

*Likelihood-based tests for evaluating
earthquake forecasts:
statistical power and result stability*

J.D. Zechar

ETH Zurich & Lamont-Doherty Earth Observatory

M.C. Gerstenberger & D.A. Rhoades

GNS Science

J.D. Zechar, M.C. Gerstenberger, & D.A. Rhoades, 2010 (in press). Likelihood-based tests for evaluating space-rate-magnitude earthquake forecasts, *Bulletin of the Seismological Society of America*, 100, doi:10.1785/0120090192.

Discussed:

Forecast format including assumptions,

Original RELM tests,

Correction and introduction of new tests,

Results of all tests applied to RELM halftime report,

Stability and statistical power

Space-rate-magnitude forecasts

Forecasts specify expected rates of earthquakes in bins of latitude-longitude-magnitude:

$$\Lambda = \{ \lambda(i, j) \mid i \in \mathbf{M}, j \in \mathbf{S} \}$$

where \mathbf{M} is the binned magnitude range of interest, and \mathbf{S} is the binned spatial range of interest (latitude-longitude cells).

Assumptions

- **A1. Poisson:** Forecast uncertainty—in each bin, and therefore for the entire forecast—is characterized by the Poisson distribution, the rate in each bin being the expected value.
- **A2. Independence:** The rate forecast in each bin is independent of all other bins (even those within the same spatial cell and nearby spatial cells).

Likelihood

When the forecast period expires, in a single bin the likelihood of the observation given the forecast is (from **A1**):

$$Pr(\omega|\lambda) = \frac{\lambda^\omega}{\omega!} \exp(-\lambda).$$

Considering all bins, the joint likelihood of the observations given the forecast is (from **A2**):

$$Pr(\mathbf{\Omega}|\mathbf{\Lambda}) = \prod_{(i,j) \in R} Pr(\omega(i,j)|\lambda(i,j)).$$

Poisson joint likelihood

There are 162 regular season Major League Baseball games. My 2008-2009 forecast:



- LA Dodgers will win 150 games,
- SF Giants will win 50 games.

In the 2008-2009 season,



- Dodgers won 96 games, and
- Giants won 88 games.



$$\Pr(\omega_1 = 95 \mid \lambda_1 = 150) * \Pr(\omega_2 = 8 \mid \lambda_2 = 50) = 1.25e-13$$

L(likelihood)-test

Q: Based on the joint likelihood, was the forecast any good at estimating the space-rate-magnitude distribution of seismicity?

A: It's hard to say.

Q: If the forecast be the right model for seismicity, what would the likelihood distribution be?

A: Let us check by simulations.

L-test

- Simulate catalogs, $\{\tilde{\Omega}\}$, that are consistent with the forecast (by Monte Carlo sampling).
- For each simulated catalog, determine the likelihood, yielding a set of simulated joint likelihoods: $\{\tilde{L}\}$, where $\tilde{L}_i = \Pr(\tilde{\Omega}_i | \Lambda)$.

With the L-test, we check if the observed joint likelihood is smaller than most of the simulated joint likelihoods:

$$\gamma = \frac{|\{\tilde{L}_x | L \leq \tilde{L}_x\}|}{|\{\tilde{L}\}|}.$$

N(umber)-test

Q: Is the forecast any good at estimating the rate of seismicity?

A: We can check using the Poisson cumulative distribution,

$$\delta = F(N_{obs} | N_{fore})$$

where N_{obs} is the number of observed earthquakes, and N_{fore} is the number of forecast earthquakes.

N-test flaw

In the case when there are zero earthquakes observed and a small forecast rate, the two-sided N-test incorrectly rejects the forecast.

- If $\lambda = 0.0015$, $\delta \sim 0.9985$.

Rather, we should consider two one-sided tests:

- $\delta_1 = 1 - F((N_{obs} - 1) | N_{fore})$, at least N_{obs}
- $\delta_2 = F(N_{obs} | N_{fore})$, at most N_{obs} .

Magnitude likelihood

Q: Is the forecast any good at estimating the magnitude distribution of seismicity?

A: We can isolate the magnitude component of the forecast and apply a test similar to the L-test.

$$\Lambda^m = \{ \lambda^s(i) | i \in M \}$$

$$\lambda^m(i) = \frac{N_{obs}}{N_{fore}} \sum_{j \in S} \lambda(i, j)$$

M(magnitude likelihood)-test

- Simulate catalogs, $\{\widetilde{\Omega}^m\}$, that are consistent with the magnitude forecast (by Monte Carlo sampling).
- For each simulated catalog, determine the magnitude joint likelihood, yielding a set of simulated joint likelihoods: $\{\widetilde{L}^m\}$, where $\widetilde{L}_i^m = \Pr(\widetilde{\Omega}_i^m | \Lambda^m)$.

With the M-test, we check if the observed magnitude joint likelihood is smaller than most of the simulated magnitude joint likelihoods:

$$\kappa = \frac{|\{ \widetilde{L}_x^m | L^m \leq \widetilde{L}_x^m \}|}{|\{ \widetilde{L}^m \}|} .$$

Spatial likelihood

Q: Is the forecast any good at estimating the spatial distribution of seismicity?

A: We can isolate the spatial component of the forecast and apply a test similar to the L-test.

$$\Lambda^s = \{ \lambda^s(j) \mid j \in \mathcal{S} \}$$

$$\lambda^s(j) = \frac{N_{obs}}{N_{fore}} \sum_{i \in M} \lambda(i, j)$$

S(patial likelihood)-test

- Simulate catalogs, $\{\widetilde{\Omega}^s\}$, that are consistent with the spatial forecast (by Monte Carlo sampling).
- For each simulated catalog, determine the spatial joint likelihood, yielding a set of simulated joint likelihoods: $\{\widetilde{L}^s\}$, where $\widetilde{L}_i^s = \Pr(\widetilde{\Omega}_i^s | \Lambda^s)$.

With the S-test, we check if the observed spatial joint likelihood is smaller than most of the simulated spatial joint likelihoods:

$$\zeta = \frac{|\{\widetilde{L}_x^s | L^s \leq \widetilde{L}_x^s\}|}{|\{\widetilde{L}^s\}|}.$$

Halftime results for RELM mainshock forecasts

Forecast	L γ	N δ_1, δ_2
Ebel-Mainshock	0.149	0.634, 0.503
Helmstetter-Mainshock	0.723	0.726, 0.391
Holliday-PI	0.992	0.996, 0.011
Kagan-Mainshock	0.974	0.982, 0.063
Shen-Mainshock	0.969	0.967, 0.107
Ward-Combo	0.998	0.999, 0.004
Ward-Geodetic81	1.000	1.000, 0.000
Ward-Geodetic85	0.987	0.993, 0.030
Ward-Geologic	0.998	0.998, 0.011
Ward-Seismic	0.993	0.997, 0.014
Ward-Simulation	0.725	0.885, 0.282
Wiemer-ALM	0.637	0.834, 0.256

Halftime results for RELM mainshock forecasts

Forecast	L γ	N δ_1, δ_2	M κ	S ζ
Ebel-Mainshock	0.149	0.634, 0.503	0.793	0.000
Helmstetter-Mainshock	0.723	0.726, 0.391	0.856	0.468
Holliday-PI	0.992	0.996, 0.011	0.717	0.025
Kagan-Mainshock	0.974	0.982, 0.063	0.660	0.755
Shen-Mainshock	0.969	0.967, 0.107	0.654	0.931
Ward-Combo	0.998	0.999, 0.004	0.695	0.767
Ward-Geodetic81	1.000	1.000, 0.000	0.706	0.718
Ward-Geodetic85	0.987	0.993, 0.030	0.705	0.724
Ward-Geologic	0.998	0.998, 0.011	0.700	0.534
Ward-Seismic	0.993	0.997, 0.014	0.706	0.700
Ward-Simulation	0.725	0.885, 0.282	0.541	0.557
Wiemer-ALM	0.637	0.834, 0.256	0.891	0.001

How stable are these results?

With respect to catalog uncertainties, we can quantify stability under the following assumptions:

- Magnitude uncertainty is well-described by a Laplace distribution with scale parameter $\nu = 0.1$ to 0.3 (Werner & Sornette 2008);
- Location uncertainty is well-described by a Gaussian distribution with standard deviation $\sigma = 5$ km;
- and measurement uncertainties are the same for all earthquakes.

We repeatedly perturb the observed catalog and compute the quantile score distribution.

Result stability

Forecast	L (γ)	N (δ_1, δ_2)	M (κ)	S (ζ)
Bird-Neokinema	1.000	1.000, 0.001	0.889	0.353
	1.00 + 0.00/ - 0.00	1.00 + 0.00/ - 0.00, 0.00 + 0.00/ - 0.00	0.92 + 0.08/ - 0.27	0.49 + 0.38/ - 0.29
	0.96 + 0.04/ - 0.18	0.98 + 0.02/ - 0.13, 0.04 + 0.16/ - 0.04	0.78 + 0.21/ - 0.39	0.15 + 0.39/ - 0.14
Ebel-Aftershock	1.000	1.000, 0.000	0.815	0.000
	1.00 + 0.00/ - 0.00	1.00 + 0.00/ - 0.00, 0.00 + 0.00/ - 0.00	0.85 + 0.12/ - 0.31	0.00 + 0.00/ - 0.00
	1.00 + 0.00/ - 0.06	1.00 + 0.00/ - 0.00, 0.00 + 0.00/ - 0.00	0.73 + 0.24/ - 0.49	0.00 + 0.00/ - 0.00
Helmstetter-Aftershock	0.949	0.937, 0.104	0.890	0.523
	0.97 + 0.03/ - 0.06	0.96 + 0.03/ - 0.07, 0.06 + 0.10/ - 0.04	0.92 + 0.08/ - 0.28	0.48 + 0.27/ - 0.26
	0.44 + 0.39/ - 0.32	0.45 + 0.39/ - 0.33, 0.64 + 0.28/ - 0.41	0.79 + 0.20/ - 0.38	0.39 + 0.41/ - 0.32
Kagan-Aftershock	0.895	0.899, 0.193	0.901	0.793
	0.95 + 0.04/ - 0.10	0.96 + 0.03/ - 0.15, 0.10 + 0.09/ - 0.06	0.88 + 0.11/ - 0.26	0.97 + 0.02/ - 0.23
	0.57 + 0.33/ - 0.42	0.54 + 0.35/ - 0.43, 0.60 + 0.34/ - 0.40	0.76 + 0.22/ - 0.49	0.99 + 0.01/ - 0.20
Shen-Aftershock	0.896	0.854, 0.262	0.908	0.981
	0.92 + 0.06/ - 0.12	0.89 + 0.08/ - 0.16, 0.20 + 0.20/ - 0.14	0.88 + 0.11/ - 0.27	0.99 + 0.01/ - 0.10
	0.43 + 0.43/ - 0.33	0.45 + 0.41/ - 0.38, 0.69 + 0.28/ - 0.42	0.76 + 0.22/ - 0.50	0.92 + 0.07/ - 0.49

How powerful are these tests?

What is the probability that they correctly reject an incorrect null hypothesis?

In our case, how well do they distinguish forecasts?

We can simulate catalogs consistent with a given Forecast A and see how often a test “rejects” Forecast B.

Power of N-test

For the N-test, no simulations are needed.

Rejection happens when $\delta_1 < \alpha_{eff}$ or $\delta_2 < \alpha_{eff}$. We can compute the probability that either occurs:

$$\sum_{\{i: \delta_1(i|\lambda_2) < \alpha_{eff}\}} Pr(i|\lambda_1) + \sum_{\{j: \delta_2(j|\lambda_2) < \alpha_{eff}\}} Pr(j|\lambda_1)$$

Power of N-test

For the N-test, no simulations are needed.

Rejection happens when $\delta_1 < \alpha_{eff}$ or $\delta_2 < \alpha_{eff}$. We can compute the probability that either occurs:

$$\underbrace{\sum_{\{i: \delta_1(i|\lambda_2) < \alpha_{eff}\}} Pr(i|\lambda_1)}_{\text{underestimation}} + \underbrace{\sum_{\{j: \delta_2(j|\lambda_2) < \alpha_{eff}\}} Pr(j|\lambda_1)}_{\text{overestimation}}$$

Here, λ_1 is the overall rate forecast of Forecast 1, and λ_2 is the overall rate forecast of Forecast 2.

Forecast rates in overlapping regions

Λ_1	Λ_2				
	1.	2.	3.	4.	5.
1. Bird- NeoKinema	27.921	17.335	27.921	15.741	15.714
2. Ebel- Aftershock	36.362	36.362	36.362	19.946	20.323
3. Helmstetter- Aftershock	17.682	12.776	17.682	9.838	9.966
4. Kagan- Aftershock	7.982	4.815	7.982	7.982	7.696
5. Shen- Aftershock	7.316	4.737	7.316	6.973	7.316

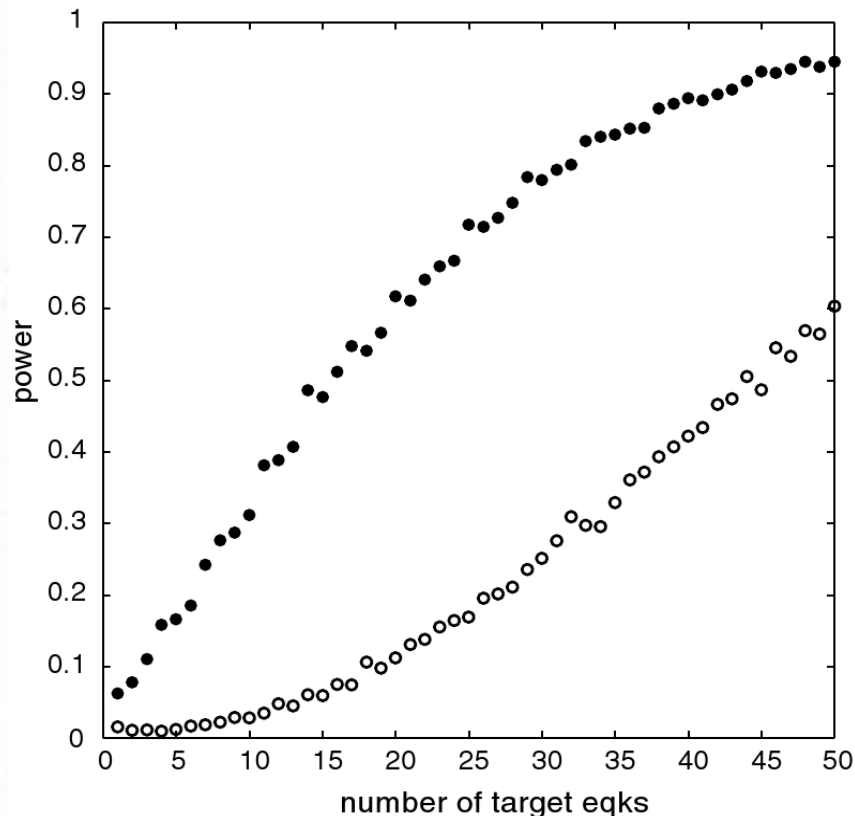
Corresponding N-test power

Λ_1	Λ_2				
	1.	2.	3.	4.	5.
1. Bird-NeoKinema	0.037	0.951	0.595	0.608	0.702
2. Ebel-Aftershock		0.038	0.998	0.989	0.996
3. Helmstetter- Aftershock			0.042	0.078	0.136
4. Kagan-Aftershock				0.031	0.030
5. Shen-Aftershock					0.041

Forecast rates in overlapping regions

Λ_1	Λ_2				
	1.	2.	3.	4.	5.
1. Bird- NeoKinema	27.921	17.335	27.921	15.741	15.714
2. Ebel- Aftershock	36.362	36.362	36.362	19.946	20.323
3. Helmstetter- Aftershock	17.682	12.776	17.682	9.838	9.966
4. Kagan- Aftershock	7.982	4.815	7.982	7.982	7.696
5. Shen- Aftershock	7.316	4.737	7.316	6.973	7.316

Statistical power



Power depends on the forecasts considered and the number of events observed. We expect low power for the M-test because most forecasts use Gutenberg-Richter distribution.

Final thoughts

Isolation of space and magnitude components can yield further insight into forecast performance and help in interpreting L-test results.

The above work could be extended if more comprehensive analysis of catalog errors were done. Can such analysis be integrated into catalog creation process?

Estimates of stability and statistical power should accompany forecast testing results.