The Q-score: a magnitude-weighted goodness-of-fit score for earthquake forecasting

Julia Jansson Valter[1]; Alejandra Arjon[2]; Francesco Serafini[3]; Frederic Schoenberg[2]

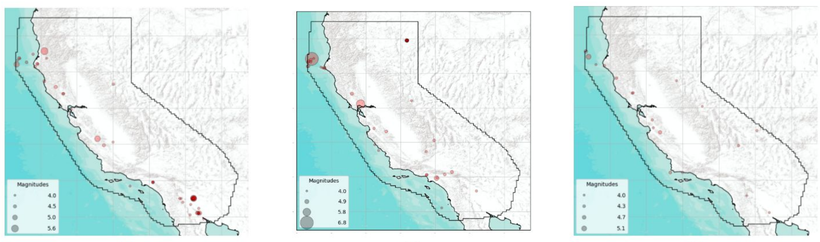
[1]Chalmers University & University of Gothenburg, [2]University of California, Los Angeles (aarjon@ucla.edu), [3]University of Bristol

Abstract

Comparing models and assessing goodness-of-fit of earthquake forecasts can be done using tests developed by CSEP. However, most current earthquake forecast evaluation metrics do not give additional weight to large magnitude events. To this end, magnitude weighted goodness-of-fit scores for earthquake forecasting have recently been introduced, such as potency weighted log-likelihood and the Q-score. We further investigate the Q-score, emphasizing model fit for the 5% largest earthquakes. Under certain regularity conditions, the expectation of the Q-score approaches 1. The Q-score is evaluated for 21 CSEP next-day gridded forecasts for California from 2012, 2014, and 2017. This is in conjunction with the log-likelihood scores to assess how well different models perform, both for predicting the largest events but also in terms of overall fit.

Introduction

The Q-score is a new magnitude-weighted goodness-of-fit score in the context of earthquake forecasting. We explore some of the properties in order to construct the conditions for a null hypothesis. The Q-score, in conjunction with the L-score, is calculated for various models through various time periods in order to have a more comprehensive evaluation of the model's performance.



(a) (b) (c)

Figure 1: ComCat catalog events above magnitude 3.95 in California. (a) 50 events in the year 2012, with a 5.6 magnitude of the largest event. (b) 44 events in the year 2014, with a 6.8 magnitude of the largest event. (c) 18 events in the year 2017, with a 5.08 magnitude of the largest event.

CSEP Forecasts

- Daily forecasts produced by 21 models ranging from the dates of 01 August, 2008 to 31 August 2018 for the California region. The cumulative rate from January 1st to December 31st is used in the evaluation of the models.
- Models have been operated in a fully prospective fashion, independently of the modelers (Zechar et al., 2010).
- Forecasts are grid-based, covering a 0.1°×0.1° latitude, longitude space grid with 0.1 magnitude bin intervals, which range from 3.95 to 8.95 and one open-end bin for magnitudes over 8.95.

Q-Score

The Q-score proposed by Schoenberg and Schorlemmer (2024, 2025) is defined as:

$$Q = \frac{\text{mean}\{\lambda_i : m_i > m^{[0.95]}\}}{\text{mean}\{\lambda(t,x,y,m)\}},$$

- Let $\Lambda(t,x,y) = \int_{\boldsymbol{M}} \lambda(t,x,y,m)dm$ be the conditional intensity aggregated over the magnitudes for the point process N.
- $A_\alpha = \{(t,x,y : m_{t,x,y} > M^{[1-\alpha]}\}$
- The Q statistic can be defined as:

$$Q = \frac{\left(\int \Lambda(t,x,y)\mathbf{1}\{(t,x,y) \in A_\alpha\}dN\right) / \left(\int \mathbf{1}\{(t,x,y) \in A_\alpha\}dN\right)}{\left(\int \Lambda(t,x,y)dN\right) / \left(\int dN\right)},$$

Results

Asymptotics of Q

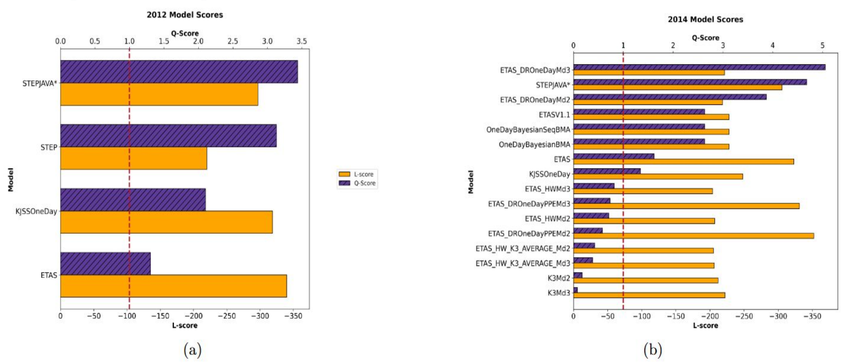
- **A1**: Assume separability of the marks
- **A2**: Assume that the ground process N is stationary and metrically transitive with finite mean density b.
- For stationary random measures: metric transitivity $\iff$ ergodicity
- **A3**: Assume that ΛN is also stationary and metrically transitive, with finite mean density $\bar{\lambda}b$, where $(\Lambda N)((0,T]) = \int_0^T \Lambda(t)dN$
- Let

$$Q_T = \frac{\left(\int_0^T \Lambda(t)\mathbf{1}\{t \in A_\alpha\}dN\right) / \left(\int_0^T \mathbf{1}\{t \in A_\alpha\}dN\right)}{\left(\int_0^T \Lambda(t)dN\right) / \left(\int_0^T dN\right)}.$$

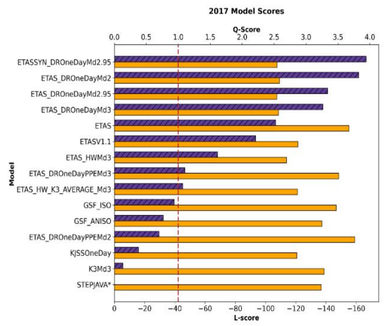
Theorem

Under **A1**–**A3**, $\lim_{T\to\infty} Q_T = 1$ *almost surely and in* L_1 *norm.*

- These assumptions help form the conditions for a null hypothesis; in practice, one may set the null hypothesis that Q = 1, and evaluate if the value of Q for a given model and a given dataset exceeds 1 by a statistically significant margin.



(a) (b)



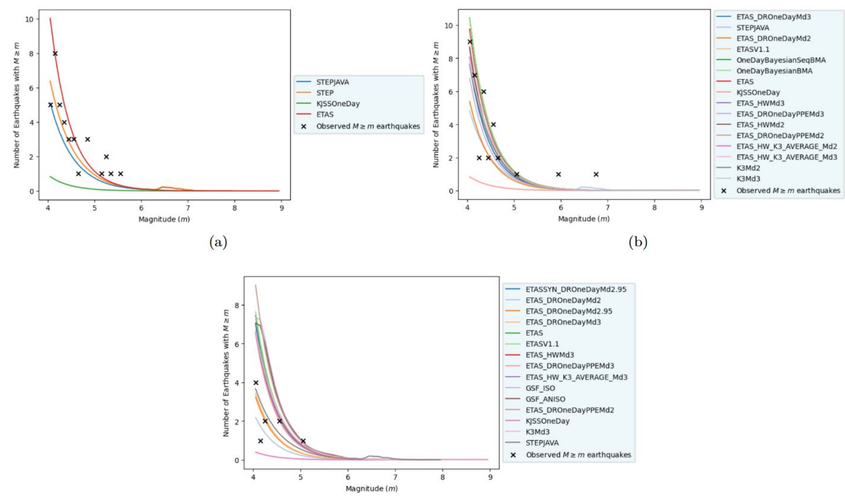
(c)

Figure 2: Q-scores and L-scores of the models for the three different years. Models are in descending order based on the Q-score. The red dashed line indicates a Q-score of 1. The purple bars are the model's Q-score, and the orange bar its corresponding L-score. NOTE: STEPJAVA* uses a waterlevel of 10^-6 to calculate the L-score. In reality, all log-likelihood scores for STEPJAVA diverge to infinity. Models with a Q-score above 1, and an L-score close to 0 should be considered good fits to the observed data. Forecast model data was obtained from https://zenodo.org/records/15076187.

- While the Q-score may be useful in highlighting models that forecast higher rates where the largest events occur, it is not a proper score and should not be used independently from other evaluation scores for a comprehensive evaluation of goodness of fit.
- The L-score gives a general goodness-of-fit score between the model's forecast and the observed events, it does not explicitly weight large magnitude events more than the other events.

Conclusion/Discussion

- The ETAS_DROneDayMd models have the highest Q and L-scores in the years where a forecast was available.
- STEPJAVA is the model with the highest Q-score in some of the time periods, but it is the only model that we analyzed for which the L-score diverges every year. The STEP and STEPJAVA models, as seen in Figure 3 seem to forecast more magnitude 6 earthquakes than the rest of the models.
- However, STEP does seem to a good fit to the observed data, evidenced by the relative high L-scores, while also having a high Q-score.
- Conducting a more thorough evaluation using other scores in addition to the Q and L-scores needs to be conducted in order to explore if the magnitude distribution of these models are affecting these scores.



(a) (b)

(c)

Figure 3: Plot of the annual magnitude distribution for the models available. (a) is for the forecasts during the year 2012, (b) is for the year 2014, and (c) is for the year 2017. The 'X' markers indicate the observed number of events.

References

Schoenberg, F. and Schorlemmer, D. (2024). Critical questions about CSEP, in the spirit of Dave, Yan, and Ilya. *Seismological Research Letters,* 95(6):3617–3625. https://doi.org/10.1785/0220240213.

Schoenberg, F. (2025). Magnitude-weighted goodness-of-fit scores for earthquake forecasting. *Spatial Statistics,* 67:100895. https://doi.org/10.1016/j.spasta.2025.100895.

Zechar, J. D., Schorlemmer, D., Liukis, M., Yu, J., Euchner, F., Maechling, P. J. and Jordan, T. H. (2010). The Collaboratory for the Study of Earthquake Predictability perspective on computational earthquake science. *Concurrency and Computation: Practice and Experience,* 22(12): 1836-1847. https://doi.org/10.1002/cpe.1519.

Take a picture to **read the full paper**

Acknowledgements

The main part of this work was done during the visit of Julia Jansson Valter at the Department of Statistics and Data Science at UCLA. This research visit was supported by The Foundation Blanceflor and Adlerbertska Forskningsstiftelsen. Research supported by Southern California Earthquake Center Fund number 25166.

Data supporting the findings of this study are openly available on Zenodo at https://doi.org/10.5281/zenodo.15076187

Earthquake catalogs and gridded forecast tools used are found in the python package pyCSEP. Code repository for the evaluation results can be found on GitHub at: https://github.com/a-arjon/Q-score-a-magnitude-weighted-goodness-of-fit-score-for-earthquake-forecasting