Comparison of two earthquake predictability evaluation approaches: Molchan error trajectory and likelihood

Jeremy Zechar[1], Thomas Jordan[1,2], Danijel Schorlemmer[1], Maria Liuikis[2]

1. Department of Earth Sciences, University of Southern California, Los Angeles, CA 90089
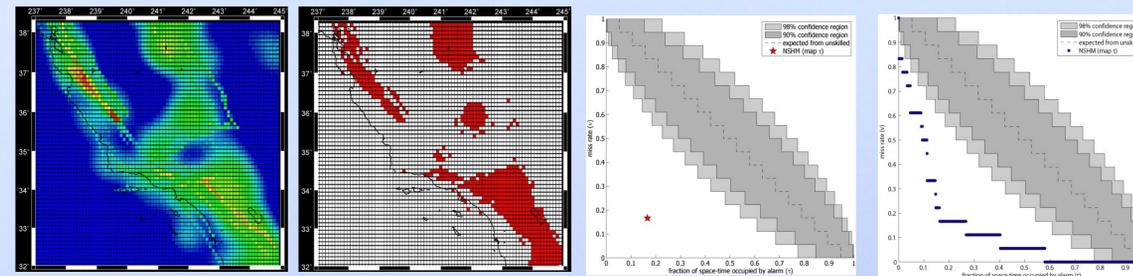2. Southern California Earthquake Center, Los Angeles, CA 90089

Abstract

The Regional Earthquake Likelihood Models (RELM) working group has begun a collaborative earthquake predictability experiment involving a dozen five-year forecasts of earthquake occurrence in a California natural laboratory. The forecasts are probabilistic in the sense that they consist in expected number of earthquakes in space-time-magnitude bins. Statistical hypothesis testing of the forecasts is achieved via three scores based on likelihood. Earthquake forecasts that do not adhere to the RELM template cannot at this time be accommodated.

In order for the Collaboratory for the Study of Earthquake Predictability (CSEP) to succeed, it is desirable to expand the scope of predictability experiments; in doing so, additional evaluation techniques must be considered. We explore a score based on the Molchan error diagram, which plots miss rate versus the fraction of space occupied by alarms, and is commonly used to assess the skill of earthquake prediction methods using a single alarm set (i.e., one point on the error diagram). We supplement the point wise approach with a cumulative performance measure based on the normalized area under an error trajectory. We call this the area skill score; a score of unity indicates perfect skill and a score of zero indicates perfect non-skill.

Both the RELM and the Molchan error trajectory techniques involve three-step hypothesis testing and both can be applied to the five-year forecasts of California seismicity. We compare the two methods both conceptually and practically – that is, by examining the results of their application to over a dozen five-year forecasts of California seismicity.

RELM Likelihood testing methodology

The RELM likelihood tests require forecasts expressed in terms of the number of expected earthquakes in latitude/longitude/magnitude bins. Upon completion of the prediction experiment, the catalog of observed earthquakes is similarly expressed. The observations and the expectations are compared via three tests based on the joint log likelihood: the L-test, the N-test, and the R-test. Each of these tests compares a measure based on the observed seismicity with the distribution of that measure assuming a forecast is a "correct" model. In order to obtain these distributions, for each forecast, we generate a set of synthetic earthquake catalogs consistent with the forecast. That is, we take the distribution of seismicity implied by a forecast and create synthetic catalogs consistent with this distribution. For the L- and R-tests, the joint log-likelihood is defined:

$$L = \sum_{i=1}^{n} - \lambda_i + \omega_i \log \lambda_i - \log \omega_i!$$

where n is the number of bins, λ_i is the forecast number of earthquakes in the ith bin and ω_i is the observed number of earthquakes in the ith bin.

In the L-test, the likelihood of a given forecast based on the observed seismicity is compared to the distribution based on synthetic catalogs. If the observed likelihood is extremely low relative to the distribution, then the observed seismicity is not consistent with the forecast and the forecast can be rejected. A quantile measure γ is defined as the fraction of synthetic catalogs for which a forecast obtains a likelihood less than the likelihood of the forecast conditional on the observed seismicity. A very low γ indicates data inconsistency; in fact, γ is a significance value. **A low value of γ allows us to reject a forecast.**

In the N-test, the total observed number of earthquakes is compared to the distribution of the total number of earthquakes in the synthetic catalogs. If the observed number of earthquakes is extreme compared to this distribution, then the observed seismicity is not consistent with the forecast and the forecast can be rejected. A quantile measure δ is chosen as the statistic upon which to base the hypothesis test. The measure δ is defined as the fraction of synthetic catalogs containing fewer earthquakes than the observed catalog. In this case, a low value of δ indicates that a forecast predicts more earthquakes than are observed – *overpredicting* – and a high value of δ indicates that a forecast predicts fewer earthquakes than are observed – *underpredicting*. **High and low values of δ allow us to reject a forecast.**

In the R-test, we consider forecasts in pairs. For a given pair and the observed seismicity, we compute the likelihood ratio, or the difference between the joint log likelihoods of the forecasts. We then compare this difference to a distribution of differences using the synthetic catalogs corresponding to one of the forecasts. For example, if we have a pair of forecasts A and B, we compute the joint log-likelihood of each L_A and L_B conditional on the observed seismicity and subtract the former from the latter and call the resulting difference $R_{B,A}$. We then repeat this process for each of the synthetic catalogs based on forecast B. If the observed likelihood ratio is extremely small relative to this distribution, forecast B can be rejected. A quantile measure α^{BA} is defined as the fraction of synthetic catalogs for which the likelihood ratio is less than the likelihood ratio conditional on the observed seismicity. **High values of α^{BA} support forecast B, while low values of α^{BA} allow us to reject forecast B.**

Figure. Conceptual process of likelihood tests: forecast in latitude/longitude bins is compared with binned observed seismicity to make a measure of L, N, and R. The forecast is also used to generate synthetic catalogs, then compared to these catalogs to generate a distribution of L, N, and R assuming the forecast is true. From this distribution and the recorded observation, γ, δ, and α are computed.

Figure. 1st frame is an example alarm function where "warmer" colors indicate higher alarm function values. 2nd frame illustrates an example derived alarm set; here red regions indicate alarms. 3rd frame shows an example (τ, ν) point corresponding to the alarm set from the previous figure. 4th frame shows an example error trajectory corresponding to the alarm function in the first frame.
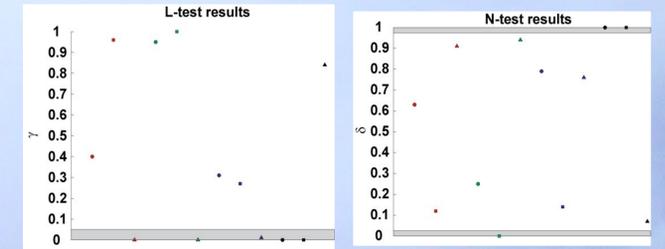
Molchan error trajectory testing

Molchan trajectory tests require only the specification of an alarm function, a function that yields alarm-based forecasts. The value of an alarm function at a given point represents the relative plausibility of a target earthquake at this point. By forecasting the number of earthquakes in geographic and magnitude cells for a fixed time period, each RELM forecast specifies an alarm function of latitude, longitude, and magnitude. Forecasts can be derived from an alarm function very simply: one can choose a single decision threshold and declare an alarm in any region where the alarm function value exceeds the threshold – this yields one set of alarms. To declare an alarm is to predict one or more target earthquakes within the specified space-time domain of the alarm. The issuance of alarms yields the following four scenarios:

- *Hit*: An earthquake occurs within the latitude/longitude/magnitude domain of an alarm.
- *False alarm*: An alarm is declared and no corresponding earthquakes occur.
- *Miss*: A target earthquake occurs at a latitude/longitude/magnitude point where no alarm is declared.
- *Correct negative*: In a given latitude/longitude/magnitude range, no alarm is declared and no earthquakes occur.

For a given alarm set and a given set of observed seismicity, we can compute the frequency of hits relative to the frequency of target earthquakes – the *hit rate* – and the *fraction of space occupied by alarms,* denoted τ. If we repeat this for many alarm sets derived from the same alarm function, we obtain a *Molchan trajectory*: the collection of all (τ, ν) points obtained for a given alarm function. The Molchan trajectory characterizes the predictive skill of the alarm function at all thresholds. If the area under the Molchan trajectory is high, this indicates high performance. We quantify this via the Area Skill Score (ASS), which is the area under the trajectory, subtracted from unity.

The key to Molchan trajectory testing is in selecting the measure of space used to compute τ. If physical space is used to compute τ, the corresponding cost map is uniform; this represents the hypothesis that target earthquakes are equally likely everywhere in the experiment space. In some sense, the measure of space is represented by an arbitrary cost map, and this cost map represents the null hypothesis.

In the case of a uniform cost map, we are testing against the hypothesis that seismicity is equally likely everywhere. As this is generally not the case in the Earth, any model with predictive skill should be able to reject the uniform hypothesis. In the uniform test, the ASS for the observed seismicity (with a uniform cost map) is compared to the distribution of ASS of randomized alarm functions. If the observed ASS is not extremely different from the distribution under randomized alarm functions, then the forecast can be rejected. **In the uniform test, an intermediate ASS allows us to reject the forecast model.**

In the case of using an arbitrary alarm function as the cost map, we are testing the hypothesis that this alarm function is consistent with the observed seismicity. In other words, does this alarm function represent a reasonable approximation of seismic distribution? If the alarm function is a good cost map, when we use the same alarm function to generate forecasts, the resultant ASS distribution will be centered about ½ and be approximately Gaussian. **In the self test, high and low values of ASS allow us to reject a forecast.**

The third Molchan trajectory test is an iteration of the self-test: we evaluate each alarm function's utility as a cost map and characterize this utility by the resultant ASSes. That is, we choose one of the forecasts as the cost map and compute the ASSes for each other forecast. If the alarm function is a good cost map, regardless of the alarm function used to generate forecasts, the resultant ASS distribution will be centered about ½ and be approximately Gaussian. **In the round-robin test, high and low values of ASS allow us to reject the forecast being used as the cost map.**

Figure. Natural laboratory used in this study.

Experiment specification

We consider M >= 4.95 earthquakes occurring between 2001 and 2005, inclusive, in the California natural laboratory depicted in the figure to the left. During the experiment, 24 earthquakes were recorded. This experiment primarily was designed to compare the two testing approaches over a five-year period and thus we chose a time period that allowed an ample number of samples.

Forecasts considered

We selected eleven of the forecasts submitted to the RELM group and one additional forecast based solely on the historical density of seismicity within the RELM bins. These models are listed in the **Table** to the right.

Forecast	Reference	Symbol
Asperity Likelihood	Schorlemmer and Wiemer 2007	● (red)
Combo 8.1	Ward, *SRL* 2007	■ (red)
Hidden Markov Model	Ebel, Chambers, Kafka, and Baglivo, *SRL* 2007	▲ (red)
HMM + Aftershocks	Ebel, Chambers, Kafka, and Baglivo, *SRL* 2007	● (green)
Geodetic 8.1	Ward, *SRL* 2007	■ (green)
Geologic 8.1	Ward, *SRL* 2007	▲ (green)
HKJ Smoothing	Helmstetter, Kagan, and Jackson, *SRL* 2007	● (blue)
Pattern Informatics	Holliday, Chen, Tiampo, Rundle, Turcotte, and Donnellan, *SRL* 2007	■ (blue)
Seismic 8.1	Ward, *SRL* 2007	▲ (blue)
Shen Geodetic	Shen, Jackson, and Kagan, 2007	● (black)
Ward Simulator	Ward, *SRL* 2007	■ (black)
Zechar RI/GR	This study	▲ (black)

Table. List of forecast models considered for this experiment. Shapes in the third column provide a legend for all test results figures.

Figure. 1st and 2nd frames show L-test and N-test results, respectively for the twelve models under consideration. Shaded area is the 5% critical region. Points falling in this region indicate that the corresponding forecast can be rejected with at least 95% confidence.
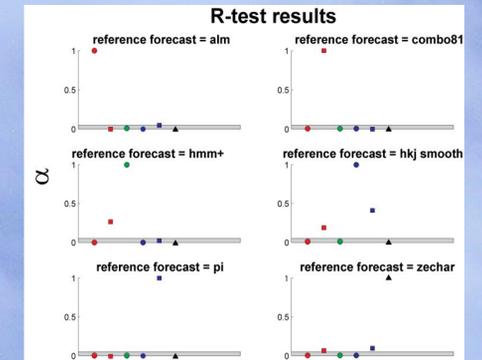
R-test results

Figure. R-tests are shown, taking as the reference model, or null hypothesis, each of the six forecasts passing the L- and N-tests. Shaded area is the 5% critical region. Points falling in this region indicate that the corresponding forecast can be chosen in favor of the reference forecast with at least 95% confidence -- that is, the reference forecast can be rejected.

Discussion

The figures above show the results of RELM likelihood tests. The L-test results indicate that the Hidden Markov Model, the Geologic 8.1, the Seismic 8.1, the Shen Geodetic, and the Ward Simulator forecasts can be rejected. The N-test results indicate that the Geodetic 8.1, Shen Geodetic, and Ward Simulator forecasts can be rejected. The R-test results give the interesting conclusion that all of the remaining forecasts can be rejected by at least one of the others with high confidence.

The figures below show the results of the Molchan trajectory tests. The uniform test results do not reject any of the forecasts; this is also the weakest test. The self-test results indicate that the HMM + Aftershocks, Geologic 8.1, Pattern Informatics, and Ward Simulator forecasts should be rejected. The round-robin test results indicate that the Combo 8.1, Hidden Markov Model, Geodetic 8.1, Seismic 8.1, and Shen Geodetic forecasts should be rejected, leaving only 3 of the original 12 as viable models.
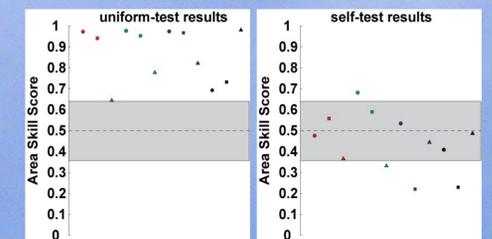
Figure. 1st frame shows uniform-test results for the twelve models under consideration. Shaded area is the 95% confidence region for random predictors based on the uniform distribution. In this case, any point falling in the shaded region does not reject a uniform null hypothesis and should itself be rejected. 2nd frame shows self-test results. In this case, any point falling outside the shaded region indicates that the corresponding forecast should be rejected.
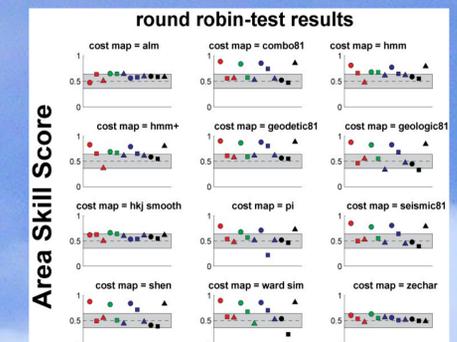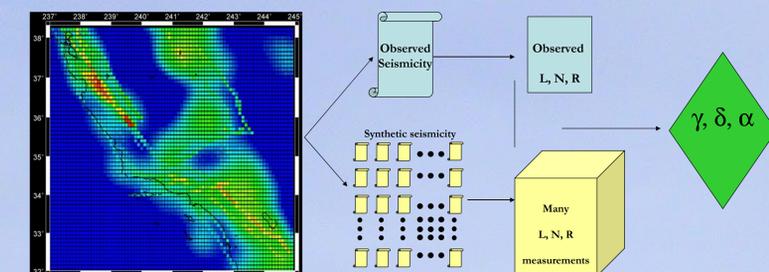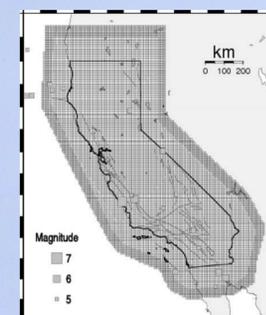
round robin-test results

Figure. Round-robin tests are shown, taking each of the dozen models as the cost map. In this case, if any point falls outside the shaded region, the forecast used as the cost map should be rejected.