Summary of Results from SCEC grant during 2020-2021, Statistical evaluation tools for earthquake simulation models.

The discrimination between competing models and the assessment of which models seem most consistent with observed seismicity have become increasingly important. Concerns with retrospective analyses, especially regarding data selection, overfitting and lack of reproducibility, have led to the development of the Regional Earthquake Likelihood Models (RELM) testing center and subsequently the Collaborative Study of Earthquake Predictability (CSEP), which was designed to evaluate and compare the goodness of fit of various earthquake forecasting models (Jordan 2006, Field 2007). Such centers, which require forecasts to be fully automatic, with no subjective or retrospective adjustments made by the modelers, are essential tools to assess the fit of models to observed seismicity and to determine which models seem best suited to earthquake forecasting. This paradigm has many benefits from a statistical perspective. The prospective nature of the experiments effectively eliminates concerns about overfitting. Furthermore, the standardized nature of the data and forecasts facilitates the comparison among different models.

One component of our research was a comparison of the fit of some of the best performing models submitted to CSEP, such as the model of Zhuang et al. (2003) and the model of Gordon (2017), using data collected and compiled after all details of the models were specified, recorded, and submitted, using one-day forecasts and earthquake data from Southern California in 2017, obtained from the CSEP repository. Forecasts were made over a grid of cells of size 0.1 latitude by 0.1 longitude within latitude range [-125.4 , -113.2] and longitude range [31.5 , 42.9], for a range of magnitude and depths. For each model, 7682 forecasted aggregated conditional rates are obtained. The seismicity data include the estimated hypocentral latitude, longitude, magnitude, depth, date, and time for each of the 19 earthquakes observed in this region in 2017. The models we used for comparison are one-day seismicity forecasting models selected from CSEP's rate-based repository that output an expected number of earthquakes in each spatial grid for each day. Max Werner and Bill Savran for their great help in providing us the data and forecasts for the models. For model evaluation, we focused in particular on residual analysis techniques for space-time point process models, including Voronoi residuals, deviance residuals, and super-thinned residuals.

A version of the Epidemic-Type Aftershock Sequence (ETAS) model of Ogata (1988, 1998) was implemented by Zhuang et al. (2003) and fits a parametric triggering function with parameters estimated by maximum likelihood, and with background rate $\mu(s)$ estimated via kernel smoothing with variable bandwidths estimated by optimizing the fit of residuals to a stationary Poisson process. The Gordon (2017) model is a version of ETAS where the spatial triggering function varies not only with magnitude but also with direction and spatial region. The model of Gordon (2017) allows for the estimation of a primary fault direction for each earthquake based on prior local seismicity using weighted least squares, and the triggering density of its aftershocks is allowed to vary relative to this estimated primary direction. In addition, the spatial triggering function is permitted to vary across subregions within California. In the Gordon (2017) model, the triggering density g is estimated nonparametrically using the Model Independent Stochastic Declustering (MISD) method of Marsan and Lengliné (2008). We showed in Gordon and Schoenberg (2021) that the fit to data of the Gordon (2017) model is similar to and overall appears slightly better than the Helmstetter et al. (2007) model shown by Schorlemmer et. al (2010) to have performed best among CSEP models by the L-tests, N-tests and R-tests.

We also proposed a new, non-parametric version of the ETAS model where each point may have its own productivity and considered the simultaneous estimation of all these productivities, without any parametric constraints on how the productivity varies over time. If n points are observed between time 0 and time T, there are thus n productivities to estimate, and we propose estimating them by maximum likelihood. The resulting estimates will then have very large variance but may be smoothed to produce more stable estimates. We discovered that these estimates can be computed simply and rapidly provided the invertibility of the adjacency matrix, G (see Schoenberg 2021). The speed with which the estimates may be obtained enables approximate standard errors for these estimates to be constructed by repeated simulation and estimation. We compared the results of our estimator using simulations and applied our method to analyze the productivities of earthquakes in the Hollister-Bear Valley region, a 35km portion of the San Andreas Fault suggested by Bruce Bolt as an example of seismic hazard calculations and studied in Schoenberg and Bolt (2000) using earthquakes of magnitude at least 3.0 and depth $\leq$ 700km from 1970-2000. We included earthquakes from 1/1/1970 to 3/6/2020 and slightly expanded the region spatially by 0.2^o in each direction to latitude 36.3 to 37.2 and longitude -120.3 to -121.2. The depth of the deepest earthquake in the catalog is just 50.84km. Data were obtained from the Northern California Earthquake Data Center (NCEDC 2014). The results indicate small but significant lack of fit of the ETAS model, so the nonparametric productivity estimate may be preferable in this case.

To evaluate model performance we use Voronoi residuals (Bray et al., 2014), Voronoi deviance residuals (Clements et al. 2011), and superthinned residuals (Clements et al. (2012). Voronoi residuals and Voronoi deviances are useful for evaluating gridded forecasts especially when a substantial proportion of pixels have very small integrated conditional intensities. Furthermore, Voronoi based residual methods offer advantages over grid based residuals in that with the former type of residuals, the spatial partition is data-driven and spatially adaptive, and the resulting distribution of residuals is usually far less skewed in such situations than residuals integrals over fixed rectangular grid cells (Bray et al. 2014).

Voronoi residuals are constructed by partitioning the space Voronoi cells, with each cell defined as the region consisting of all locations that are closer to the observed event than to any of the other points. Each Voronoi cell C_i has only one point inside it by construction. Hence, a raw Voronoi residual for each cell C_i is given simply by $\hat{R}_i = 1 - \int\limits_{C_i} \hat{\lambda}(s,t)dtds$. Voronoi residuals were shown to be considerably less skewed than pixel residuals in Bray et al. (2014). One difficulty when plotting Voronoi residuals is the determination of an appropriate color scale. Bray et al. (2014) proposed using a probability integral transformation (PIT) to scale the Voronoi residuals uniformly, which is computationally intensive since it requires repeated simulation of the model under consideration. Here, we instead fit a homogeneous Poisson process model, with rate fit by maximum likelihood, and use the standardized Voronoi residuals for this null model as a scale by which to judge the residuals of alternative models.

The competing models were compared using Voronoi deviance analysis, where one considers the difference between the log-likelihoods of the two point process models over Voronoi cells. Instead of simply comparing the observed to the forecasted seismicity within each pixel, the difference between the log-likelihoods of two competing models is examined. Large Voronoi deviance residuals indicate places where one model fit substantially better than the other, and the sign of the residual indicates which model had superior fit.

Figure 1 displays the Voronoi residuals for the Zhuang et al. (2003) and Gordon (2017) models, as well as deviances between the two models. The overall log-likelihood and information gain for the Zhuang et al. (2003) model are -13.721 and -0.722, retrospectively, while the corresponding values for the Gordon (2017) model are -8.865 and -0.467 respectively, indicating better performance of

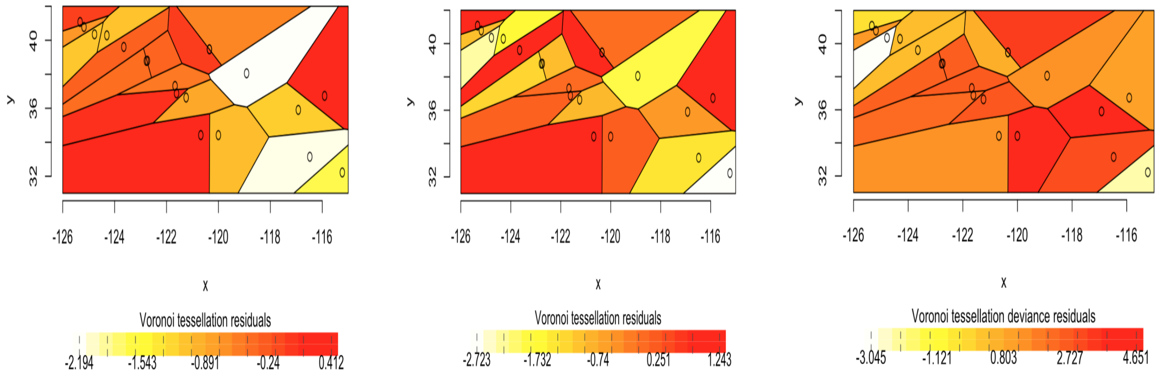

Figure 1: Voronoi residuals for Zhuang et al. (2003) model (top), Voronoi residuals for Gordon (2017) model (middle), and Voronoi deviance residuals for the Gordon (2017) model compared to the Zhuang et al. (2003) model (bottom). Positive deviances indicated locations where the Gordon (2017) model fit better than the Zhuang et al. (2003) model.

the Gordon (2017) model overall. Both models appeared to adequately forecast the conditional intensity around the Sierra earthquake [latitude 39.47^o, longitude -120.35^o] on June 27, 2017, near the Mohawk Valley fault zone, and around the Santa Clara earthquake [latitude 37.31^o, longitude -121.67^o] on October 10, 2017, between Silver Creek fault and San Jose fault, as well as the Paicines earthquake [latitude 36.63^o, longitude -121.24^o] on November 13, 2017 in San Benito on the Paicines fault in the Calaveras fault zone. The Gordon (2017) model also appears to forecast adequately around the earthquakes occurring at [latitude 34.42^o, longitude -120.00^o] on May 17, 2017, at [latitude 36.88^o, longitude -121.61^o] on March 31, 2017, near the San Andreas fault zone, and at [latitude 35.90^o, longitude -116.92^o] on August 22, 2017, near Death Valley. The Zhuang et al. (2003) model seemed to forecast more accurately around the Collayomi fault zone area. Both models appeared to have failed to anticipate the two earthquakes at [latitude 41.08^o, longitude -125.33^o] on July 29, 2017 and [latitude 34.42^o, longitude -120.68^o] on May 17, 2017 near the Pacific Ocean, as well as the earthquake at [latitude 32.23^o, longitude -115.23^o] on March 22, 2017 near the Mexico-California border. While both the Gordon (2017) model and the Zhuang et al. (2003) model forecast well in the central part of Southern California, the Gordon (2017) model appears to improve upon the Zhuang et al. (2003) forecasts near the San Andreas Fault zone. On the other hand, in the periphery of California, where both models forecast more poorly, it seems that the Zhuang et al. (2003) model makes more reasonable forecasts while the Gordon (2017) model has a tendency to overpredict seismicity, particularly in the vicinity of the earthquake at [latitude 40.35^o, longitude -124.77^o] on March 6, 2017 near the West Coast.

We also used super-thinned residuals to compare the goodness of fit for two models. In super-thinning a given model with estimated conditional intensity $\hat{\lambda}$, as proposed in Clements et al. (2012), one first thins the point process N, keeping each point (s_i, t_i) independently with probability $\min\{k/\hat{\lambda}(s_i, t_i), 1\}$ to obtain a thinned residual process Z_1. Next, one simulates a Cox process Z_2 directed by $\max\{k - \hat{\lambda}(s, t), 0\}$. The points of the residual point process $Z = Z1 + Z2$, obtained by superposing the thinned residuals and the simulated Poisson process, are called super-thinned residuals. Because Z is homogeneous Poisson with rate k if and only if the modeled conditional intensity is correct almost everywhere (Clements et al. 2012), one may inspect the points in Z for uniformity as a way of assessing the goodness-of-fit of the model.

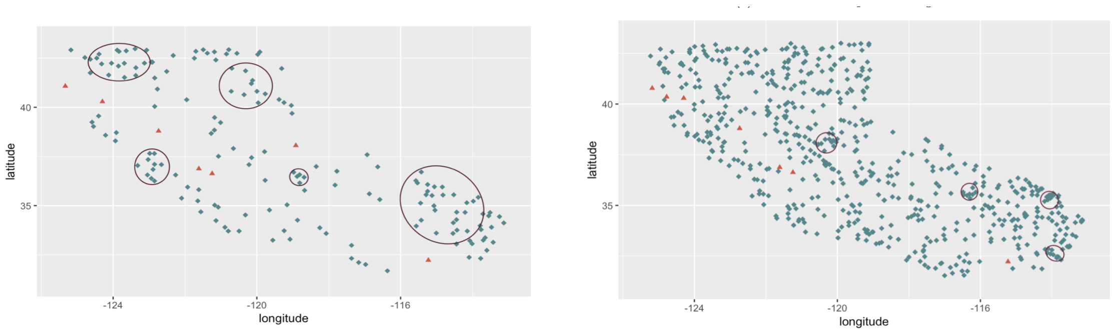

Figure 2: Super-thinned residuals for Zhuang et al. (2003) model (left) and Gordon (2017) model (right). Red triangles indicate observed events after thinning and blue squares indicate superposed points.

Figure 2 displays the super-thinned residuals for the Zhuang et al. (2003) and Gordon (2017) models. The super-thinned residuals corresponding to the Zhuang et al. (2003) model exhibit noticeable clustering, while for the Gordon (2017) model the super-thinned residuals appear nearly uniformly distributed, despite some minor clustering at small radii. The super-thinned residuals for the Zhuang et al. (2003) model indicate underprediction of seismicity around the Calumet Fault around [latitude 34.3^o, longitude 115.5^o], around the Paleo-subduction zone near [36.8^o, -123^o], in the mountain areas around [42.3^o, -124.1^o], and near the East side of the Madeline Plains around [40.8^o, -120.1^o]. There are also noticeable gaps near the Willard fault [33.5, -117.2^o], the Rinconada fault [35.6^o, -120.7], near the Tahoe Valley Fault zone [38.9^o, -120.0^o], and near Concord Fault [38.0^o, -122.0^o], suggesting areas where the Zhuang et al. (2003) model overpredicted seismicity.

The super-thinned residuals of the Gordon (2017) model exhibit some noticeable clustering, especially close to the Red Pass fault [35.2^o, -116.3^o] and in the Homer Mountain area [35.0^o, -114.9^o], indicating under-prediction of seismicity. There may also be gaps indicating over-prediction of seismicity of the Gordon (2017) model near the Cambria Fault zone [35.5^o, -121.0^o], the La Nacion Fault zone [32.6^o, -117.0^o], and near Mission Fault [37.7^o, 122.0^o].

Our main finding was that point process residual methods such as Voronoi deviance residuals and superthinned residuals are quite powerful, capable of identifying substantial differences in quality of fit and highlighting key spatial-temporal areas where improvement may be desired, even when using a small dataset and when comparing two models that fit and forecast earthquake occurrences very accurately. Certainly, when comparing or evaluating models with more obvious departures from the data, such lack of fit would be far easier to detect. Importantly, our analysis was prospective rather than retrospective, in line with the goals of CSEP. Retrospective analyses are frought with difficulties, as among other problems, publication bias and overfitting issues often yield analyses suggesting newer, more complex models will outperform older, simpler models, only to be debunked later. Here, both models were fully implemented in CSEP prior to the collection of the data used in their comparison, rendering the issue of overfitting essentially moot.

We found slightly superior fit of the Gordon (2017) model compared to the version of the ETAS model implemented by Zhuang et al. (2003), though both models fit the data overall very well.

The results are in agreement with the retrospective analysis in Gordon (2017) and suggest that use of the nonparametric estimation method of Marsan and Lengliné (2008), and taking strike angle information into account, tends to result in more accurate forecasts of seismicity. In particular, the Gordon (2017) model appears to improve upon the Zhuang et al. (2003) forecasts near the San Andreas Fault zone, though it may overpredict seismicity in several off-fault locations, especially near the West Coast of California.

References

Bray A, Wong K, Barr CD, and Schoenberg FP (2014). Voronoi residual analysis of spatial point process models with applications to California earthquake forecasts. *Ann. Appl. Stat.* 8(4), 2247-2267.

Clements RA, Schoenberg FP, and Schorlemmer D (2011). Residual analysis methods for space-time point processes with applications to earthquake forecast models in California. *Annals of Applied Statistics* 5(4), 2549-2571.

Clements RA, Schoenberg FP, and Veen A (2012). Evaluation of space-time point process models using super-thinning. *Environmetrics* 23(7), 606-616.

Gordon, J.S., and Schoenberg, F.P. (2021). A nonparametric Hawkes model for forecasting California seismicity. *BSSA*, in review.

Marsan D and Lengliné O (2008). Extending earthquakes' reach through cascading. *Science*, 319(5866):1076-1079.

Ogata Y (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association* 83(401), 9-27.

Schorlemmer D, Gerstenberger M, Wiemer S, and Jackson D (2010). First results of the Regional Earthquake Likelihood Models experiment. *Pure and Applied Geophysics* 167(8-9), 859-876.

Zhuang J, Ogata Y, and Vere-Jones D (2003). Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association* 97(458), 369-380.