Poisson Tests of Declustered Catalogs

Brad Luen
Department of Mathematics
Reed College
Portland, OR, USA

Philip B. Stark
Department of Statistics
University of California
Berkeley, CA, 94720-3860 USA

Accepted 2012 ???? ??; Received 2012 January ??; in original form 2011 January 29

SUMMARY
The commonly enunciated reason to decluster catalogs is so that the remaining "main" events will be consistent with a spatially inhomogeneous, temporally homogeneous Poisson process (SITHP) model. But are they? Conclusions depend on the declustering method, the catalog, the magnitude range, and the statistical test. Gardner and Knopoff's (1974) conclusion that 1932–1971 southern California events with $M \geq 3.8$ are Poissonian after declustering apparently results from their use of a test with low power. That test ignores space, is insensitive to long-term rate variations, is relatively insensitive to seismicity rate fluctuations on the scale of weeks, and uses an inaccurate approximation to the null distribution of the test statistic. Better temporal tests and a novel spatio-temporal test show that SITHP does not fit $M \geq 3.8$ 1932–1971 or 1932–2010 Southern California Earthquake Center (SCEC) catalogs declustered using Gardner and Knopoff's windows in a linked-window or a mainshock-window algorithm. For $M \geq 4.0$, SCEC catalogs declustered using the Gardner-Knopoff windows in a linked-window method are far closer to SITHP, while catalogs declustered using those windows in a mainshock-window method are inconsistent with SITHP. Reasenberg's (1985) declustering method applied to southern California seismicity produces catalogs inconsistent with SITHP, even for events with $M \geq 4.0$.
If enough events are deleted from a catalog, the remainder always will be consistent with SITHP. This suggests posing declustering as an optimization problem: Delete the fewest events such that those left pass a particular test or suite of tests for SITHP. While that optimization problem is combinatorially complex, inexpensive suboptimal methods are surprisingly effective: Declustered catalogs can be consistent with temporal tests of SITHP at significance level 0.05 and have 50% to 80% more events than window-declustered catalogs that are inconsistent with SITHP. But tests that incorporate spatial information reject the SITHP hypothesis for those declustered catalogs, illustrating the importance of using spatial information.

Key words: Earthquake interaction, forecasting, and prediction; Probabilistic forecasting; Declustering; Statistical seismology

1 INTRODUCTION

We study the most common declustering methods, *mainshock-window* and *linked-window* declustering. There are also stochastic declustering methods, which use chance to decide whether to remove a particular event (Zhuang et al. 2002; Vere-Jones 1970); the "waveform similarity approach" (Barani et al. 2007); and others. See Davis and Frohlich (1991) and Zhuang et al. (2002) for taxonomies.

Mainshock-window methods remove the earthquakes in a space-time window around every "mainshock," suitably defined. Mainshock-window methods can be thought of as punching a hole in the catalog after each mainshock. The hole is the window. Gardner and Knopoff's windows (Knopoff and Gardner 1972; Gardner and Knopoff 1974) are common in mainshock-window declustering. They are larger in space and time the larger the shock is.

Linked-window methods calculate a space-time window for every event in the catalog, not just mainshocks. In linked-window methods, an event is in a *cluster* if and only if it falls within the window of at least one other event in that cluster. Linked-window declustering replaces each cluster with a single event—for instance, the first, the largest, or an "equivalent event." The most widely used linked-window method was developed by Reasenberg (1985). Reasenberg's windows are larger in space but shorter in time the larger the shock is.

Earthquake catalogs are often declustered using window methods as a precursor to modeling the remaining events as a realization of a spatially inhomogeneous, temporally homogeneous Poisson process (SITHP). Tests of the null hypothesis that declustered catalogs are realizations of a SITHP have not rejected the null hypothesis, leading some studies to conclude that declustered catalogs are Poisson. For instance, the title of Gardner and Knopoff (1974) is "Is the sequence of earthquakes in Southern California, with aftershocks removed, Poissonian?" The abstract: "Yes."

Their claim seems to be based on a *multinomial chi-square test* described below in section 3.1. The assumptions of the multinomial chi-square test are false when the SITHP hypothesis is true. Moreover, not rejecting the null hypothesis does not imply that the null hypothesis is true: Failure to reject could be a Type II error, especially if the test has little power against plausible alternatives—which we show is the case.

No test has good power against every alternative, so we compare the multinomial chi-square test with several other tests of the Poisson hypothesis: conditional chi-square, Brown-Zhao, and Kolmogorov-Smirnov, described in sections 3.2, 3.3, and 3.4. Among these, the Kolmogorov-Smirnov test is most sensitive to long-term variations in the rate of seismicity. The multinomial chi-square (with P-value estimated by simulation rather than the chi-square approximation) is most sensitive to local departures from Poisson behavior—but Poisson behavior and constant rate are not the same thing, nor is "Poisson behavior" necessarily a good proxy for unpredictability. The conditional chi-square and Brown-Zhao tests are similar and are more sensitive to variations in the short-term rate of seismicity. An omnibus test that combines these four temporal tests using Bonferroni's inequality rejects the hypothesis that the 1932–1971 Southern California Earthquake Center (SCEC) catalog of events with magnitude 3.8 and above, declustered with four window methods, follow a SITHP. (In contrast, Shearer and Stark (2012) show that global catalogs of large events declustered using crude window methods pass essentially the same suite of temporal tests for SITHP behavior.)

A declustering method designed to pass these tests—*deTest*—can pass these four temporal tests and retain 50% to 80% more events in the SCEC catalog than methods that do not pass the tests. Finding the smallest number of events to remove from a catalog so that the rest pass a tests of SITHP is a combinatorially complex problem. Instead, deTest uses an inexpensive, "greedy" approach that still performs well compared with window methods. We do not propose deTest as a method for declustering catalogs for seismological purposes, only to show that it is not hard to pass tests for temporal Poisson behavior, and that spatial information is important.

The multinomial chi-square, conditional chi-square, Brown-Zhao, and Kolmogorov-Smirnov tests use only the times of events, not event locations. Section 5 presents a nonparametric permutation test that uses locations as well as times to test the hypothesis that event times in declustered catalogs are conditionally exchangeable given event locations. That hypothesis is implied by the hypothesis that declustered catalogs are a realization of a SITHP. The test, based on ideas in Romano (1988, 1989), generally finds small P-values for this weaker hypothesis. We believe this test is new to seismology.

2 THE POISSON NULL HYPOTHESIS

Consider a fixed spatial domain S and time interval $(0, T]$. In a spatially inhomogeneous, temporally heterogeneous Poisson process (SITHP) on the spatiotemporal domain $S \times (0, T]$, the number of events in disjoint subsets of $S \times (0, T]$ are independent Poisson random variables. Within any subset, the locations of the events are independent of the times of events. The space-time rate is the product of the in homogenous marginal spatial rate and the uniform temporal rate.

Let N denote the (random) number of events in a SITHP on $S \times (0, T]$. Denote the random locations and times of the N events by $\{(X_i, Y_i, T_i)\}_{i=1}^{N}$. The times between successive events are marginally independent and identically distributed (iid) exponential random variables. The number of events in any subset of S in disjoint subsets of $(0, T]$ are independent Poisson random variables with means proportional to the durations of the intervals; this is the basis of the multinomial chi-square test described in section 3.1. Conditional on $N = n$ (that is, given the number of events that actually occur), the times $\{T_i\}_{i=1}^{n}$ are (marginally) iid uniform random variables; this is the basis of the Kolmogorov-Smirnov test described in section 3.4. Since the conditional distribution of times given $N = n$ is iid uniform, the number of events in K equal-length disjoint time intervals whose union is $(0, T]$ is has a multinomial conditional joint distribution with equal category probabilities. This is the basis of the conditional chi-square test and the Brown-Zhao test.

In describing the tests below, we assume that we are given a declustered catalog with n events. The longitude, latitude, and time of the ith event are (x_i, y_i, t_i), $i = 1, \ldots, n$. Events are not necessarily in chronological order. We do not consider earthquake depths. We study the null hypothesis that the points $\{(x_i, y_i, t_i)\}_{i=1}^{n}$ are a realization of a SITHP.

3 TEMPORAL TESTS

3.1 The multinomial chi-square test (MC)

We believe that the chi-square test of the hypothesis that declustered catalogs are realizations of a homogeneous temporal Poisson process used by Gardner and Knopoff (1974) and Barani et al. (2007) was a *multinomial chi-square test* (Brown and Zhao 2002). It works as follows:

(i) Pick $K \geq 1$. Partition the study period into K disjoint time intervals of length T/K. Count the events in each interval:

$$N_k \equiv \#\{i : t_i \in ((k-1)T/K, kT/K]\}, \quad k \in \{1, \ldots, K\}. \tag{1}$$

(ii) Estimate the theoretical rate of events per interval. We believe Gardner and Knopoff (1974) and Barani et al. (2007) used the estimate

$$\hat{\lambda} = n/K. \tag{2}$$

(iii) Pick $C \geq 2$, the number of "categories," such that the expected number of intervals that fall into each category, assuming that events follow a Poisson process with rate $\hat{\lambda}$, is at least 5. That is, choose the smallest integer C such that

$$E_c \geq 5 \quad \forall c \in \{0, \ldots, C-1\},^{\star} \tag{3}$$

where

$$E_c \equiv \begin{cases} Ke^{-\hat{\lambda}} \cdot \frac{\hat{\lambda}^c}{c!}, & c = 0, 1, \ldots, C-2 \\ K - \sum_{j=0}^{C-2} E_j, & c = C-1. \end{cases} \tag{4}$$

For $k \in \{1, \ldots, K\}$, interval k is in category $c \in \{0, \ldots, C-2\}$ if it contains c events; interval k is in category $C-1$ if it contains $C-1$ or more events. Let O_c be the number of intervals observed to be in category c.

(iv) Calculate the chi-square statistic:

$$\chi_m^2 \equiv \sum_{c=0}^{C-1} \frac{(O_c - E_c)^2}{E_c}. \tag{5}$$

Take the nominal P-value to be

$$P \equiv \Pr\{X \geq \chi_m^2\}, \tag{6}$$

where X is a random variable with a chi-square distribution with d degrees of freedom. We believe that Gardner and Knopoff (1974) used $d = C - 2$.

The nominal and true P-values depend on arbitrary choices: K, C, d, and the method of estimating λ. Moreover, the true P-value depends on whether these choices are made before or after looking at the data.

Let $I_k = c$ if the number of events in the kth interval is in category c. In the basic chi-square test for goodness of fit, C is fixed before observing the data, and the null hypothesis is that (i) $\{\Pr\{I_k = c\}\}_{c=0}^{C-1}$ are known and do not depend on k, and (ii) $\{I_k\}_{k=1}^{K}$ are independent (Lehmann 2005). The numbers of intervals in the C categories, $\{O_c\}_{c=0}^{C-1}$, then have a multinomial joint distribution. The null distribution of the chi-square statistic converges to a chi-square distribution with $C-1$ degrees of freedom as the number K of data increases—but the finite-sample distribution is only approximately chi-squared.

In testing whether declustered catalogs are Poisson, C generally is not fixed ahead of time: It is chosen after looking at the data to estimate λ, for instance, so that the expected number of intervals in each category exceeds some minimum, such as 5. Moreover, neither (i) nor (ii) is true in testing whether declustered catalogs are Poisson. (i) is false because the hypothesis that declustered seismicity is Poisson does not completely specify the category probabilities $\{\Pr\{I_k = c\}\}_{c=0}^{C-1}$. Instead, those probabilities are estimated from an estimate $\hat{\lambda}$ of the marginal temporal rate λ of the Poisson process. Estimating $\{\Pr\{I_k = c\}\}_{c=0}^{C-1}$ from the data changes the distribution of the chi-square statistic; moreover, the theoretical value of those probabilities conditional on the observed rate is different from the values used in practice, which are (estimated) unconditional probabilities.

(ii) is false too: Conditional on the estimated temporal rate, the random variables $\{I_k\}_{k=1}^{K}$ are not independent because they are related through the total number of earthquakes—an ingredient in estimating the rate. For instance, if $n \geq C-1$ and $I_k = 0$, $k = 1, \ldots, K-1$, we would know that $I_K = C-1$. The joint distribution of $\{O_c\}_{c=0}^{C-1}$ is not multinomial when the Poisson null hypothesis is true, and the chi-squared statistic, as calculated to test declustered catalogs, may not have even approximately a chi-square distribution.

Hence, we calibrate the P-value for the multinomial chi-square test using a simulation that takes into account estimating the rate of events from the data, choosing C on the basis of that estimate, and calculating the category probabilities in an inconsistent way (using the observed number of events to estimate the probabilities, but ignoring that conditioning in computing the probabilities). The simulation conditions on the observed number of events, and takes the times of the events to be independent, identically distributed (iid) uniform random variables.

3.2 The conditional chi-square test (CC)

The MC test described in the previous subsection assesses whether the numbers of intervals with various numbers of events agree well with the numbers expected for iid uniformly distributed event times. MC uses C categories of possible values of the number of events per interval. An interval with $C-1$ events is in the same category as one with $C+10$ events. For this and other reasons, MC is not as sensitive to overdispersion—apparent fluctuations in the rate of seismicity—as some other tests.

The *conditional chi-square test* (or *Poisson dispersion test*) uses the fact that, conditional on the total number of events, the joint distribution of the numbers of events in the windows is multinomial with equal category probabilities. The test statistic is

$$\chi_c^2 \equiv \sum_{k=1}^{K} \frac{(N_k - \hat{\lambda})^2}{\hat{\lambda}}. \tag{7}$$

This is proportional to the variance of the counts across windows. If the Poisson hypothesis is true, the distribution of χ_c^2 is approximately chi-square with $K-1$ degrees of freedom. The conditional chi-square test involves choosing the

[⋆] We assume that $E_0 \geq 5$. If not, categories might be combined so that the expected number of intervals in each category is at least 5. See, for instance, Shearer and Stark (2012). The reason to have a lower threshold of 5 is that many textbooks state that the chi-square approximation to the null distribution of the chi-square statistic holds when the expected number of counts in every category is at least 5, so we suspect that seismologists generally use this rule to select the categories. Note that this method of selecting the categories for the chi-square test makes the number of categories and their definitions depend on the observed data through $\hat{\lambda}$. This can make the actual P-value differ from the nominal P-value based on the chi-square distribution. We compare the nominal P-value with the P-value estimated by simulation, which takes into account the conditioning and does not rely on the chi-square approximation.

number of intervals K but not C, and, unlike the multinomial chi-square test, it uses the information $N = n$ in a consistent way. While the multinomial chi-square test tries to look at the detailed distribution of the number of events per interval, the conditional chi-square test looks only at the variability of the observed number of events across intervals. High variability—overdispersion—is a sign that the process is not a homogeneous Poisson process.

3.3 The Brown-Zhao test (BZ)

Brown and Zhao (2002) proposed an alternative test. Let $Y_k \equiv \sqrt{N_k + 3/8}$ and $\bar{Y} \equiv \sum Y_k/K$. The Brown-Zhao (BZ) test statistic is

$$\chi^2_{BZ} \equiv 4 \sum_{k=1}^{K} (Y_k - \bar{Y})^2. \quad (8)$$

Under the Poisson hypothesis, the statistic χ^2_{BZ} has a distribution that is approximately chi-square with $K - 1$ degrees of freedom. The chi-square approximation to the null distribution of χ^2_{BZ} tends to be better than the chi-square approximation to the null distribution of χ^2_c or χ^2_m (Brown and Zhao 2002). Like CC, the BZ test requires choosing K but not C, and uses the information $N = n$ in a consistent way. Also like CC, the BZ test rejects when there is high variability of the observed numbers of events in different intervals—i.e., when the counts are overdispersed.

3.4 The Kolmogorov-Smirnov test (KS)

The Kolmogorov-Smirnov test compares the empirical cumulative distribution function (cdf) $\hat{F}_n(x)$ of a random variable to a fixed reference cumulative distribution function $F(x)$ (Lehmann 2005). The KS test rejects when

$$D_n \equiv \sup_x \left| \hat{F}_n(x) - F(x) \right| \geq C(n, \alpha). \quad (9)$$

In seismology, the KS test has been used to assess the uniformity of declustered earthquake sequences preceding mainshocks (Matthews and Reasenberg 1988; Reasenberg and Matthews 1988).

If declustered earthquakes follow a SITHP, then conditional on $N = n$, the times $\{T_i\}_{i=1}^n$ are iid uniform on $(0, T]$. Their common cumulative distribution function is $F(x) = t/T$. Hence, conditional on $N = n$,

$$D_n = \sup_t \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{t_i \leq t} - t/T \right|. \quad (10)$$

Unlike the chi-square tests, the KS test has no ad hoc choices analogous to K, C, and d. The KS test has asymptotic power 1 against the alternative that the data are iid with any fixed distribution $G \neq F$.

3.5 Power

MC, CC, and BZ ignore the order of the K intervals. This causes them to have low power against some kinds of clustering that violate the Poisson hypothesis, such as long-term variations in the rate of seismicity.

To see why, consider any particular catalog. Divide the study period into K disjoint equal-length windows. Now, rearrange the windows so that the first has the most events, the second has the next most events, and so on, to form a new catalog. This catalog has a monotonically decreasing rate of seismicity, which would be very unlikely if declustered seismicity followed a homogeneous Poisson process. However, the test statistics for MC, CC, and BZ would have the same values for this rearranged catalog as they did for the original catalog.

In contrast, KS would tend to reject the null hypothesis for the new data: The empirical cdf would be far above t/T in the early part of the rearranged catalog. KS is more sensitive to long-term rate variations than the other tests are, but less sensitive to short-term variations. KS, CC, and BZ are sensitive to whether the rate varies with time: to clustering. For instance, if events were equispaced in time, none of those tests would reject the null hypothesis. MC would reject the Poisson hypothesis if events were equispaced, given enough data.

None of these tests uses spatial information; in section 5 we propose a test that does. A process can be non-Poisson in space-time yet have a homogenous Poisson marginal temporal distribution; see, e.g., Luen (2010, Chapter 3), so using spatial information can increase power, as we show empirically.

4 DECLUSTERING METHODS

We consider the following five window-based declustering algorithms:

GKl (Gardner-Knopoff linked) Remove every event that is in the window of some other event (Gardner and Knopoff 1974).

GKlb (Gardner-Knopoff linked, biggest) Divide the catalog into clusters as follows: An event is in a given cluster if and only if it is in the window of at least one other event in the cluster. In every cluster, remove all events except the largest (Gardner and Knopoff 1974).

GKm (Gardner-Knopoff mainshock) Consider the events in chronological order. If the ith event is in the window of a preceding larger shock that has not already been deleted, delete it. If a larger shock is in the window of the ith event, delete the ith event. Otherwise, retain the ith event (Knopoff and Gardner 1972).

Rl (Reasenberg linked) Reasenberg's method (Reasenberg 1985).

dT deTest, described below.

GKl, GKlb, and Rl are linked-window methods. Gardner and Knopoff (1974) found that GKl and GKlb gave similar results for 1932–1971 Southern California seismicity. GKm is a mainshock-window method.

deTest is not a window method. It has no physical basis, not even a heuristic one. It is not intended to be a method for declustering catalogs for seismological purposes, only a "straw man" to show two things:

(i) A declustered catalog can have rather more events than window-based declustering methods leave, and still pass a test for temporally homogeneous Poisson behavior.

(ii) Using spatial and temporal data by testing for conditional exchangeability of times given the locations (described below) can be more powerful than testing only for temporal homogeneity.

We assume that K is given. Declustering a catalog to make the result pass the MC, CC, or BZ test is constrained by the number of intervals among the K in the original catalog that have no events, since declustering can delete events but not add them. The number of intervals with no events gives an implicit estimate of the rate of a Poisson process that the declustered catalog can be coerced to fit well: If seismicity followed a homogenous Poisson process with theoretical rate λ events per interval, the chance that an interval would contain no events is $e^{-\lambda}$. So,

$$\hat{\lambda} \equiv -\log\left(\frac{\#\text{ intervals with no events}}{K}\right) \tag{11}$$

is a natural estimate of the rate of events per interval. deTest tries to construct a catalog with about this rate that passes all four tests described above (MC, CC, BZ, and KS). If seismicity followed a homogeneous Poisson distribution with theoretical rate per interval $\hat{\lambda}$, then the expected number of intervals in the declustered catalog with at least c events would be

$$G_c \equiv \sum_{i=c}^{\infty} K \times \frac{\hat{\lambda}^i e^{-\hat{\lambda}}}{i!}. \tag{12}$$

deTest constructs a catalog in which $[G_c]$ intervals contain c or more events, where $[x]$ denotes the integer closest to x.

deTest is defined by the following algorithm, which starts with an empty catalog and adds events from the original catalog until the result has approximately the correct expected number of intervals with each number of events (ensuring that it will pass the first three tests), then removes events until the catalog passes the KS test.

(i) Count the events in the raw catalog in each of the K intervals.

(ii) Define $\hat{\lambda}$ by equation 11 and G_c by equation 12.

(iii) Let $c = 1$. From each interval in the raw catalog that has at least one event, include one event selected at random from that interval.

(iv) Let $c \leftarrow c+1$. If $[G_c] = 0$, go to step (vi). Otherwise, go to step (v).

(v) This step adds events to the declustered catalog until $[G_c]$ intervals have at least c events, while trying to keep the KS statistic small. Let N_t be the number of events in the current declustered catalog that have occurred by time t. Find the element t_m of the set $t \in \{T/K, 2T/K, \ldots, T\}$ at which $N_t/N_T - t/T$ is minimized. Adding an event before time t_m will tend to reduce the KS statistic. Find the set of intervals that

a. contain $c-1$ events in the current declustered catalog;

b. contain at least c events in the raw catalog.

From this set, select the interval prior to t_m but closest to t_m (If no interval is prior to t_m, choose the first interval in the set.) Choose an event at random from the events in the selected interval that have not yet been added to the declustered catalog, and add that event to the declustered catalog. Repeat this step until $[G_c]$ intervals contain c events.[†] Return to step (iv).

(vi) Find the KS P-value. If it is above the target significance level, find a time t at which the empirical cdf differs maximally from the uniform cdf. Either t is infinitesimally before an event or t is the time of one or more events. If t is just before an event, t/T is larger than the empirical cdf. In that case, delete the event after time t at which the empirical cdf minus the uniform cdf is largest. If t is the time of an event, the empirical cdf at t is larger than t/T. In that case, delete an event at time t. Repeat this step until the KS P-value is below 0.05.

5 SPATIO-TEMPORAL TESTS

5.1 A weaker null hypothesis: conditionally exchangeable times

The marginal distribution of event times for a SITHP is Poisson, so if the hypothesis that declustered event times follow a Poisson distribution is rejected, so is the hypothesis that event times and locations follow a SITHP. Moreover, SITHPs can have events arbitrarily close together. But catalogs declustered with window methods have a minimum spacing between events: If a catalog contains two events very close in space and time, the later event will fall within the window of the former, and one or both of them will be deleted. However, catalogs declustered using window methods may still have some properties of SITHPs.

To try to salvage part of the SITHP hypothesis, we develop a test of a weaker condition implied by SITHP: the hypothesis that times are *conditionally exchangeable* given event locations. Let Π be the set of all $n!$ permutations of $\{1, \ldots, n\}$. We say a process has *conditionally exchangeable times* if, conditional on the locations,

$$\{T_1, \ldots, T_n\} \overset{d}{=} \{T_{\pi(1)}, \ldots, T_{\pi(n)}\} \tag{13}$$

for all permutations $\pi \in \Pi$. (The notation $\overset{d}{=}$ means "has the same probability distribution as.") If event times are conditionally iid given event locations, they are conditionally exchangeable given event locations. Since event times in SITHPs are conditionally iid uniform given event locations, SITHPs have conditionally exchangeable times given event locations.

Under the hypothesis of conditionally exchangeable times, conditional on the set of locations $\{(x_i, y_i)\}_{i=1}^n$ and, separately, on the set of times $\{t_i\}_{i=1}^n$, all one-to-one assignments of times to locations have the same chance. If events close in space tend to be close in time—the kind of clustering real seismicity exhibits—times are not conditionally exchangeable. If events close in space tend to be distant in time—which can result from deleting events in windows—times are not conditionally exchangeable.

We test the hypothesis that times are conditionally

[†] To ensure that there are at least $[G_c]$ intervals that contain c events could in principle require modifying the rule for deciding which intervals to include at each stage, so that things "telescope" correctly. In practice, we have not found it necessary to complicate the algorithm in that way.

exchangeable by adapting abstract methodology of Romano (1988, 1989). Let $\hat{P}_n$ be the empirical distribution of the times and locations of the n observed events: $\hat{P}_n$ assigns probability $1/n$ to each observed (time, location) pair $((x_i, y_i), t_i)$. For each permutation π of $\{1, \ldots, n\}$, let $\hat{P}_{\pi n}$ be the distribution that assigns probability $1/n$ to each pair $((x_i, y_i), t_{\pi(i)})$.

If the null hypothesis of conditionally exchangeable times holds, then conditional on the times and (separately) the locations, the empirical distribution was just as likely to have been $\hat{P}_{\pi n}$ as it was to be $\hat{P}_n$. Consider a test statistic ϕ that can be computed from the empirical distribution of the data. Such a statistic is called a *functional statistic*. If the null hypothesis is true, then conditional on the times and the locations, all values of $\phi(\hat{P}_{\pi n})$ as π varies over the $n!$ permutations of $\{1, \ldots, n\}$ were equally likely. We can test hypotheses (conditional on the times and the locations) by determining whether $\phi(\hat{P}_n)$ is surprisingly large compared to those $n!$ values. If $\phi(\hat{P}_{\pi n}) \geq \phi(\hat{P}_n)$ for a fraction P of the $n!$ permutations, then the P-value of the null hypothesis is P.

We now define the functional test statistic ϕ we will use. It measures the "distance" between the empirical distribution $\hat{P}_n$ and a transformation $\tau \hat{P}_n$ of the empirical distribution that satisfies the null hypothesis by construction. In particular, we take $\tau \hat{P}_n$ to be the distribution that assign probability $1/n^2$ to the n^2 pairs $((x_i, y_i), t_j)_{i,j=1}^n$. For $\tau \hat{P}_n$, times and locations are independent, and hence times are conditionally exchangeable. Thus $\tau \hat{P}_n$ satisfies the null hypothesis.

We now define the measure of "distance" that ϕ uses. A set $V \subset R^3$ is a *lower-left quadrant* if, for some (x_0, y_0, t_0), it is of the form:

$$V = \{(x, y, t) \in R^3 : x \leq x_0 \text{ and } y \leq y_0 \text{ and } t \leq t_0\}. \quad (14)$$

Let $\mathbf{V}$ be the set of all lower-left quadrants. The test statistic ϕ is the supremum (over all lower-left quadrants $V \in \mathbf{V}$) of the difference between the probability $\hat{P}_n$ assigns to V and the probability that $\tau \hat{P}_n$ assigns to V:

$$\phi(\hat{P}_n) \equiv \sup_{V \in \mathbf{V}} |\hat{P}_n(V) - (\tau \hat{P}_n)(V)|. \quad (15)$$

In principle, one could enumerate all $n!$ distributions $\hat{P}_{\pi n}$, calculate the $n!$ values of $\phi(\hat{P}_{\pi n})$, and find the fraction of such values that are greater than or equal to $\phi(\hat{P}_n)$ to determine the P-value. But since $n!$ is enormous, it is more practical to estimate the P-value by comparing $\phi(\hat{P}_n)$ to the values of $\phi(\hat{P}_{\pi n})$ for a large random sample of permutations π. The larger the number of random permutations, the smaller the variability of the estimated P-value. Alternatively, it is still conservative simply to redefine the test to examine only a smaller, pre-determined subset of permutations, since in any such subset all the permutations are equally likely, conditional on the times and on the locations. The P-value for that test is the fraction of that subset of permutations for which the test statistic is at least as large as $\phi(\hat{P}_n)$.

We implemented a test for conditionally exchangeable times in R (`http://cran.r-project.org/`). Code is available online at `http://statistics.berkeley.edu/~stark/Code/Quake/permutest.r`. Appendix A describes the algorithm.

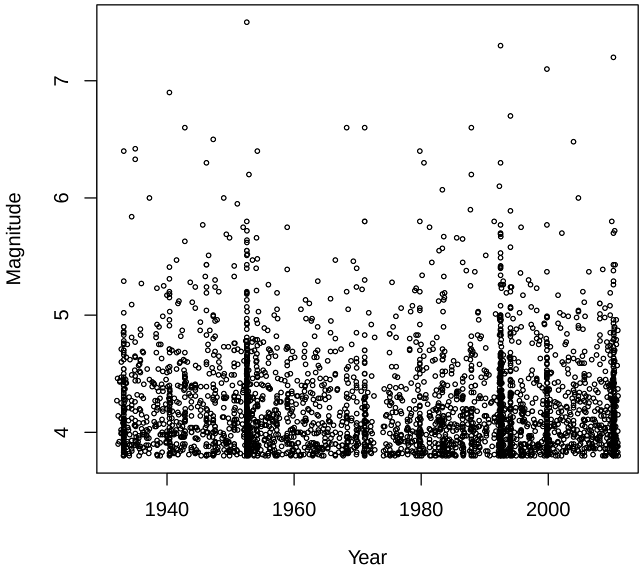


Figure 1. Time versus magnitude plot of the 3,368 events with magnitude 3.8 and larger in the 1932–2010 SCEC catalog of Southern California events.

6 DATA AND RESULTS

We assessed whether four subsets of the SCEC catalog[‡] (1932–1971 and 1932–2010, each for $M \geq 3.8$ and $M \geq 4.0$), declustered using the five methods in section 4, are consistent with SITHP. The 1923–1971, $M \geq 3.8$ subset was chosen to be as similar as possible to the catalog used by Gardner and Knopoff (1974), for comparison with their results.

Gardner and Knopoff (1974) performed multinomial chi-square tests on a number of catalogs declustered using GKl. Among other things, they report results for a catalog of earthquakes with $M \geq 3.8$ occurring in the "Southern California Local Area" from 1932–1971. That raw catalog had 1,751 events; the declustered catalog had 503 events. They divided the forty-year period into ten-day intervals, found O_c and estimated E_c for some range of c, and found nominal P-values using the chi-square distribution with 2 degrees of freedom. They did not state C, how they estimated λ, nor whether they used $d = C - 1$ or $d = C - 2$ in their tests. They found a P-value of 0.0599, and hence did not reject the hypothesis that declustered catalogs are Poisson at significance level 0.05.

The SCEC catalog contains 3,368 events with magnitude at least 3.8 between 1932 and 2010, of which 1,556 occurred between 1932 and 1971. Figure 1 shows a time-magnitude plot for these events. We declustered that catalog using GKl, GKlb, and GKm with the Gardner and Knopoff (1974) windows and using Rl and dT. (We used

[‡] `http://www.data.scec.org/eq-catalogs/date_mag_loc.php` (last accessed 23 September 2011).

Stefan Wiemer's ZMAP package for MATLAB[§] to apply Reasenberg's method, with the default parameter values taumin=1 day, taumax=10 days, P1= 0.95, xk= 0.5, xmeff=1.5, rfact= 10, epicentral error= 1.5 km, and depth error= 2 km.) The declustered catalogs contained 437, 424, 544, 985, and 608 events, respectively. Figure 2 shows cumulative frequency plots for the original and declustered catalogs for 1932–2010. Figure 3 maps the events in the original catalog and the events that remain after declustering using each of the methods. Figure 4 shows the 1932–2010 SCEC catalog of 3,368 events of magnitude 3.8 and above, declustered using the same five methods. Those declustered catalogs contained 913, 892, 1,120, 2,046, and 1,615 events, respectively.

We applied MC (using both the χ^2 approximation and simulation to approximate the null distribution), CC, BZ, and KS to the declustered catalogs. We combined these four tests (using the simulation P-value rather than the χ^2 approximation) to obtain a composite level 0.05 temporal test of the SITHP hypothesis, using Bonferroni's equality: We rejected the SITHP hypothesis if any of these four tests gave a P-value less than 0.0125. If the null hypothesis is true, the chance of a Type I error is no greater than $4 \times 0.0125 = 0.05$.

Results, reported in Table 1, varied. For 1932–1971, $M \geq 3.8$, the catalog most similar to that studied by Gardner and Knopoff (1974), none of the window-declustered catalogs appears to be Poisson, contradicting Gardner and Knopoff (1974). For the four window declustering methods, the KS test rejects the Poisson hypothesis at level 0.0125; the KS test does not reject the Poisson hypothesis for deTest. The other tests reject the Poisson hypothesis at level 0.0125 for GKm and Rl. For 1932–1971 and 1932–2010 $M \geq 4.0$, the Poisson hypothesis is rejected for GKm and Rl. For 1932–2010 $M \geq 3.8$, the Poisson hypothesis is rejected for all methods except deTest.

Table 1 also gives results for the permutation test of the hypothesis that event times are conditionally exchangeable given event locations. As discussed above, this hypothesis is weaker than SITHP; nonetheless, incorporating spatial information can lead to more power to reject the SITHP hypothesis when that hypothesis is in fact false. This is evident in the results for deTest. deTest only tries to pass the temporal tests, which it succeeds in doing for all four catalogs, despite the fact that it retains more events than all the other methods but Reasenberg's. Unsurprisingly, it fails the spatio-temporal test for all four catalogs: The spatio-temporal behavior of catalogs declustered by deTest is not consistent with the hypothesis that times are conditionally exchangeable. Of course, one could devise an analog of deTest to produce declustered catalogs that pass the permutation test; we have not tried.

7 DISCUSSION

Conclusions about whether declustered catalogs are consistent with the SITHP hypothesis depend not only on the declustering method but also on the catalog—and on the statistical test. The multinomial chi-square test commonly used to assess whether declustered catalogs have Poisson temporal behavior relies on ad hoc tuning constants, lacks theoretical justification, can have a significance level larger than its nominal significance level, and has low power against many plausible alternatives. Comparing the nominal P-value with P-values (conditional on the number of events) estimated by simulation using SCEC data (Table 1) shows that the χ^2 approximation to the P-value can be too low by at least 2.4% in seismological applications. The multinomial chi-square test is sensitive to departures from Poisson behavior within intervals, but not to clustering *per se*. In particular, the multinomial chi-square test discards information about the time order of the intervals, which reduces its power to detect long-term rate variations.

Compared with the multinomial chi-square test, the conditional chi-square test and the Brown-Zhao test have better theoretical justification and fewer ad hoc tuning constants, but they still require an arbitrary choice of the number of intervals into which to divide the study period. Both tests condition on the number of events; they test whether the conditional distribution of times given the number of events is iid uniform. Both are sensitive to variation of the observed rate of events across intervals—to clustering on the scale of the intervals. Although it might be more powerful against some alternatives, for the data we studied, the Brown-Zhao test never gave a smaller P-value than the conditional chi-square test, to which it is closely related. The primary advantage of the Brown-Zhao test over the conditional chi-square test in this application seems to be that the chi-square approximation to the distribution of the test statistic is more accurate than it is for the conditional chi-square test. This might not matter much since the P-values can be estimated by simulation regardless.

The Kolmogorov-Smirnov test of the Poisson hypothesis also conditions on the number of events and tests whether times are conditionally iid uniform. In contrast to the other three temporal tests, it has no ad hoc tuning constants and—because keeps the entire observation period intact—has more power against long-term rate variations than the other three tests, which divide the study period into shorter intervals and ignore the temporal order of the shorter intervals. This is evident in Table 1: P-values for the Kolmogorov-Smirnov test are most often the smallest. However, the tests are to some extent complementary: the chi-square tests sometimes give small P-values when the Kolmogorov-Smirnov test does not. The conditional chi-square test seems preferable to the multinomial chi-square test in that it has a firmer theoretical foundation and requires fewer ad hoc choices, although it does not have power against some of the same alternatives, for instance, periodic or nearly periodic seismicity. (Given enough data, the multinomial chi-square test will reject the null hypothesis if events are nearly equispaced, but the conditional chi-square test, the Brown-Zhao test, and the Kolmogorov-Smirnov test will not.) Using the Kolmogorov-Smirnov test in conjunction with the conditional chi-square test and combining the results using Bonferroni's inequality seems like a good compromise. That is, if one wishes to test at significance level α, reject the null hypothesis if either test has a P-value less than $\alpha/2$.

Such a composite test shows that 1932–1971 SCEC seis-

[§] `http://www.earthquake.ethz.ch/software/zmap` (last accessed 23 September 2011).

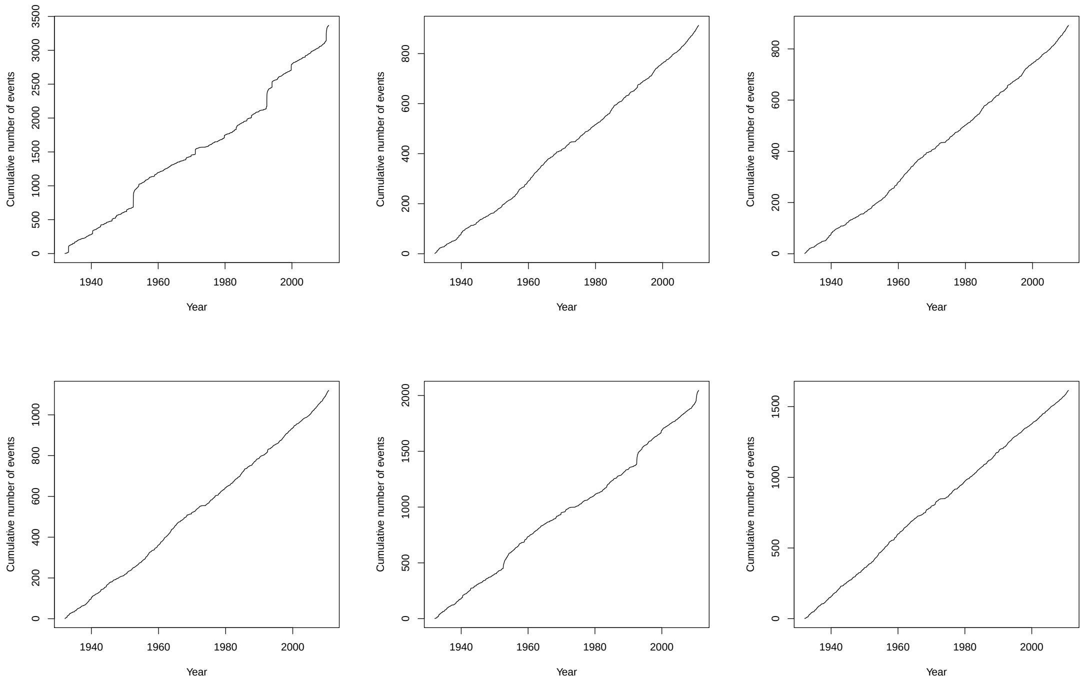
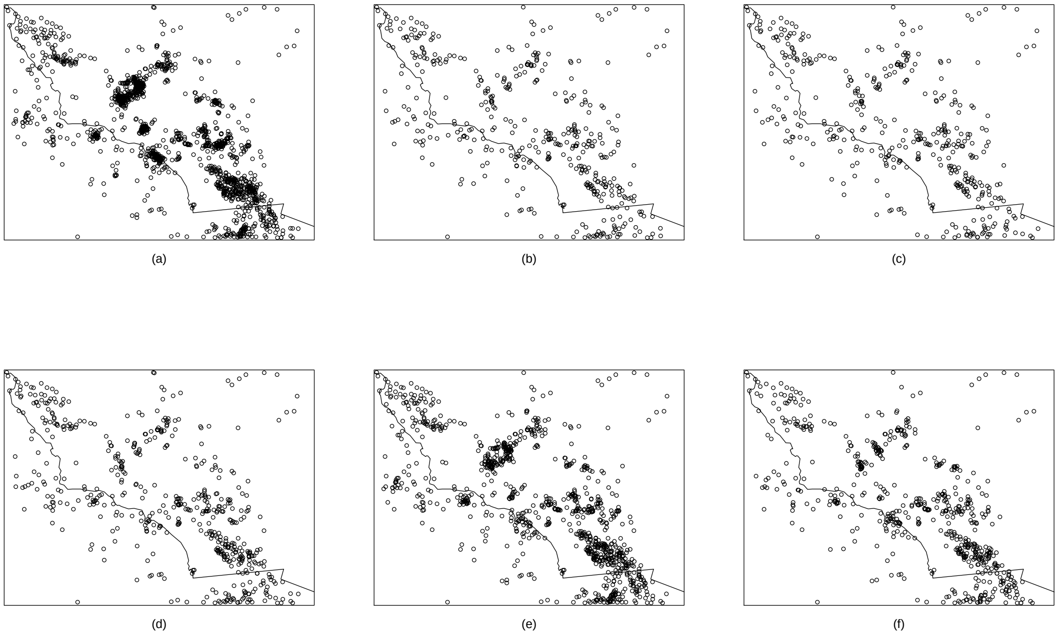


Figure 2. Cumulative temporal distribution of events in the 1932–2010 before and after declustering. (a): All 3,368 events with magnitude 3.8 (b): The 913 events that remain after declustering using GKl. (c): The 892 events that remain after declustering using GKlb. (d): The 1,120 events that remain after declustering using GKm. (e): The 2,046 events that remain after declustering using Rl. (f): The 1,615 events that remain after declustering using dT.


(a)

(b)

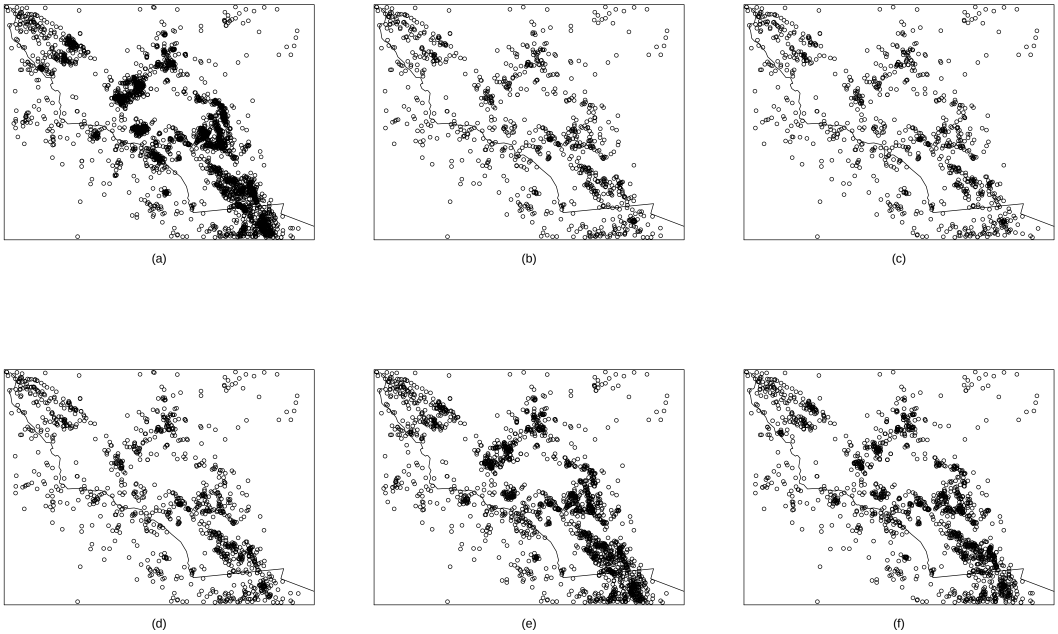
(c)

(d)

(e)

(f)

Figure 3. (a): 1932–1971 SCEC catalog of 1,556 events of magnitude 3.8 or greater in Southern California. (b): The 437 events that remain after declustering using GKl. (c): The 424 events that remain after declustering using GKlb. (d): The 544 events that remain after declustering using GKm. (e): The 985 events that remain after declustering using Rl. (f): The 608 events that remain after declustering using dT.


(a) (b) (c)

(d) (e) (f)

Figure 4. (a): 1932–2010 SCEC catalog of 3,368 events of magnitude 3.8 or greater in Southern California. (b): The 913 events that remain after declustering using GKl. (c): The 892 events that remain after declustering using GKlb. (d): The 1,120 events that remain after declustering using GKm. (e): The 2,046 events that remain after declustering using Rl. (f): The 1,615 events that remain after declustering using dT.

micity with $M \geq 3.8$, declustered using standard window methods, is not consistent with the Poisson hypothesis. The opposite conclusion by Gardner and Knopoff (1974) seems to have resulted from their choice of tests: the multinomial chi-square. It is surprising that the 1932–2010 SCEC data declustered using Gardner-Knopoff windows is more consistent with the Poisson hypothesis, since the Gardner-Knopoff method was derived for the earlier data.

Moreover, it is hard to explain why increasing the threshold magnitude from 3.8 to 4.0 makes as much difference as it does. It would be expected to increase P-values somewhat simply because it reduces sample size, but that does not appear to be all that is at play: The Kolmogorov-Smirnov test seems to reject the Poisson hypothesis because the rate of small events is too low in the earlier part of the catalog. This might be explained by catalog incompleteness—that events of magnitude 3.8–4.0 are more often missing from the earlier catalog—but according to Hutton et al. (2010) the SCEC catalog has been essentially complete above magnitude 3.25 from its earliest days. The accuracy of magnitude and location estimates in the early part of the catalog might contribute to the difference.

All four of these tests—multinomial chi-square, conditional chi-square, Brown-Zhao, and Kolmogorov-Smirnov—condition on the total number of events. None uses spatial information, and it is the spatio-temporal distribution of seismicity that matters. A test that uses spatial information could be much more powerful against some alternatives. In a spatially inhomogeneous, temporally homogeneous Poisson process (SITHP), two events may occur arbitrarily close to one another with strictly positive probability. Catalogs declustered using window methods can never have events very close in space and time.

But declustered catalogs may still have properties in common with SITHP. For instance, the times might be conditionally exchangeable given the locations. As a special case, knowing the location of an event might give no information about the time of the event. A novel permutation test can be used to assess whether event times are conditionally exchangeable given event locations. The power of incorporating spatial information is evident in the fact that catalogs declustered using deTest pass all the temporal tests, but fail the spatio-temporal test for exchangeability. (We do not seriously propose deTest as a method for declustering catalogs for seismological purposes, only as a "straw man" to show that spatial information matters.)

"Ok, so why do you decluster the catalog?" asks the online FAQ for the Earthquake Probability Mapping Application of the USGS.[¶] The answers: "to get the best possible estimate for the rate of mainshocks," and "the methodology [of the Earthquake Probability Mapping Application] requires a catalog of independent events (Poisson model), and declustering helps to achieve independence." The evidence presented here suggests that appropriate statistical tests can easily distinguish regional catalogs declustered using window methods from "independent events" that follow

[¶] `http://earthquake.usgs.gov/learn/faq/?faqID=280`, last accessed 29 September 2011.

Years	Mag (events)	Meth	n	MC		CC	BZ	KS	Romano		Reject?	
				χ^2	Sim				P	CI	Time	Space-time
1932–1971	3.8 (1,556)	GKl	437	0.087	0.089	0.069	0.096	0.011	0.005	[0.003, 0.007]	Yes	Yes
		GKlb	424	0.636	0.656	0.064	0.108	0.006	0.000	[0.000, 0.001]	Yes	Yes
		GKm	544	0	0	0	0	0.021	0.069	[0.063, 0.076]	Yes	No
		Rl	985	0	0	0	0	0.003	0	[0.000, 0.001]	Yes	Yes
		dT	608	0.351	0.353	0.482	0.618	0.054	0.001	[0.000, 0.006]	No	Yes
	4.0 (1,047)	GKl	296	0.809	0.824	0.304	0.344	0.562	0.348	[0.318, 0.378]	No	No
		GKlb	286	0.903	0.927	0.364	0.385	0.470	0.452	[0.421, 0.483]	No	No
		GKm	369	<0.001	<0.001	0	0	0.540	0.504	[0.473, 0.535]	Yes	No
		Rl	659	0	0	0	0	0.001	0	[0.000, 0.004]	Yes	Yes
		dT	417	0.138	0.134	0.248	0.402	0.051	0	[0.000, 0.004]	No	Yes
1932–2010	3.8 (3,368)	GKl	913	0.815	0.817	0.080	0.197	0.011	0.214	[0.189, 0.241]	Yes	No
		GKlb	892	0.855	0.855	0.141	0.204	0.005	0.256	[0.229, 0.284]	Yes	No
		GKm	1120	0	0	0	0	0.032	0.006	[0.002, 0.013]	Yes	Yes
		Rl	2046	0	0	0	0	0	0	[0.000, 0.004]	Yes	Yes
		dT	1615	0.999	1.000	0.463	0.466	0.439	0	[0.000, 0.004]	No	Yes
	4.0 (2,169)	GKl	606	0.419	0.421	0.347	0.529	0.138	0.247	[0.221, 0.275]	No	No
		GKlb	592	0.758	0.768	0.442	0.500	0.137	0.251	[0.224, 0.279]	No	No
		GKm	739	0	0	0	0	0.252	0.023	[0.015, 0.034]	Yes	Yes
		Rl	1333	0	0	0	0	0	0	[0.000, 0.004]	Yes	Yes
		dT	1049	0.995	0.999	0.463	0.465	0.340	0.001	[0.000, 0.006]	No	Yes

Table 1. *P*-values for tests of the null hypotheses that subsets of the 1932–2010 SCEC catalog of events, declustered using GKl, GKlb, GKm, Rl, and dT, have a homogeneous Poisson distribution in time or have a temporally homogeneous, spatially heterogeneous distribution in space and time. Column 1 gives the catalog year range. Column 2 is the magnitude threshold and the number of events before declustering. Column 3 is the declustering method. The number of events that remain after declustering is n. "χ^2" is the nominal *P*-value for a multinomial chi-square test using the chi-square approximation to the the null distribution of the test statistic. "Sim" is the *P*-value for a multinomial chi-square test estimated by simulation that includes conditioning on the observed number of events to estimate the rate of the process and to define the categories. "CC" is the *P*-value for the conditional chi-square test. "BZ" the the *P*-value for the method of Brown and Zhao (2002). Values in columns "Sim", "CC" and "BZ" are estimated using 10^5 simulated catalogs; sampling error in those estimated *P*-values are on the order of 0.16%. "KS" is the *P*-value for the Kolmogorov-Smirnov test that event times are iid uniform given the number of events. "Romano" is the permutation test for conditional exchangeability of times of events given their locations. Romano *P* is the *P*-value estimated from 1,000–10,000 simulations. Romano CI are confidence intervals for the Romano *P*-values based on the number of simulations performed in each case. "Reject" is "Yes" for "Time" if the simulation *P*-value for any of the four temporal tests is less than 0.0125 (using the simulation *P*-value rather than the χ^2 *P*-value for the MC test). "Reject" is "Yes" for "Space-time" if the *P*-value for the Romano test is less than 0.05.

a Poisson model, even approximately. In contrast, Shearer and Stark (2012) show that global catalogs of large events declustered using window methods are harder to distinguish from realizations of Poisson processes.

ACKNOWLEDGMENTS

We note with sadness that Leon Knopoff, a giant in geophysics who pioneered the methods assessed here, passed away this year. We are grateful to Steven N. Evans and Peter M. Shearer for helpful conversations, and to Yehuda Ben-Zion and anonymous referees for helpful comments.

APPENDIX A: ALGORITHM TO TEST THE HYPOTHESIS THAT TIMES ARE CONDITIONALLY EXCHANGEABLE

R code that implements the algorithm to test whether times are conditionally exchangeable is available at `http://statistics.berkeley/edu/~stark/Code/Quake/permutest.r`. The algorithm has the following steps:

(i) Sort the catalog of longitudes, latitudes, and times in time order. Label the sorted points $\{x_i, y_i, t_i\}$ for $i \in \{1, \ldots, n\}$. Find the longitude and latitude ranks of every event.

(ii) Find the empirical spatial measure of all lower-left quadrants in $\mathbf{R}^2$ with corners

$$(y_i, x_j), \quad 1 \le i, j \le n. \tag{A.1}$$

In the online R code, this spatial distribution is stored in the matrix `xy.upper`. The entry indexed by (i, j) is

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(y_i \le y, x_j \le x). \tag{A.2}$$

This is the number of events in the catalog with latitude less than the latitude of the ith event in the catalog and with longitude less than the longitude of the jth event in the catalog.

(iii) Find the absolute differences between the empirical measure and the empirical null measure for the n^3 lower-left quadrants with corners

$$(x_j, y_i, t_k), \quad i, j, k \in \{1, \ldots, n\}.$$

Find the maximum value of all these differences; this is the test statistic ϕ. To reduce storage requirements, the code finds the distances for quadrants with corners (x_j, y_i, t_k) for every value of k successively; that is, it finds

$$\phi = \max_k \left[\max_{j,i} \left| \hat{P}(V(j,i,k)) - (\tau \hat{P})(V(j,i,k)) \right| \right],$$

where $V(j,i,k)$ is the lower-left quadrant with corner (x_j, y_i, t_k).

(iv) Set the iteration counter h to 0.

(v) Increment h. Create a random permutation of $\{1, \ldots, n\}$. Apply this permutation to the locations, leaving times fixed. (One could permute the times instead of the locations, but that would require the re-sorting the catalog into temporal order after each permutation.) The spatial measure has not changed, but its indexing has; apply the permutation to both the rows and the columns of `xy.upper`.

(vi) As in step (iii), find the absolute differences between the empirical measure and the empirical null measure of the n^3 lower-left quadrants. Let ϕ_h be the maximum value of all these distances.

(vii) Determine whether to stop or to return to step (v). (We might simply stop when $h = 10,000$, or we might apply Wald's sequential probability ratio test (Wald 1945) to determine whether, on the basis of the random permutations taken so far, it is possible to conclude whether $P \leq \alpha$.) If the algorithm stops, estimate the P-value as

$$\hat{P} = \frac{1}{H} \#\{h : \phi_h \geq \phi\},$$

where H is the total number of iterations.

REFERENCES

S. Barani, G. Ferretti, M. Massa, and D. Spallarossa. The waveform similarity approach to identify dependent events in instrumental seismic catalogues. *Geophys. J. Intl.*, 168(1):100–108, 2007.

L.D. Brown and L.H. Zhao. A test for the Poisson distribution. *Sankhyā : The Indian Journal of Statistics*, 64:611–625, 2002.

S.D. Davis and C. Frohlich. Single-link cluster analysis of Earthquake aftershocks: Decay laws and regional variations. *J. Geophys. Res.*, 96:6335–6350, 1991.

J.K. Gardner and L. Knopoff. Is the sequence of earthquakes in Southern California, with aftershocks removed, Poissonian? *Bull. Seis. Soc. Am.*, 64(15):1363–1367, 1974.

K. Hutton, J. Woessner, and E. Hauksson. Earthquake monitoring in southern california for seventy-seven years (1932-2008). *Bulletin of the Seismological Society of America*, 100 (2):423–446, 2010. doi: 10.1785/0120090130. URL `http://www.bssaonline.org/cgi/content/abstract/100/2/423`.

L. Knopoff and J.K. Gardner. Higher seismic activity during local night on the raw worldwide earthquake catalogue. *Geophys. J. Intl.*, 28(3):311–313, 1972.

E.L. Lehmann. *Testing Statistical Hypotheses.* Springer, New York, 3rd edition, 2005.

B. Luen. *Earthquake prediction: Simple methods for complex phenomena.* PhD thesis, University of California, Berkeley, 2010.

M.V. Matthews and P.A. Reasenberg. Statistical methods for investigating quiescence and other temporal seismicity patterns. *Pure Appl. Geoph.*, 126(2-4):357–372, 1988.

P.A. Reasenberg. Second-order moment of central California seismicity, 1969-1982. *J. Geophys. Res.*, 90(B7):5479–5495, 1985.

P.A. Reasenberg and M.V. Matthews. Precursory seismic quiescence: A preliminary assessment of the hypothesis. *Pure Appl. Geoph.*, 126(2-4):373–406, 1988.

J.P. Romano. A bootstrap revival of some nonparametric distance tests. *J. Am. Stat. Assoc.*, 83:698–708, 1988.

J.P. Romano. Bootstrap and randomization tests of some nonparametric hypotheses. *Ann. Stat.*, 17:141–159, 1989.

P.M. Shearer and P.B. Stark. The global risk of big earthquakes has not recently increased. *Proc. Nat. Acad. Sci.*, 109(3):717–721, 2012. doi: 10.1073/pnas.1118525109. URL `http://www.pnas.org/content/early/2011/12/12/1118525109.full.pdf+html`.

D. Vere-Jones. Stochastic models for earthquake occurrence. *J. Roy. Stat. Soc., Ser. B*, 32:1–62, 1970.

A. Wald. Sequential tests of statistical hypotheses. *Ann. Math. Stat.*, 16:117–186, 1945.

J. Zhuang, Y. Ogata, and D. Vere-Jones. Stochastic declustering of space-time earthquake occurrences. *J. Am. Stat. Assoc.*, 97 (458):369–380, 2002.