events Harvard CMT catalog, holding the alarms and the locations and magnitudes of events in the catalog fixed. Col. 7: fraction of permutations in which the number of successful predictions was $\geq$ observed number. Col. 8: upper bound on fraction of study region (in space and time) covered by alarms.

Method succeeds far beyond chance. Why?

Null hypothesis does not model the dependence of predictions on seismicity.

That dependence, plus clustering, gives 'surprising' success rates.