

Testing Skill in Earthquake Predictions

Richard H. Jones

Department of Preventive Medicine and Biometrics

School of Medicine, Box B-119

University of Colorado Health Sciences Center

Denver, Colorado 80262

Richard.Jones@uchsc.edu

Alan L. Jones

Department of Geological Sciences and Environmental Studies

State University of New York at Binghamton

jones@sunquakes.geol.binghamton.edu

June 6, 1996

Abstract

A statistical method is presented for testing whether forecasts of events which have known probabilities show skill. When an earthquake forecast is made for a given magnitude and time and space window, a prior probability of an earthquake occurring is estimated by searching an earthquake catalog. A scoring method based on information theory is developed to test whether a number of forecasts show forecasting skill. The forecaster may or may not have knowledge of the prior probabilities. A method is developed for calculating exact p-values for a series of predictions. The method is tested on both simulated data, and on a series of actual forecasts.

1 Introduction

When evaluating the success of a method to predict earthquakes one should always consider the prior probability that an event will happen in the prediction time, space, and magnitude windows based on past catalogs. The sci.geo.earthquakes newsgroup on the Internet is a location where many people place predictions. However, few of these state clear windows for the predictions, but some do. An experiment was run to compare the success of people making predictions against the null hypothesis that the predictions are no better than chance given the prior probabilities. If the probability of success of each prediction is the same, one can determine if a statistically significant result has been obtained by using the binomial distribution. When the probability of success for the predictions vary, it is not straight forward to evaluate if a statistically significant result has been obtained. We present a method for scoring a predictor to test if skill is involved.

2 The Experiment

Some of the people on the Internet newsgroup sci.geo.earthquakes making predictions declared their success by taking the ratio of successful predictions to the total number of predictions. Those contributing to the newsgroup complained that many of the predictions

were for very likely events such as a magnitude 3.0 earthquake in the Cape Mendocino region of California within a given week. However, no one declared just how likely this was so the predictions continued and the claims of great success continued. In addition, after the time window closed, the predictor often claimed success because an event was “close” to one or more of the windows. One of us (ALJ) decided that this would be a useful forum to demonstrate the null hypothesis referring to the work of Jackson (1994) and Richardson (1994). Whenever a clearly-defined prediction was made, a follow-up posting to the newsgroup would be placed declaring what the probability of success was by chance due to the historical seismicity for that region. After the time window closed, the reports from the National Earthquake Information Center (NEIC) and other providers of hypocenter information would be consulted to see if the prediction was successful.

3 Method to Determine the Prior Probabilities

When the experiments began, the probability of success was computed very simply: the total number of events in the space-magnitude window for the period 1960 to the present was determined. This number was divided by the total number of days since 1960 to the present to get the expected number of events per day. This was then multiplied by the number of days in the prediction to obtain the expected number of events. A Poisson distribution for the events was assumed and the probability of at least one event was computed as

$$p = 1 - e^{-\mu},$$

where μ is the expected number of events in the time window.

When this method was used and documented on the newsgroup, a participant on the newsgroup objected that the Poisson distribution is not a good assumption for earthquakes since there is significant cluster due to after shocks. A new method was developed. In this method, all of the similar time periods in the catalog were scanned to see how many had at least one event. An example, will help to clarify this method. Say the prediction is for a 10-day period. The catalogs were scanned for the occurrence of at least one event in the time period from 1960/01/01 through 1960/01/10 and a hit recorded if there was. This was repeated for the period 1960/01/11 through 1960/01/20 and so on. The ratio of periods with at least one event was divided by the total number of such periods in the time period 1960 to the present. For a 10-day prediction window, there are over 1200 such time periods. Figure 1 shows the probability of at least one earthquake as a function of the length of the prediction time window calculated using the Poisson assumption and the second method referred to as the cluster method. The Poisson method over-predicts the probability of an earthquake in a time interval.

Insert Figure 1 about here

This second method was in place a short while before someone made a prediction of an event off of Cape Mendocino only a few days after an event of magnitude 6.5. When the probability of success of 68% was posted to the newsgroup, A. Michael (personal communication) stated that the probability was closer to 100% due to after shocks of the event. He suggested that the Rosenberg-Jones (1994) formula be used to factor in the probability of an after shock. This was done and that particular prediction was found to

have a probability of 99% of at least one event. The R-J formula predicted that there would be 4 after shocks in the time-space-magnitude window. Two occurred.

4 Catalogs Used

The catalogs used were obtained from the NEIC Epic CD-ROM for the period 1960 through 1992. For more recent events, hypocenters were obtained from the University of California at Berkeley for northern California, Caltech for southern California, and the NEIC for world-wide events and U.S.A events (see Table 1). The catalogs are considered complete to the magnitude values given.

Catalog	Dates	Magnitudes
World	1960-1995	≥ 5.0
U.S.A.	1960-1995	≥ 4.0
California	1960-1995	≥ 3.0

Table 1: Catalogs

5 Some Actual Predictions

In 1994 DDG began making predictions on the newsgroup. His predictions were followed for this experiment starting in February 1995. His predictions were of the form of a clearly stated time window and a number of rings centered on his home in southern California. A typical prediction was of the form:

EQ Prediction for 96/01/20 UTC:

Time window

UTC Time: 2300 HOURS 20 JAN 1996 THRU 1700 HOURS 31 JAN 1996

PST Time: 1500 HOURS 20 JAN 1996 THRU 0900 HOURS 31 JAN 1996

Magnitude-vs-Range-Envelope:

```

From => 0 and < 76 miles => 4.5
From => 76 and < 151 miles => 4.8
From => 151 and < 251 miles => 5.1
From => 251 and < 451 miles => 5.4
From => 451 and < 701 miles => 5.7
From => 701 and < 1226 miles => 6.0
From => 1226 and < 1926 miles => 6.3
From => 1926 and < 3326 miles => 6.6
From => 3326 and < 5526 miles => 6.9
From => 5526 and < 9526 miles => 7.2
          < less than
          => equal to or greater

```

Ranges are from Saugus, California located at:

Latitude 34o 26' 20"N Longitude 118o 31' 22"W
 34.43877N 118.52272W

In the earlier predictions, his magnitude windows were closed. That is, for each ring he stated a minimum and a maximum magnitude. Therefore, if an event occurred in a window which was larger than his maximum, it could not be declared a hit. Later in the experiment he changed to open-ended magnitude windows such as in this prediction.

Some of his predictions had 10 rings. Others had only a few of the inner rings. The radii of the rings did not vary from prediction to prediction. When all 10 rings were used, the total area encompassed by the rings covered 87% of the earth's surface.

DDG did not, and as of this writing, has not revealed his method of prediction but only to say it is mostly electro-magnetic in nature. In addition, he has some un-specified "indicators." For this experiment, we were not concerned with his method but were more interested in developing a method for evaluating predictions.

DDG's first prediction failed but he followed this failure with 12 straight successes in the time period 1995/02/21 through 1995/11/23. Even though this seemed remarkable, we still needed a method to evaluate if this was statistically significant. Table 2 shows a summary of his predictions along with the *a priori* probabilities. Note that the time intervals for three of these predictions overlap the time interval for the previous prediction. This violates the assumption of statistical independence of the predictions, but will be ignored for this example.

Dates	Prob	Success?
1995/02/21 - 03/02	80%	No
1995/03/07 - 03/17	80%	Yes
1995/04/04 - 04/14	50%	Yes
1995/04/09 - 04/19	66%	Yes
1995/04/24 - 05/01	90%	Yes
1995/06/06 - 06/14	58%	Yes
1995/06/20 - 06/26	77%	Yes
1995/06/23 - 06/30	45%	Yes
1995/08/29 - 09/01	60%	Yes
1995/08/29 - 09/07	90%	Yes
1995/09/26 - 10/03	70%	Yes
1995/10/07 - 10/14	57%	Yes
1995/11/15 - 11/23	63%	Yes
1996/01/09 - 01/17	64%	No
1996/01/17 - 01/19	17%	No
1996/01/20 - 01/31	47%	No
1996/02/02 - 02/04	3%	No

Table 2: DDG's predictions

6 An Information Score

A method for scoring probability forecasts of weather events based on information theory was given by *Brelsford and Jones* [1967]. In this problem, the forecaster predicts the probability of an event such as rain the next day, and the next day observes whether it rained or not. The forecaster gets a loss of $-\ln p$, where p is the predicted probability of the event that occurred. If the forecast was a probability of rain of .6, the loss would be $-\ln .6$ if it rains and $-\ln .4$ if it does not. These scores are summed across forecasts, and the forecaster with the highest score loses. The score is based on the $-\log$ likelihood function.

This is the same problem considered by *Kagan and Jackson* [1995] with a sign change since they use the log likelihood function, so a higher score is better. Using the notation of Kagan and Jackson, for prediction i , let c_i be 1 if an earthquake occurs, and 0 if an earthquake does not occur, and let p_i be the predicted probability of an earthquake. Kagan and Jackson's L test has the score

$$L = \sum_{i=1}^n c_i \ln(p_i) + \sum_{i=1}^n (1 - c_i) \ln(1 - p_i).$$

These likelihood based scores are useful for comparing scores of different forecasters, or, in the case considered by Kagan and Jackson, two sets of probability forecasts generated by different theories.

Predicting an earthquake as a yes/no event is a similar but different problem. Here we are interested in testing whether the yes/no forecasts show skill. In this problem, there are four possibilities. The prediction can be that the event will occur or will not occur, and the result is that the event does or does not occur. If the prior probability of an event is p , and the prediction is that the event will occur, the score for a correct forecast is $-\ln p$, a positive number. If the forecast is wrong, the score is $\ln(1 - p)$, a negative number. If the forecast is that the event will not occur, the score for a correct forecast is $-\ln(1 - p)$, and the score for an incorrect forecast is $\ln(p)$. If there is no skill, and the forecast is that the event will occur, the expected value of this score is

$$-p \ln p + (1 - p) \ln(1 - p).$$

If the forecast is that the event will not occur, the expected value of this score is

$$-(1 - p) \ln(1 - p) + p \ln p.$$

Subtracting the expected score from the score gives a score whose mean value is zero if there is no skill. These scores are shown in Table 3. The variance of each score is

$$p(1 - p)[\ln(p(1 - p))]^2.$$

Assuming forecasts are independent, the score and variance are summed over forecasts and divided by the square root of the summed variances. Because of the central limit theorem, under certain conditions, the result should be a standard normal distribution if there is no skill.

		Outcome	
		1	0
Prediction	1	$-(1-p)\ln[p(1-p)]$	$p\ln[p(1-p)]$
	0	$(1-p)\ln[p(1-p)]$	$-p\ln[p(1-p)]$

Table 3: Scores for Predictions

7 Calculating exact p-values

Because a fairly large number of trials may be necessary before the asymptotic normal distribution of the information score holds, an exact method was derived to calculate the p-value. For a series of forecasts, prior probabilities and actual outcomes, the information score is calculated. Holding the forecasts and prior probabilities fixed, all possible outcomes are enumerated. If there are n forecasts, there are 2^n possible outcomes. For every outcome, the information score is calculated. If this score is greater than or equal to the actual score for the actual outcomes, the probability of this particular outcome is calculated and summed. This calculates the probability that a score as large or larger than the actual score could have occurred by chance. The probability of a given outcome is the product of the probabilities of each of the events in that outcome. A single probability will be p if the outcome is 1 and $1-p$ if the outcome is 0. It is practical to calculate these exact p-values for up to about 25 forecasts.

8 Simulations

Several simulations were carried out to compare the p-values for the exact method with the p-values from the asymptotic method based on the normal distribution as a function of the number of predictions. Two ‘no skill’ methods were simulated. The first assumed that the forecaster had access to the prior probabilities and forecast the more probable event. The second ‘no skill’ method assumed that the forecaster did not have prior knowledge and just flipped a coin. The results were the same for both these simulations.

For the prior probabilities, a random number was drawn from a uniform distribution with the range .01 to .99. This was to avoid the end points of 0 and 1 where a score of 0 is given for a correct forecast and a score of $-\infty$ is given for a wrong forecast. The outcome was simulated with the probability of an earthquake being this prior probability. The number of forecasts by a forecaster was varied from 1 to 15. For each number of forecasts, 10,000 simulations were carried out, and the difference between the exact p-value and asymptotic p-value calculated. Since most interest is in small p-values, simulations with an asymptotic p-value above 0.1 were discarded. This is because large p-values are not usually looked at very seriously. There are approximately 1000 simulations kept out of 10,000 for each n . The results are shown in Table 4. Note that as the number of predictions increases, the absolute differences between the p-value for the asymptotic method and the exact p-value decreases to 0.0082 with the positive sign indicating that the approximate method gives a higher p-value on average. This makes the approximate method based on

n	Bias	s.d.
5	-0.0904	0.0207
6	-0.0730	0.0222
7	-0.0579	0.0230
8	-0.0435	0.0264
9	-0.0288	0.0254
10	-0.0156	0.0224
11	-0.0065	0.0197
12	-0.0012	0.0187
13	0.0011	0.0160
14	0.0037	0.0144
15	0.0052	0.0129
16	0.0066	0.0121
17	0.0074	0.0116
18	0.0076	0.0110
19	0.0082	0.0099
20	0.0082	0.0092

Table 4: 10,000 simulations for each value of n , the number of forecasts. Only forecasts with an approximate p-value of 0.1 or less are kept. The ‘Bias’ is the asymptotic p-value minus the exact p-value averaged over the 10,000 simulations. ‘sd’ is the standard deviation of these differences

an asymptotic normal distribution conservative, giving, on average, a higher p-value. This also indicates that it is best to use the exact method for up to 20 or 25 forecasts. The standard deviation of 0.0092 indicates that the variation about this bias is about ± 0.018 (\pm two standard deviations).

Table 5 shows DDG’s predictions and the exact p-value for all the predictions up to prediction n . With a series of correct predictions, the p-value falls until it reaches a p-value of 0.04, and then, with incorrect forecasts the p-values increase again. Since these are multiple tests that are correlated, this example is inconclusive as to whether these forecasts show skill.

9 Discussion and Conclusions

In this paper, a method has been developed to test whether a series of earthquake forecasts shows skill when the prior probabilities of an earthquake for a given time, space and magnitude window are known. The forecasts can be ‘yes, an earthquake will occur’, or ‘no, an earthquake will not occur’. The score is based on information theory, and is independent of whether the forecaster knows the prior probabilities or not. If the forecaster has no skill and just flips a coin, he/she will occasionally hit a low probability event correctly giving a large score. If the forecaster knows the prior probabilities, but has no skill and just forecasts the more probable event he/she will never get a large value of the score as when a rare event is correctly forecast. These two no skill forecasts average out and do not affect the null hypothesis of the test that the forecaster has no skill. If the forecaster does show skill, the method should produce a small p-value.

A FORTRAN subroutine for calculating the exact p-value given a series of n fore-

n	p	pred	outcome	exact p-value
1	0.80	1	0	1.0000
2	0.80	1	1	0.9600
3	0.50	1	1	0.8000
4	0.66	1	1	0.6368
5	0.90	1	1	0.5731
6	0.58	1	1	0.4122
7	0.77	1	1	0.3428
8	0.45	1	1	0.2009
9	0.60	1	1	0.1358
10	0.90	1	1	0.1223
11	0.70	1	1	0.0918
12	0.57	1	1	0.0585
13	0.63	1	1	0.0399
14	0.64	1	0	0.1044
15	0.17	1	0	0.1326
16	0.47	1	0	0.2035
17	0.03	1	0	0.2164

Table 5: DDG's predictions and the exact p-value for the first n predictions

casts with the corresponding prior probabilities and the actual outcomes is given in the Appendix.

Acknowledgment

The first author is partially supported by the Geophysical Statistics Project, National Center for Atmospheric Research, Boulder, Colorado (<http://www.cgd.ucar.edu/stats>)

Appendix: FORTRAN Code for Exact p-value

```

subroutine exact(n,p,pred,y,score,prob)
real p(n),q(50)
integer pred(n),binary(50)
c
c subroutine to calculate the exact p-value for a series
c of zero-one forecasts.
c input to subroutine
c n is the number of forecasts
c p is a vector length n of the prior probabilities
c pred is an integer array of predictions, 0 or 1.
c y is an integer array of actual outcomes, 0 or 1.
c output by subroutine
c score is calculated by subroutine test
c prob is the exact p-value
c
call test(n,p,pred,y,score)

```

```

do 60 i=1,n
  binary(i)=0
  q(i)=1.0-p(i)
60 continue
prob=0.0
iimax=2**n
do 110 ii=1,iimax
  call test(n,p,pred,binary,s)
  if(s.ge.score-.000001)then
c
c calculate probabilities
c
      pb=1.0
      do 80 i=1,n
        if(binary(i).eq.1)then
          pb=pb*p(i)
        else
          pb=pb*q(i)
        endif
80      continue
      prob=prob+pb
    endif
c
c add one to binary number
c
      if(ii.lt.iimax)then
        do 90 i=1,n
          if(binary(i).eq.0)then
            binary(i)=1
            goto 100
          else
            binary(i)=0
          endif
90      continue
100     continue
      endif
110 continue
return
end
c
subroutine test(n,p,pred,y,score)
real p(n)
integer pred(n),y(n)
c
c subroutine to calculate the information score for a series
c of zero-one forecasts with known prior probabilities.
c output from subroutine

```

```

c   score is the information score
c
  var=0.0
  score=0.0
  do 50 i=1,n
    q=1.0-p(i)
    fac=log(p(i)*q)
    if(y(i).eq.1)then
      if(pred(i).eq.1)score=score-q*fac
      if(pred(i).eq.0)score=score+q*fac
    endif
    if(y(i).eq.0)then
      if(pred(i).eq.1)score=score+p(i)*fac
      if(pred(i).eq.0)score=score-p(i)*fac
    endif
    var=var+p(i)*q*fac**2
  50 continue
  sd=sqrt(var)
  score=score/sd
  return
  end

```

References

- Brelsford, W. M. and R. H. Jones, Estimating Probabilities, *Monthly Weather Review* 95, 570-576, 1967.
- Jackson, D. D., Earthquake Prediction Methods, Paper S31B-1, AGU Fall Meeting, Dec. 5-9, 1994.
- Kagan, Y. Y. and D. D Jackson, New seismic gap hypothesis: Five years after, *J. Geophys. Res.*, 100, B3, 3943-3959, 1995.
- Richardson, R. M., The Role of the Null Hypothesis in Earthquake Prediction, Paper S31B-2, AGU Fall Meeting, Dec. 5-9, 1994.
- Rosenberg, P. A. and L. M. Jones, Earthquake Aftershocks: Update, *Science*, 265, Aug. 26, 1994.

Figure Caption

Figure 1. Comparison of Poisson assumption and the cluster method of estimating the probability of an earthquake in a given time and space window. The cluster method scans the catalog for the given space window and the same time window as the prediction to estimate the probability of an earthquake.