Statistica Sinica Preprint No: SS-2017-0403	
Title	Detection and replenishment of missing data in marked point processes
Manuscript ID	SS-2017-0403
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0403
Complete List of Authors	Jiancang Zhuang Ting Wang and Koji Kiyosugi
Corresponding Author	Jiancang Zhuang
E-mail	zhuangjc@ism.ac.jp
Notice: Accepted version subject to English editing.	

Detection and replenishment of missing data in marked point processes

Jiancang Zhuang[1], *Ting Wang*[2], *and Koji Kiyosugi*[3]
[1] *Institute of Statistical Mathematics*
10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
[2] *Department of Mathematics and Statistics, University of Otago, Dunedin, New Zealand*
[3] *Organization of Advanced Science and Technology, Kobe University*
1-1 Rokkodai-cho, Nada-ku, Kobe 657-8501, Japan

January 16, 2019

Abstract

Records of geophysical events, such as earthquakes and volcanic eruptions, are usually modeled as marked point processes. These records often have missing data, resulting in underestimation of the corresponding hazards. We propose a computational approach for replenishing missing data in the records of temporal point processes with time-separable marks. The basis of this method is that, if such a point process is completely observed, it can be transformed into a homogeneous Poisson process approximately on the unit square $[0, 1]^2$ by a biscale empirical probability integral transformation (BEPIT). This approach includes three key steps: (1) Transforming the process onto $[0, 1]^2$ using the BEPIT, and finding a time-mark range that likely contains missing events; (2) Estimating a new empirical distribution function based on the data in the time-mark range in which the events are supposed to be completely observed; (3) Generating events in the missing region. We test this method on a synthetic dataset, and apply it to the records of volcanic eruptions of the Hakone Volcano in Japan and the aftershock sequence following the 2008 Wenchuan Mw7.9 earthquake in Southwest China. The results show that this algorithm provides a useful way to estimate missing data and to replenish incomplete records of marked point processes. The replenished data provide more robust estimates of the hazard function.

1 Introduction

Many geophysical processes, such as earthquakes and volcanic eruptions, occur at random times and/or locations, and are often described naturally by point-process models (e.g., Vere-Jones, 1970; Zhuang et al., 2002; Wang and Bebbington, 2012, 2013). Point-process models and related theories are also widely used in many other fields, such as crime, disease, and fire (Diggle and Rowlingson, 1994; Schoenberg et al., 2007; Mohler et al., 2011). With the development of advanced technology for recording these natural and social phenomena, the amount of data has increased significantly. However, the degree of completeness of these records varies, and in many cases, small events are often missed in the early period of observation. For example, smaller aftershocks are less likely to be recorded than larger aftershocks during the period immediately following a large earthquake (Ogata and Katsura, 1993; Omi et al., 2013). Other examples include missing data in volcanic eruption records (Kiyosugi et al., 2015) and in the field of communication in social networks (Zipkin et al., 2015). Missing data limit our efficient use of these records, often resulting in biased estimates. However, statistical tools for analyzing incomplete point process data are not well developed.

Geophysicists have been searching for reliable methods to obtain more complete earthquake catalogs. For example, waveform-based detection methods for small earthquakes within an aftershock sequence have been proposed (e.g., Enescu et al., 2007, 2009; Peng et al., 2007; Marsan and Enescu, 2012; Hainzl, 2016). However, even these methods cannot recover all missing aftershocks. An alternative is to switch to energy-based descriptions (Sawazaki and Enescu, 2014); that is, instead of regarding it as a process of events with different magnitudes, the process of earthquake occurrences is regarded as a stream of energies released by earthquakes. However, methods related to such descriptions remain underdeveloped.

Based on the empirical law that the distribution of earthquake magnitudes follows the Gutenberg–Richter magnitude–frequency relation (Gutenberg and Richter, 1944), Ogata and others investigated why events were missing from earthquake catalogs (Ogata and Vere-Jones, 2003; Iwata, 2008, 2013, 2014). They used a Bayesian method to make probabilistic earthquake forecasts, with missing earthquakes taken into account (Ogata, 2006; Omi et al., 2013, 2014, 2015).

In most of the aforementioned studies, when dealing with missing events in a point process, the full structure of the model or the distribution of marks are assumed to be known. However, owing to incomplete records and other reasons, on most occasions, the information available on the process or the mark distribution is limited. Thus, a preferable method for evaluating the missingness should be based on as few assumptions as possible, especially when the temporal structure and the distribution of marks are unknown. Zhuang et al. (2017) used a stochastic algorithm to restore missing aftershocks in the aftershock sequences following several earthquakes in

Kumamoto, Japan (April 14, 2016, $M6.5$; April 15, 2016, $M6.4$; April 16, 2016, $M7.3$). This method can be used to restore missing data in the records of a more general temporal point process with time-separable marks using information from the parts of the process that are completely observed. In Zhuang et al. (2017), the mathematical background is not well addressed. In this study, we explain in detail the mathematics related to this fast algorithm and discuss its asymptotic properties.

In the following sections, we first introduce the biscale empirical probability integral transformation (BEPIT) and then analyze the completely observed process with time-separable marks after the transformation. Based on the results of this transformation, we restore the empirical distributions from an incomplete record using an iterative algorithm. The algorithm is explained using a simulated dataset, and then consistency and asymptotic normality are derived. Finally, we apply the algorithm to investigate the incomplete eruption record of the Hakone volcano in Japan, and the aftershock sequence of the Wenchuan Mw7.9 earthquake that occurred in Southwest China on May 28, 2008.

2 Concepts, methodology, and illustration

2.1 Mark-separable temporal point process and biscale empirical probability integral transformation

Mathematically, a marked temporal point process N is a random subset of discrete points on the space $\mathbb{R} \times \mathbb{M}$, say $\{(t_i, m_i) : i = 1, 2, \cdots, n\}$, which includes a finite or countable number of elements, and satisfies the following two conditions (Karr, 1991): (a) for any bounded subset $A \subset \mathbb{R}$, $\Pr\{N(A \times \mathbb{M}) \equiv \#[N \cap (A \times \mathbb{M})] \ < \infty\} = 1$, where $\#[\ ]$ represents the number of elements in a set; and, (b) for each i, m_i is a random variable on $\mathbb{M}$. In our study, we assume: (a) the marks are continuous random variables, and (b) the point process is simple (i.e., $\Pr\{\max_{t\in\mathbb{R}} N(\{t\}\times\mathbb{M}) \le 1\} = 1$), such that there are no overlapping events on the time axis.

A marked temporal point process is often specified by its conditional intensity function, which is defined by

$$\lambda(t,m)\,\mathrm{d}t\,\mathrm{d}m = \mathbf{E}\left[N([t,t+\mathrm{d}t)\times(m,m+\mathrm{d}m)\mid\mathcal{H}_t\right], \tag{1}$$

where $\mathcal{H}_t$ denotes the history of N up to time t, but not including t. The conditional intensity can be decomposed as

$$\lambda(t,m) = \lambda_g(t)\,g(m|t),$$

where $\lambda_g(t) = \int_{\mathbb{M}} \lambda(t,m)\,\mathrm{d}m$ is called the conditional intensity of the ground point process N_g induced by N on $\mathbb{R}$, defined by $N_g(A) = N(A\times\mathbb{M})$, and $g(m|t)$ is the probability density function of the event mark at time t. An important property of the conditional intensity is that if a temporal point process N has conditional intensity $\lambda(t)$, then the transformation

$$t_i \to \tau_i = \int_0^{t_i} \lambda(u)\,\mathrm{d}u \tag{2}$$

transforms N into a Poisson process $N' = \{\tau_i : i = 1, 2, \cdots\}$ (see, e.g., Ogata, 1988; Schoenberg, 2003; Daley and Vere-Jones, 2003).

For the above conditional intensity, when the mark distribution is separable from the occurrence times, i.e.,

$$\lambda(t,m) = \lambda_g(t)\,g(m), \tag{3}$$

the marks of this point process is said time-separable. Point-process models with time-separable marks have been widely used in many research areas. In seismology, most practical versions of earthquake forecasting models explicitly assume that the magnitude distribution is separable from time (see, e.g., Ogata and Zhuang, 2006; Zhuang et al., 2002, 2004; Zhuang, 2011; Werner et al., 2011; Ogata et al., 2013). In volcanology, Bebbington (2014) suggested that there is not enough evidence of a universal dependence of eruption size on time. In forecasting, time-independent size distributions are used frequently (e.g., Passarelli et al., 2010).

Other ways to specify point-process models include moment intensity functions, Papangelou intensities, and Palm intensities. Traditionally, when a point process is specified in these ways, it refers to a spatial point process. A point process can be completely determined by its likelihood (terminologically, the local Janossy density, see Daley and Vere-Jones, 2003, 2008). This gives the joint probability density/mass function of the total number and each location of the particles in the process, assuming that the particles are indistinguishable. If one of the following three is known: (1)the moment intensities of all orders, (2) the conditional intensity, and (3) the Papangelou intensity, then the likelihood is also known (i.e., the point process is completely determined). Here, we refer to Daley and Vere-Jones (2003, 2008) and Møller and Waagepetersen (2003) for the relations between the Janossy density and three other types of intensities. In this study, as we see in the following sections, the method for replenishing missing data in a marked point process does not depend on any specific form of the conditional intensity. Therefore, it can be applied to spatial point processes as well if the ground space is one-dimensional and the conditional intensity is mark-separable.

Before testing for missing data in a record of a marked point process and replenishing the record, we need to know what a complete record looks like.

Given a series of i.i.d. observations on X, $x_1, x_2, \cdots, x_n$, for a fixed x, the empirical cdf

$$\tilde{F}_X(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(x_i < x)$$

converges almost surely to $F_X(x)$ and, thus, $\tilde{F}_X(X_j)$, $j = 1, 2, \cdots, n$, converges to a unit uniform distribution. We call transformation $x \to \tilde{F}_X(x)$ the empirical probability integral transformation induced by $\{x_1, x_2, \cdots, x_n\}$. In a general marked point process N in $[0, T]$, the occurrence times of an arbitrary event may depend on the occurrence times and/or marks of other events. But the empirical probability integral transformation still results in an approximate unit uniform distribution since the transformation does not require the explicit formulation of the conditional intensity.

Suppose $N = \{(t_i, m_i) : i = 1, 2, \cdots, n\}$ is a realization of a temporal marked point process in a time-mark domain $[0, T] \times \mathbb{M}$, where $\mathbb{M}$ is the space of marks. Consider the following biscale empirical transformation (BEPIT):

$$\begin{aligned} \Gamma_N : \quad [0, T] \times \mathbb{M} \quad &\to \quad [0, 1] \times [0, 1] \\ (t, m) \quad &\to \quad (t', m') = \left(\tilde{F}(t), \tilde{G}(m)\right), \end{aligned} \tag{4}$$

where $\tilde{F}$ and $\tilde{G}$ are the empirical cdfs of $\{t_i : i = 1, 2, \cdots, n\}$ and $\{m_i : i = 1, 2, \cdots, n\}$, respectively. If the marks of the events in the process are separable from the occurrence times, then $\{t'_i : i = 1, 2, \cdots, n\}$ and $\{m'_i : i = 1, 2, \cdots, n\}$, which are the images of $\{t_i : i = 1, 2, \cdots, n\}$ and $\{m_i : i = 1, 2, \cdots, n\}$, respectively, approximately form a homogeneous Poisson process on $[0, 1] \times [0, 1]$. It is straightforward to show the independence between $\tilde{F}(t)$ and $\tilde{G}(m)$ and, thus, given the total number of events N, the number of events in a cell of area $s \subseteq [0, 1] \times [0, 1]$ is a random variable from a binomial distribution $B(N, s)$, which can be approximated by a Poisson distribution with mean Ns. The smaller s gets, the better this approximation.

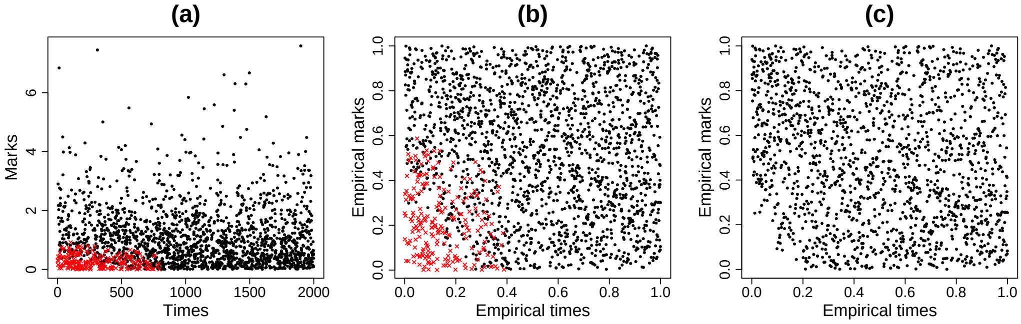

Figure 1: A synthetic dataset of a marked point process. (a) Marks versus occurrence times. (b) Empirical marks versus empirical occurrence times of all the synthetic events under the transformation Γ_N. (c) Empirical marks versus empirical occurrence times for the observed incomplete record under the transformation $\Gamma_{N_{\text{obs}}}$. The red crosses in (a) and (b) represent the missing events.

In the following discussions, we only consider the case of mark-separable Poisson processes. This is because, for the case of a more general process, say N, with a conditional intensity $\lambda(t, m)$, we can transform it into a Poisson process N' with a constant intensity by using the marked version of the transformation in (2), $(t_i, m_i) \in N \to (\tau_i, m_i) \in N'$, where $\tau_i = \int_0^{t_i} \int_{\mathbb{M}} \lambda(t, m)\, \mathrm{d}m\, \mathrm{d}t$. Since such a transformation does not change the chronological order of the events or the mark-separable property of the process, the BEPIT transforms N and N' into the same point patterns.

Example 1. *In Figure 1(a), we simulate a Poisson process N (the combination of black and red points) with a temporal rate $\lambda = 1$ on [0, 2000], and marks following an exponential distribution with mean 1, i.e.,*

$$g(x) = \begin{cases} \mathrm{e}^{-x}, & x > 0; \\ 0, & \textit{otherwise.} \end{cases}$$

Figure 1(b) shows that under transformation (4), N is transformed into an approximately homogeneous Poisson process, say N', which has rate $\lambda = 2000$ and i.i.d. marks uniformly distributed in $[0, 1]$.

2.2 Detection of missing data

When events in part of an observed time-mark range are missing, deterministically or in probability, the separability between the occurrence times and the marks of the observed events is usually destroyed. In addition, the image of the observed N_{obs} mapped by the above BEPIT $\Gamma_{N_{\text{obs}}}$, as defined in (4), may not be a homogeneous process.

Example 2. *Consider the simulated data in Example 1 (Figures 1(a)). Assume the missing probability is*

$$\begin{aligned} q(t,m) &= \Pr\{\textit{an event occurring at } (t,m) \textit{ is missing}\} \\ &= \begin{cases} \min\left[1, \frac{(1000-t)(1-m)}{800}\right], & \textit{if } 0 < t < 800,\ m < 0.3, \\ 0, & \textit{otherwise.} \end{cases} \end{aligned} \tag{5}$$

If we thin the original process N (the combination of the red and black points) in Figure 1(a) with this missing probability, then the red points are deleted (i.e., they are missing from the record). Denote the remaining events (i.e., the observed process) as N_{obs}. Figure 1(c) shows that the image of the observed data of the process under the BEPIT $\Gamma_{N_{\text{obs}}}$ is not homogeneous.

In the above biscale transformation, we do not need to know the exact forms of $g(m)$, λ_g, or q. This method only uses the conditions that the original process is mark-separable, and that the process of missing events is time- and mark-dependent. Thus, for a temporal point process N with time-separable marks, we can test whether there are data missing from its observed record, N_{obs}, by testing the homogeneity of the image $\Gamma_{N_{\text{obs}}}(N_{\text{obs}})$ of the observed data N_{obs} in the biscale transformed domain, when the missing values are time- and mark-dependent. After using the BEPIT $\Gamma_{N_{\text{obs}}}$ to map N_{obs} onto $[0,1]^2$, we divide the overall area of $[0,1]^2$ into L sub-regions of equal areas, $L = L_1 \times L_2$ cells. Here, L_1 is the number of cells along the transformed time domain and L_2 is the number of cells along the transformed mark domain. Then, we calculate the following statistics:

$$R = \frac{\min\{C_1, C_2, \cdots, C_L\}}{\max\{C_1, C_2, \cdots, C_L\}}, \text{ and } D = \max\{C_1, C_2, \cdots, C_L\} - \min\{C_1, C_2, \cdots, C_L\}, \tag{6}$$

where $C_1, C_2, \cdots, C_L$ are the numbers of events falling within each of the L cells. These two statistics are analogous to test statistics for homogeneous multinomial distributions, where "homogeneous" means that each category of the possible outputs has the same probability (Johnson, 1960; Johnson and Young, 1960; Corrado, 2011).

Suppose that $[0, 1]^2$ is divided into $L = L_1 \times L_2$ cells with equal areas, i.e., $[0, 1]^2 = \bigcup_{j=1}^{L_2} \bigcup_{i=1}^{L_1} [(i-1)/L_1, i/L_2] \times [(j-1)/L_2, j/L_2)$, L_1 and L_2 being positive integers. For any point process N on $[0, 1]^2$, if N is a homogenous Poisson process, then the numbers of events in the above L cells, $C_1, C_2, \cdots, C_L$, form a homogeneous $(n, \mathbf{p})$-multinomial random vector, with $\mathbf{p} = (1/L, 1/L, \cdots, 1/L)$. However, if N is obtained by applying the BEPIT to a completely observed mark-separable point process, then the row sum of C_i in the kth row $(1 \le k \le L_1)$, and the column sum of C_i in the jth column $(1 \le j \le L_2)$ are fixed to $\lfloor kn/L_1 \rfloor - \lfloor (k-1)n/L_1 \rfloor$ and $\lfloor jn/L_2 \rfloor - \lfloor (j-1)n/L_2 \rfloor$, respectively, where $\lfloor x \rfloor$ denotes the integer part of x, and n is the total number of events in N. Such constraints do not hold for the homogeneous multinomial distribution. Since the distributions of R and D are complicated, we obtain them by simulation: (1) with n fixed, simulating n events uniformly distributed in $[0, 1]^2$; (2) applying the BEPIT to these n simulated events; (3) with the specified parameters, L_1 and L_2, calculating R and/or D for the transformed points.

Example 3. *We use a simulation to test for missing data in the original and the thinned point processes, as shown in Figures 1(a) and (c), respectively. We simulate 500,000 sequences of the marked Poisson process as defined in Example 1 with the number of events in each simulation the same as those in Figure 1(a). For each simulated sequence, we apply the BEPIT (4) (which results in an image similar to the combination of red and black points in Figure 1(b)). Then, we divide the unit square image into five-by-five cells with equal sizes, and calculate R and D, as defined in (6). After that we plot the empirical cumulative distribution function of the 500,000 values of R and D, as shown in Figures 2(a). To test the thinned process, we simulate another 500,000 sequences of the marked point process with the total number of events in each simulation the same as those in Figure 1(c). The cumulative distributions of R and D are shown in 2(b). We can see that the hypothesis that there are no missing data in the observed (thinned) process can be rejected with a significance level below 0.001 (*$p \leq 2 \times 10^{-6}$*, Figure 2(b)), while, for the original process, the p-values associated with R and D (0.396 and 0.700, respectively) provide no evidence for rejection.*

2.3 Imputation method and algorithm

We start with a heuristic example to explain the algorithm. As shown in Figure 3, suppose that N is a homogeneous point process on $[0,1]^2$, and that events in the domain S are completely unobservable. Let N_{obs} $=\{(x_i, y_i) : (x_i, y_i) \in N \backslash S\}$. Then the empirical distributions of the x- and y-coordinates are, respectively,

$$\tilde{F}_X(x) = \frac{\sum_{i:(x_i,y_i)\in N\backslash S} w_{x,i} I(x_i \leq x)}{\sum_{i:(x_i,y_i)\in N\backslash S} w_{x,i}} \tag{7}$$

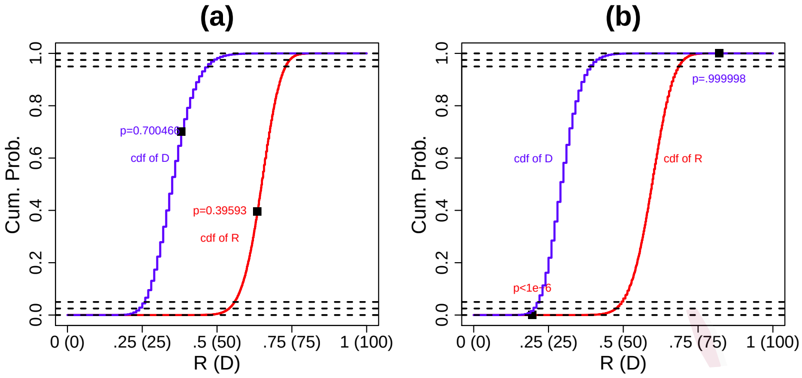

Figure 2: Statistical tests of the existence of missing data on (a) all the events and (b) the observed events in the synthetic point process, with cumulative distribution functions of R (red curve) and D (blue curve). R and D are defined in (6), with $L = L_1 \times L_2$, $L_1 = L_2 = 5$. The cumulative distribution functions in (a) and (b) are obtained from 500,000 simulations with the same numbers of events as in Figures 1(a) and (c), respectively. The black dots in (a) and (b) are the statistics R and D, calculated for the original process in 1(a) and (c), respectively.

and

$$\tilde{F}_Y(y) = \frac{\sum_{i:(x_i,y_i)\in N\setminus S} w_{y,i} I(y_i \le y)}{\sum_{i:(x_i,y_i)\in N\setminus S} w_{y,i}}, \tag{8}$$

where

$$w_{x,i} = \frac{1}{1 - \int_0^1 I((x_i, y) \in S)\mathrm{d}y}, \qquad w_{y,i} = \frac{1}{1 - \int_0^1 I((x, y_i) \in S)\mathrm{d}x}. \tag{9}$$

In most cases, N is not homogeneous in $[0, 1]^2$, and the variation of the event density in S should be considered. Equation (9) should then be

$$w_{x,i} = \frac{1}{1 - \int_0^1 I((x_i, y) \in S)\mathrm{d}F_y(y)}, \qquad w_{y,i} = \frac{1}{1 - \int_0^1 I((x, y_i) \in S)\mathrm{d}F_X(x)}. \tag{10}$$

Since F_Y and F_X are unknown, we replace them by $\tilde{F}_Y$ and $\tilde{F}_X$, respectively, i.e.,

$$w_{x,i} = \frac{1}{1 - \int_0^1 I((x_i, y) \in S)\mathrm{d}\tilde{F}_y(y)}, \qquad w_{y,i} = \frac{1}{1 - \int_0^1 I((x, y_i) \in S)\mathrm{d}\tilde{F}_X(x)}. \tag{11}$$

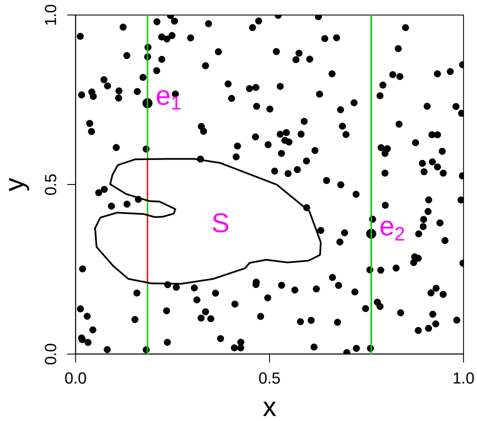

Figure 3: A heuristic estimation of the empirical distribution with missing points. Suppose that, among events $e_i = (x_i, y_i)$, $i = 1, 2, \cdots, N$, events that fall in S cannot be observed. To estimate the empirical distribution $\tilde{F}_X(x)$ of x_i, $i = 1, 2, \cdots, N$, weights need to be assigned to each observed point. That is, when N is uniform, $\tilde{F}_X(x) = \sum_{i=1}^N w_{x,i} I(x_i < x) / \sum_{i=1}^N w_{x,i}$, where $w_{y,i} = 1 - \int_0^1 I((x_i, y) \in S)\, \mathrm{d}y$. In this figure, $w_{x,1}$ is the total length of the green part of the vertical line segments crossing over e_1, and $w_{x,2} = 1$ since the vertical line segment crossing e_2 has no intersection with S.

The above equation, together with (7) and (8), form a solvable equation system. We introduce below an algorithm to solve this equation system.

Firstly, the missing region S needs to satisfy the following condition:

Condtion 1. *The projections of* $([0, T] \times M) \backslash S$ *(i.e., the sub-region in which no event is missing) on the* t*- and* m*-axes cover the entire observation period and the entire range of possible marks, respectively.*

This requirement is to ensure that the empirical distributions of $\{t_i\}$ and $\{m_i\}$ can be restored. With Condition 1 satisfied, when a record is incomplete, we can determine the area, say S, outside of which the record is complete. This can be done either in the original mark-time plot based on prior knowledge of the data quality or in the BEPIT domain based on the statistics R or D.

The algorithm to replenish the record includes three key steps: (1) transforming the process onto $[0, 1]^2$ using the BEPIT to find a time-mark range that likely contains all the missing events; (2) estimating a new empirical distribution function based on the data in the time-mark range, inside which events are supposed to be completely observed; (3) generating events in the missing region.

Initial settings. *Given the dataset* $N_{\text{obs}} = \{(t_i, m_i) : i = 1, 2 \cdots, n\}$ *observed in* $[0, T] \times M$ *and a time-mark range* S*, known to include the missing events, suppose that* S *satisfies Condition 1.*

Step 1. We project the observed data and the range S *that contains the missing data onto* $[0, 1]^2$ *using the BEPIT in (4). Explicitly, set*

$$(t_i^{(1)}, m_i^{(1)}) = \Gamma^{(1)}_{N_{\text{obs}}}(t_i, m_i) \tag{12}$$

where

$$\Gamma^{(1)}_{N_{\text{obs}}}(t, m) = \left(\tilde{F}^{(1)}(t),\, \tilde{G}^{(1)}(m)\right) = \left(\frac{1}{n}\sum_{j=1}^{n} \mathbf{1}(t_j < t),\ \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}(m_j < m)\right). \tag{13}$$

Denote $S^{(1)}$ *as the image of* S *under the transformation* $\Gamma^{(1)}_{N_{\text{obs}}}$.

Step 2. Starting from $\ell = 1$*, repeat the following iterative computation until convergence (e.g.,* $\max\{|t_i^{(\ell+1)} - t_i^{(\ell)}|, |m_i^{(\ell+1)} - m_i^{(\ell)}|\} < \epsilon$*), where* ϵ *is a given small positive number:*

$$(t_i^{(\ell+1)}, m_i^{(\ell+1)}) = \Gamma^{(\ell+1)}_{N_{\text{obs}}}(t_i^{(\ell)}, m_i^{(\ell)}; S^{(\ell)}), \qquad i = 1, 2, \cdots, n, \tag{14}$$

$$S^{(\ell+1)} = \Gamma^{(\ell+1)}_{N_{\text{obs}}}(S^{(\ell)}; S^{(\ell)}), \tag{15}$$

where

$$\Gamma^{(\ell+1)}_{N_{\text{obs}}}(t, m; A) = \left(\frac{\sum_{j=1}^{n} w_1^{(\ell)}(t_j^{(\ell)}, m_j^{(\ell)}, A)\,\mathbf{1}(t_j^{(\ell)} < t)}{\sum_{j=1}^{n} w_1^{(\ell)}(t_j^{(\ell)}, m_j^{(\ell)}, A)},\ \frac{\sum_{j=1}^{n} w_2^{(\ell)}(t_j^{(\ell)}, m_j^{(\ell)}, A)\,\mathbf{1}(m_j^{(\ell)} < m)}{\sum_{j}^{n} w_2^{(\ell)}(t_j^{(\ell)}, m_j^{(\ell)}, A)}\right) \tag{16}$$

with the weights defined by

$$w_1^{(\ell)}(t,m,A) = \frac{\mathbf{1}\left((t,m) \notin A\right)}{1 - \int_0^1 \mathbf{1}\left((t,m') \in A\right)\, \mathrm{d}G^{(\ell)}(m')} \tag{17}$$

$$w_2^{(\ell)}(t,m,A) = \frac{\mathbf{1}\left((t,m) \notin A\right)}{1 - \int_0^1 \mathbf{1}\left((t',m) \in A\right)\, \mathrm{d}F^{(\ell)}(t')}, \tag{18}$$

for any regular region $A \subset [0,1]^2$. Denote the results upon convergence by $N^*_{\text{obs}} = \{(t_i^*, m_i^*) : i = 1, 2 \cdots, n\}$ and S^*.

Step 3. Generate a random number K from a negative binomial distribution, with parameters $(k, 1 - |S^*|)$, where $|S^*|$ is the area of S^* and

$$k = \sum_{i=1}^{n} \mathbf{1}((t_i^*, m_i^*) \notin S^*) = \#(N^*_{\text{obs}} \setminus S^*).$$

Step 4. Generate K random events independently, identically, and uniformly distributed in S^*. Denote these newly generated events by N^*_{rep}.

Step 5. For each event in N^*_{obs}, say, (t_j, m_j), that falls in S^*, sequentially remove from N^*_{rep} the event that is the closest to (t_j, m_j).

Step 6. Convert the resulting N^*_{rep} from the last step to the original observation space $[0,T] \times M$ through linear interpolation:

$$s_j = \mathrm{LI}\left(s_j^*;\, [0,\, t_1^*,\, t_2^*,\, \cdots,\, t_n^*, 1],\, [0,\, t_1,\, t_2,\, \cdots,\, T]\right), \tag{19}$$

$$v_j = \mathrm{LI}\left(v_j^*;\, [0,\, m_1^*,\, m_2^*,\, \cdots,\, m_n^*],\, [0,\, m_1,\, m_2,\, \cdots,\, m_n]\right), \tag{20}$$

for each $(s_j^*, v_j^*) \in N^*_{\text{rep}}$, where $\mathrm{LI}(x, A, B)$ represents the linear interpolation value of x, conditional on the function values for each component in A being locations corresponding to each component in B. Denote the set consisting of all (s_j, v_j) by N_{rep}.

Final output. *Return N_{rep}.*

Example 4. *The above algorithm is applied to the thinned dataset in Example 2. The output from Steps 4 to 6 is shown in Figures 4(b)-(c). The final output for our simulation example is shown in Figure 4(d). Tests using statistics R and D in (6) give p-values of 0.605 and 0.718, respectively, providing no evidence to reject the hypothesis that the replenished dataset is complete (Figure 4(e)). Figure 4(f) compares the cumulative numbers of events in the original, the observed, and the replenished processes, showing that the replenishing algorithm recovers the missing data to some extent.*

Notes:

(1) Equation (13) is the BEPIT that we mentioned in the previous section. If the data are completely recorded, $\{(t_i^{(1)}, m_i^{(1)}),\ i = 1,\ 2 \cdots,\ n\}$ form an approximately homogeneous process on $[0,1]^2$. As we can see in Figure 2(b), the sparseness of points around the lower, left corner implies that smaller events are missing in the earlier period. Rather than choosing S in Figure 1(a), it is more convenient to specify $S^{(1)}$ directly in Figure 2(a) or (b).

(2) Step 2 is carried out based on the fact that the transformation $\Gamma_{N_{\text{obs}}}$ and $S^{(1)} = \Gamma_{N_{\text{obs}}}(S)$ can be quite different from Γ_N, owing to the missing data. The iteration in this step helps us construct a biscale transformation as close as possible to the BEPIT yielded by the complete data (i.e., $\Gamma^*_{N_{\text{obs}}} \approx \Gamma_N$). At the same time, the corresponding area that contains the missing data, S^*, is restored. This can be seen by comparing Figures 1(b) and 4(b)

Step 2 essentially solves F^* and G^* in the following equations:

$$F^*(t) = \frac{\sum_{j=1}^n w_1(t_j, m_j, S)\,\mathbf{1}\,(t_j < t)}{\sum_{j=1}^n w_1(t_j, m_j, S)}, \tag{21}$$

$$G^*(m) = \frac{\sum_{j=1}^n w_2(t_j, m_j, S)\,\mathbf{1}\,(m_j < m)}{\sum_{j=1}^n w_2(t_j, m_j, S)}, \tag{22}$$

where

$$w_1(t, m, S) = \frac{\mathbf{1}\,((t, m) \notin S)}{1 - \int_M \mathbf{1}\,((t, m') \in S)\,\mathrm{d}G^*(m')} \tag{23}$$

$$w_2(t, m, S) = \frac{\mathbf{1}\,((t, m) \notin S)}{1 - \int_M \mathbf{1}\,((t', m) \in S)\,\mathrm{d}F^*(t')}. \tag{24}$$

If we define $\Gamma^*_{N_{\text{obs}}}(t, m) = (F^*(t), G^*(m))$ as a mapping from $[0, T] \times M$ to $[0, 1]^2$, then $\Gamma^*_{N_{\text{obs}}}(t, m)$ directly maps N_{obs} to N^*_{obs} and S to S^*.

The existence of a solution in the iteration given by (21) to (24) and the asymptotic property of the solution are given in the supplementary materials.

(3) Steps 3 and 4 are based on the following fact: given a homogeneous Poisson process with an unknown occurrence rate, if there are k events falling within an area of S_1, then the number of events falling in the complementary area S_2 follows a negative binomial distribution with parameter $(k, |S_1|/(|S_1| + |S_2|))$ (e.g., DeGroot, 1986, 258–259).

(4) In step 5, given the existing events observed in S, we should keep them and remove the same number of simulated points.

One advantage of the algorithm is that if S is unknown, we can use the mark-time plot of $N^{(1)}$, as in Figure 2(b), to decide $S^{(1)}$ by justifying which region is likely to contain the missing events, and then continue with Step 2. Once the replenishment is done, S can be obtained by substituting the coordinate of each point on the boundary of S^* into (19) and (20).

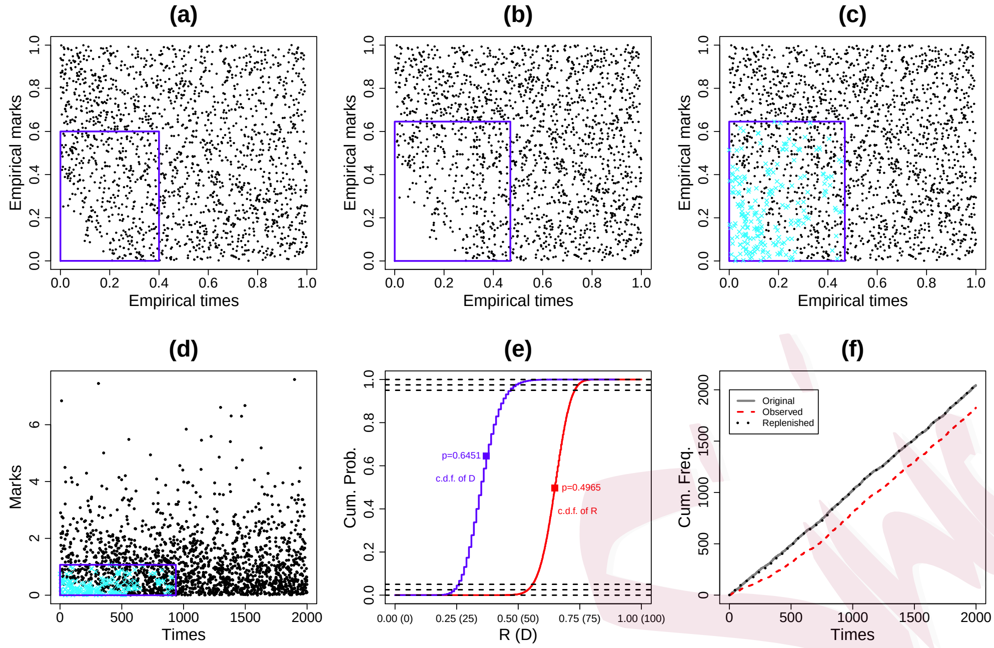

Figure 4: An application of the proposed replenising algorithm to the synthetic dataset. (a) Rescaled marks versus rescaled occurrence times of the observed events (black dots), with the biscale transformation $\Gamma_{N_{\text{obs}}}$ based on the observed process. The blue polygon is the missing area, $S^{(1)}$. (b) Rescaled marks versus rescaled occurrence times of the observed events (black dots), with the rescaling $\Gamma^*_{N_{\text{obs}}}$ based on the events outside of S. The blue polygon is the missing area after transformation $\Gamma^*_{N_{\text{obs}}}$, i.e., S^*. (c) Rescaled marks versus rescaled occurrence times of the observed and replenished events (blue dots) (i.e., newly generated events after removing events that are closest to any of those observed in S, with the rescaling $\Gamma^*_{N_{\text{obs}}}$ based the empirical distributions of the events outside S. (d) Marks versus occurrence times of the observed synthetic events and the replenished events. (e) Cumulative distribution functions of R (red curve) and D (blue curve) for testing missing data in the replenished dataset in (c). (f) Cumulative frequencies versus occurrence times for the original, observed, and replenished processes.

2.4 More simulations

To illustrate the overall behavior of the above replenishing algorithm, we repeat the algorithm many times, with S fixed, for the following two cases: (1) Simulating a Poisson process with λ= 2000; (2) Simulating Poisson processes with rate λ drawn from a uniform distribution within [100, 3000]. Both simulations have the same missing probability functions, as given by

(5). Figures 5(a) and (b) give the comparison between the true numbers of missing events and the number of the replenished events for cases (1) and (2), respectively. In Figure 5(a), since λ is fixed, the number of replenished events is independent of the true number of missing events, and has a larger variance. Some statistics related to these simulations, including the mean numbers and variances of the missing and the replenished points, the mean of relative differences, and the relative difference of means in 500 and 2000 simulations are given in Table 1. In particular, the near-zero relative deviation of the mean number of the replenished events shows that the proposed method is consistent. Here, the larger values of the mean relative deviation of the number of replenished events from the number of missing events illustrate the nature of the uncertainty related to the problem. Such uncertainty is produced not only by the randomness of the numbers of replenished and missing events, but also by the uncertainty in the estimation of the occurrence rate in the process from the events in the non-missing part. In Figure 5(b), the expected number of replenished events in many repeated simulations is close to the number of missing events. Moreover, the relative deviation decreases when the number of missing events (or λ) increases. These results imply that this algorithm replenishes the missing events reasonably well. Also, when λ or the number of events in the process is quite small, there are some outputs that the number of replenished events (when the number of missing events is less than 50 in Figure 5(b)), which is simply calculated by the number of simulated events in S in Steps 3 and 4 minus the number of observed events in S, is negative. This indicates that the existence of missing data in these situations cannot be quantified probabilistically.

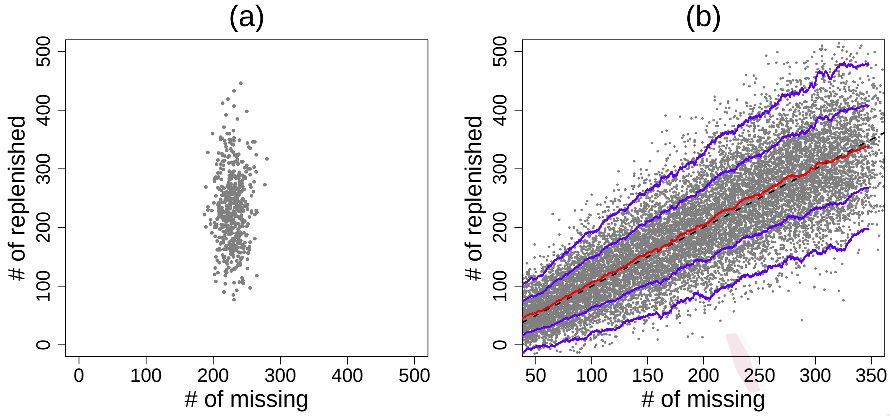

Figure 5: Comparison between the number of true missing events and the number of replenished events. (a) $\lambda = 2,000$ fixed. (b) λ is drawn from a uniform distribution between 100 and 3000. The dashed line represents the case where the numbers of missing and replenished events are equal. The blue and red curves represent the running mean and the corresponding single and double standard deviation bands.

Table 1: Statistics related to Figure 5(a). $\#m$: number of missing points; $\#r$: number of replenished points; $\bar{\cdot}$: mean value; $\sigma(\cdot)$: standard deviation.

#simu.	$\overline{\#m}$	$\sigma(\#m)$	$\overline{\#r}$	$\sigma(\#r)$	$\overline{\left[\frac{\|\#m-\#r\|}{\#m}\right]}$	$\frac{\|\overline{\#m}-\overline{\#r}\|}{\overline{\#m}}$
500	228.274	14.929	232.006	63.926	0.226	0.016
2000	227.712	14.719	230.860	62.145	0.224	0.014

3 Application

3.1 Volcanic eruption record

In this example, we analyze the record of eruptions from the Hakone volcano. The Hakone volcano is an active volcano located at the northern boundary zone of the Izu-Mariana volcanic arc in central Japan (Yukutake et al., 2010; Honda et al., 2014). Data on Japanese explosive eruptions are compiled from the Smithsonian's Global Volcanism Program database (Siebert and Simkin, 2002), the Large Magnitude Explosive Volcanic Eruptions database (LaMEVE database, Crosweller et al., 2012), and additional Japanese databases (Machida and Arai, 2003; Committee for Catalog of Quaternary Volcanoes in Japan (ed), 2000; Geological Survey of Japan, AIST

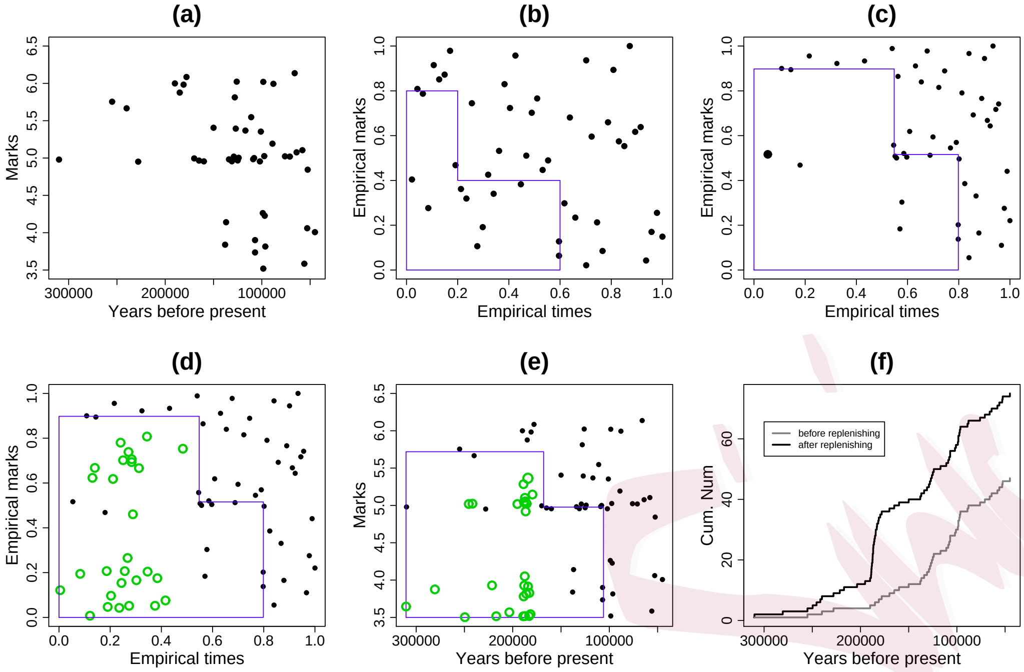

Figure 6: Results from applying the replenishment algorithm to volcanic eruption data. (a) Marks versus occurrence times of the eruption events. (b) Empirical distribution of marks versus that of occurrence times. (c) Rescaled marks versus rescaled occurrence times, with the rescaling based on the empirical distributions of the events outside of S. (d) Rescaled marks versus rescaled occurrence times of the observed and replenished events (i.e., newly generated events after removing events that are closest to any of those observed in S), with the rescaling based the empirical distributions of the events outside of S. (e) Marks versus occurrence times of the observed and replenished events. (f) Cumulative numbers of events against occurrence times. The blue polygon is the area S and its corresponding mappings in which the missing events fall. The green dots are the replenished events.

(ed), 2013; Hayakawa, 2010).

For the Hakone volcano, 46 of 54 compiled events have an eruption magnitude (M = $\log_{10}$[erupted mass in kg] $-$ 7; see Pyle (2015)) equal to or larger than 4 (Table S1 in the supplementary materials). Figure 6(a) shows the eruption magnitudes versus occurrence times of these 46 events. Figure 6(b) shows the empirical distribution, transformed following Step 1 of the algorithm. Based on this plot, the polygon boundaries of S are determined based on the following assumptions. First, events of empirical marks $<$ 0.8

($M < 5.7$) are missing before the empirical time = 0.2 (165 ka). Second, the recording of larger events improves after the empirical time = 0.2 (165 ka), though events of empirical marks < 0.4 ($M < 5.0$) are still missing. Third, the recording of events improves further, and there are no missing events after the empirical time = 0.6 (105 ka). The results from running the replenishing algorithm are shown in Figures 6(c) to 6(e).

The estimated cumulative number of events for the replenished dataset shows a remarkable jump of around 180 ka (Figure 6(f)). This jump is caused by the replenished events synthesized around 180 ka (Figure 6(e)) based on the cluster of four large events ($M \sim 6$) at 178 ka, 181 ka, 185 ka and 190 ka (Figure 6(a); Hayakawa, 2010). The ages of the events at the Hakone volcano are still not fully agreed in the literature. For example, Yamamoto (2015) assumed that the ages of these eruptions are about 135 ka, 135 ka, 180 ka and 215 ka, respectively. Therefore, the reliability of the jump of the cumulative number of events (Figure 6(f)) is a problem in volcanological dating of event ages. In addition, estimating the tephra volume and rounded eruption magnitude is also a problem in volcanology (Brown et al., 2014). For example, the analyzed dataset has clusters of events with magnitude 4 and 5 (Figure 6(a)) and, therefore, the replenished events around 180 ka are also clustered around magnitudes 4 and 5 (Figure 6(e)).

Note that it is difficult to determine the exact period of under-recording in the eruption history of each volcano. Kiyosugi et al. (2015) showed that there are still a lot of eruptions missing in the overall Japanese database, even for the last 100,000 years. Therefore, the polygon shape (Figure 6(b)) that we used suggests that our replenished data have the same completeness level as the data outside the polygon. Our method is one possibility of considering the under-recording of events in volcanic hazard assessments of explosive eruptions using geological records.

3.2 Earthquake catalog: missing aftershocks

It is well known that, immediately after a large earthquake, many aftershocks cannot be recorded because the seismic waveforms generated by the aftershocks cannot be distinguished from the overlapping waveforms generated by the mainshock on seismographs. In this section, we study the earthquake catalog from Southwest China, between January 1, 1990, and April 20, 2013, in a space range of $26° - 34°N$ and $97° - 107°E$ with minimum magnitude 3.0 (Figure S2 in supplementary materials). This dataset is selected from the Chinese Earthquake catalog compiled by the China Earthquake Data Center (CEDC) (URL: http://data.earthquake.cn/index.html). The Wenchuan Mw 7.9 (Ms 8.0) earthquake, which occurred on May 12, 2008, was one of the two largest seismic events in China during the last 50 years. There are 6,249 events in the selected space and time range, among which 3,754 events occurred after the Wenchuan earthquake, indicating low seismicity level above magnitude 3 in the study region before 2008. There are many aftershocks missing immediately after the mainshock. In particular, events of magnitudes between 3 and 4 are not properly recorded for a period of about one-and-a-half months after the mainshock. The majority of the events after May 12, 2008, can be taken as clustering events triggered by the Wenchuan mainshock. When analyzing seismicity in this area, Jia et al. (2014) and Guo et al. (2015) adopted a relatively high magnitude threshold of 4.0 to avoid biases in estimates caused by missing events, with 5,217 of the 6,249 events being ignored.

This example is quite different from the previous example and the simulated data. The missing range can be well specified before replenishment: the missing values are known immediately after the occurrence of the mainshock, and the monitoring ability for events between magnitudes 3 and 4 are restored one and half months later. The results are illustrated in Figure 7. We can see that missing events take up about half the total number of events.

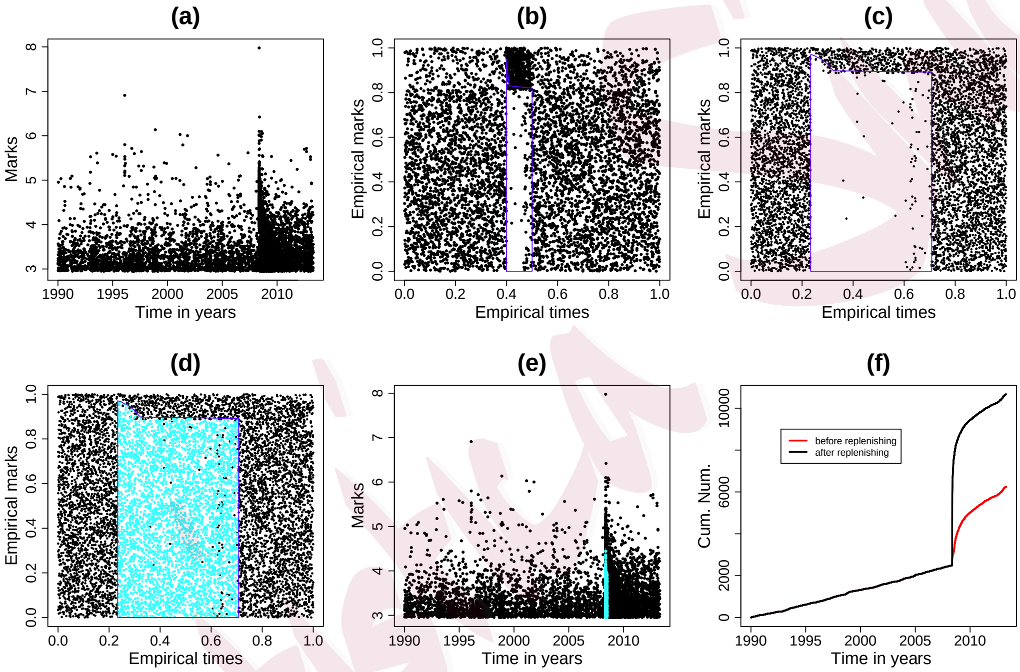

Figure 7: Results from applying the replenishment algorithm to the earthquake data from Southwest China. (a) Marks versus occurrence times of the earthquake events. (b) Empirical distribution of marks versus that of occurrence times. (c) Rescaled marks versus rescaled occurrence times, with the rescaling based on the empirical distributions of the events outside of S. (d) Rescaled marks versus rescaled occurrence times of the observed and replenished events (i.e., newly generated events after removing events that are closest to any of the observed in S), with the rescaling based on the empirical distributions of the events outside S. (e) Marks versus occurrence times of the observed and replenished events. (f) Cumulative numbers of events against occurrence times. The blue polygon is the area S and its corresponding mappings in which the missing events fall. The blue dots are replenished events.

In seismology, the frequency of aftershock occurrences in an aftershock sequence can be modeled by the empirical Omori-Utsu formula (e.g., Utsu et al., 1995)

$$\lambda(t) = \frac{K}{(t+c)^p}, \tag{25}$$

where K is an index proportional to the number of earthquakes excited by the mainshock, c is related to the period after the mainshock, from which the aftershock rate drops slowly, and p is the power related to the decay rate of aftershocks. Utsu et al. (1995) discussed how the parameters c and p change with the cutoff magnitude threshold, and hypothesized that such changes are caused by the fact that small aftershocks in an early stage of the sequence are missing from the catalog. We fit the above Omori-Utsu formula to both the original and the replenished catalogs (Table 2) and obtain the maximum likelihood estimates of parameters. The results show that after the replenishment, the Omori parameters c and p no longer change. We also fit the Omori formula to the original dataset, but only consider earthquakes that occurred at least 54 days after the mainshock. In this case, though c and p are slightly different from the estimates for the replenshed data from the starting time, they do not change much when the magnitude threshold changes from 2.95 to 4.15 (Table S2 in the supplementary materials). These results confirm numerically Utsu et al. (1995)'s hypothesis that missing small events in the early stage of an aftershock sequence causes the instability of the estimate of the Omori-Utsu formula.

4 Conclusions and Discussions

In this study, we proposed a method for replenishing missing data in marked temporal point processes, based on only the assumption that the marks of the

Magnitude threshold	Replenished dataset $[t_{\text{main}}, T]$			Orig. dataset $[t_{\text{main}}, T]$		
	$\hat{K}$	$\hat{c}$	$\hat{p}$	$\hat{K}$	$\hat{c}$	$\hat{p}$
2.95	804.4	.1140	1.003	82.29	.0553	.6205
3.05	639.2	.1131	1.003	80.31	.0596	.6547
3.15	511.5	.1134	1.001	79.25	.0660	.6872
3.25	412.9	.1110	.9965	79.04	.0737	.7185
3.35	327.3	.1067	.9926	78.80	.0825	.7555
3.45	260.3	.1141	.9925	80.67	.0991	.7986
3.55	213.8	.1142	.9953	83.33	.1177	.8407
3.65	171.6	.1135	.9907	85.73	.1360	.8799
3.75	135.9	.1132	.9911	90.18	.1642	.9278
3.85	111.2	.1029	.9941	95.17	.1935	.9708
3.95	100.0	.1241	1.015	103.2	.2383	1.023
4.05	74.12	.1082	1.013	79.20	.1938	1.027
4.15	60.65	.1266	1.026	62.92	.1690	1.034

Table 2: Results from fitting the Omori-Utsu formula to the original and the replenished datasets of earthquakes from Southwest China, with different magnitude thresholds. t_{main}: occurrence time of the mainshock; T: end of the time interval.

events are separable from the occurrence times, regardless of how the events interact on the time axis. The key point of this method is an algorithm that iteratively estimates the missing area in the transformed domain according to the parts where data are completely recorded. This method is applied to the eruption record of the Hakone volcano in Japan and the earthquake catalog from Southwest China, including the aftershock zone of the 2008 Mw7.9 Wenchuan earthquake. The results show that the proposed method helps us evaluate the influence of missing data and correct the bias caused by missing data in our conclusion.

Detection of the missing area In our two examples, the missing area is determined by visual inspection of the biscale transformed data for the historical records of the Hakone volcano and by prior information on the seismic network for the Wenchuan aftershock sequence. In most cases, such missing area needs to be determined by the experience of data analysts or information on the data from other sources. However, it is possible to turn the replenishing algorithm into an automated algorithm.

Starting from $S' = \emptyset$, we divide the unit square into small cells in the biscale transformed domain obtained by applying the transformation defined in (9) to (13). Then, we carry out the statistical tests based on the statistics R or D on the cells that do not intersect S', as discussed in Section 3. If the test shows that missing cells exist, then we merge these cells into S'. Such steps are iterated until no more cells are added to S'. Since this topic belongs to the scope of data processing algorithms, we did not include it in this statistical article.

Separability of marks As discussed earlier, the applicability of this algorithm depends on whether the mark distribution is separable from the occurrence time. If such dependence is known explicitly as a probability density function, say $g(m \mid t)$, we can directly use the cdf that corresponds to f in Steps 1 and 2 in the algorithm (i.e., $m_i^{(\ell)} = G(m_i \mid t_i)$ for $\ell \geq 1$). Of course, such dependence should also be considered when transforming the marks of replenished events from $[0, 1]$ to the original mark space. If the mark is dependent on the time, but we do not know how it depends on the time, together with the existence of missing events, the replenishment/imputation problem becomes unidentifiable.

Another case that is worth discussing is when the mark distribution is known and does not depend on time. We can again use the cdf of the marks in Steps 1 and 2 directly in the algorithm (i.e., by setting $m_i^{(\ell)} = G(m_i)$ for

$\ell \geq 1$). Such missing data can also be estimated using Bayesian methods, as in Ogata and Katsura (1993), and then replenished by direct simulation.

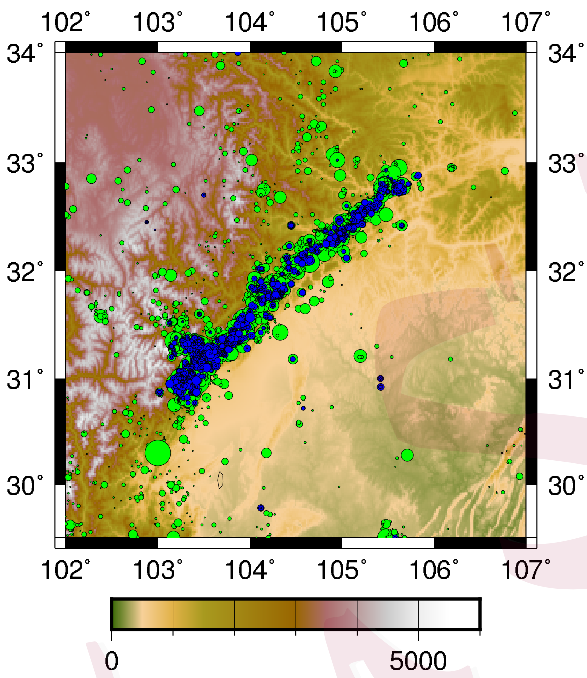

Figure 8: Epicenter map of imputed earthquakes (solid blue circles) for the Wenchuan aftershock sequence.

Imputation of locations This method is powerful for marked temporal point processes, but it cannot be extended easily to high-dimension or spatiotemporal cases. This is because in most cases, the process is not homogeneous in space. However, it is still possible case by case. For example, in replenishing the Wenchuan aftershock sequence, we can use the clustering feature of earthquakes. A simple replenishing algorithm is as follows. For each simulated event, find a fixed number, for example, 50, of events closest to it in time in the observed process. Then we construct a Delaunay tessellation network for these 50 events and select with equal probabilities one of the Delaunay triangles, and put this simulated event randomly and uniformly in this selected triangle. An example of the imputed locations of the missing aftershocks of the Wenchuan earthquake is shown in Figure 8. For a spatially inhibitive process, different methods should be used.

In summary, the method proposed in this study is useful in dealing with missing data problem in point-process observations, such as volcano eruption records and historical or short-term earthquake catalogs.

Acknowledgement

This project is supported by the Royal Society of New Zealand Marsden Fund (contact UOO1419). JZ is also partially supported by Grants-in-Aid No. 2530052 for Scientific Research (C) from the Japan Society for the Promotion of Science. Helpful discussions with David Harte from GNS New Zealand and Boris Baeumer from Otago University are gratefully acknowledged. The authors also thank the AE and three anonymous reviewers for their encouragement and constructive comments.

References

Bebbington, M. S. (2014). Long-term forecasting of volcanic explosivity. *Geophysical Journal International*, 197:1500–1515.

Brown, S. K., Crosweller, H. S., Sparks, R. S. J., Cottrell, E., Deligne, N. I., Guerrero, N. O., Hobbs, L., Kiyosugi, K., Loughlin, S. C., Siebert, L., and Takarada, S. (2014). Characterisation of the quaternary eruption record: analysis of the large magnitude explosive volcanic eruptions (LaMEVE) database. *J Appl Volcanol*, 3:5.

Committee for Catalog of Quaternary Volcanoes in Japan (ed) (2000). Catalog of quaternary volcanoes in Japan (in Japanese).

Corrado, C. J. (2011). The exact distribution of the maximum, minimum and the range of multinomial/dirichlet and multivariate hypergeometric frequencies. *Statistics and Computing*, 21(3):349–359.

Crosweller, H., Arora, B., Brown, S. K., Cottrell, E., Deligne, N., Guerrero, N., Hobbs, L., Kiyosugi, K., C., L. S., Lowndes, J., Nayembil, M., Siebert, L., Sparks, R. S. J., Takarada, S., and Venzke, E. (2012). Global database on large magnitude explosive volcanic eruptions (LaMEVE). *Journal of Applied Volcanology*, 1:4.

Daley, D. D. and Vere-Jones, D. (2003). *An Introduction to Theory of Point Processes – Volume 1: Elementary Theory and Methods (2nd Edition)*. Springer, New York, NY.

Daley, D. D. and Vere-Jones, D. (2008). *An Introduction to Theory of Point Processes – Volume II: General Theory and Structure (2nd Edition)*. Springer, New York, NY.

DeGroot, M. H. (1986). *Probability and Statistics (Second ed.)*. Addison-Wesley.

Diggle, P. J. and Rowlingson, B. S. (1994). A conditional approach to point process modelling of elevated risk. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 157(3):433–440.

Enescu, B., Mori, J., and Miyazawa, M. (2007). Quantifying early aftershock activity of the 2004 mid-Niigata prefecture earthquake (Mw6.6). *Journal of Geophysical Research: Solid Earth*, 112(B4):B004629.

Enescu, B., Mori, J., Miyazawa, M., and Kano, Y. (2009). Omori-Utsu law c-values associated with recent moderate earthquakes in Japan. *Bulletin of the Seismological Society of America*, 99(2A):884–891.

Geological Survey of Japan, AIST (ed) (2013). Catalog of eruptive events during the last 10,000 years in japan, version 2.1 (in japanese). Technical report.

Guo, Y., Zhuang, J., and Zhou, S. (2015). An improved space-time ETAS model for inverting the rupture geometry from seismicity triggering. *Journal of Geophysical Research: Solid Earth*, 120(5):3309–3323. 2015JB011979.

Gutenberg, B. and Richter, C. F. (1944). Frequency of earthquakes in California. *Bull. Seis. Soc. Am.*, 34:184–188.

Hainzl, S. (2016). Rate-dependent incompleteness of earthquake catalogs. *Seismological Research Letters*, 87(2A):337–344.

Hayakawa, Y. (2010). Hayakawafs 2000-year eruption database and one million-year tephra database, http://www.hayakawayukio.jp/database/.

Honda, R., Yukutake, Y., Yoshida, A., Harada, M., Miyaoka, K., and Satomura, M. (2014). Stress-induced spatiotemporal variations in anisotropic structures beneath hakone volcano, japan, detected by s wave splitting: A tool for volcanic activity monitoring. *Journal of Geophysical Research: Solid Earth*, 119(9):7043–7057.

Iwata, T. (2008). Low detection capability of global earthquakes after the occurrence of large earthquakes: Investigation of the Harvard CMT catalogue. *Geophysical Journal International*, 174(3):849–856.

Iwata, T. (2013). Estimation of completeness magnitude considering daily variation in earthquake detection capability. *Geophysical Journal International*, 194(3):1909–1919.

Iwata, T. (2014). Decomposition of seasonality and long-term trend in seismological data: A Bayesian modelling of earthquake detection capability. *Australian & New Zealand Journal of Statistics*, 56(3):201–215.

Jia, K., Zhou, S., Zhuang, J., and Jiang, C. (2014). Possibility of the independence between the 2013 Lushan earthquake and the 2008 Wenchuan earthquake on Longmen Shan Fault, Sichuan, China. *Seismological Research Letters*, 85(1):60–67.

Johnson, N. L. (1960). An approximation to the multinomial distribution some properties and applications. *Biometrika*, 47(1-2):93–102.

Johnson, N. L. and Young, D. H. (1960). Some applications of two approximations to the multinomial distribution. *Biometrika*, pages 463–469.

Karr, A. (1991). *Point Processes and Their Statistical Inference.* Marcel Dekker, Inc., New York and Basel.

Kiyosugi, K., Connor, C. B., Sparks, R. S. J., Crosweller, H. S., Brown, S. K., Siebert, L., Wang, T., and Takarada, S. (2015). How many explosive eruptions are missing from the geologic record? analysis of the quaternary record of large magnitude explosive eruptions in japan. *Journal of Applied Volcanology*, 4:17.

Machida, H. and Arai, F. (2003). *Atlas of Tephra in and around Japan, revised edition.* University of Tokyo Press, Japan (in Japanese).

Marsan, D. and Enescu, B. (2012). Modeling the foreshock sequence prior to the 2011, MW9.0 Tohoku, Japan, earthquake. *Journal of Geophysical Research: Solid Earth*, 117(B6):B06316.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108.

Møller, J. and Waagepetersen, R. P. (2003). *Statistical Inference and Simulation for Spatial Point Processes*. Chapman and Hall/CRC.

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

Ogata, Y. (2006). Monitoring of anomaly in the aftershock sequence of the 2005 earthquake of M7.0 off coast of the western Fukuoka, Japan, by the ETAS model. *Geophysical Research Letters*, 33:L01303.

Ogata, Y. and Katsura, K. (1993). Analysis of temporal and spatial heterogeneity of magnitude frequency distribution inferred from earthquake catalogues. *Geophysical Journal International*, 113(3):727–738.

Ogata, Y., Katsura, K., Falcone, G., Nanjo, K., and Zhuang, J. (2013). Comprehensive and topical evaluations of earthquake forecasts in terms of number, time, space, and magnitude. *Bulletin of the Seismological Society of America*, 103(3):1692–1708.

Ogata, Y. and Vere-Jones, D. (2003). Examples of statistical models and methods applied to seismology and related earth physics. In Lee, W. H., Kanamori, H., Jennings, P. C., and Kisslinger, C., editors, *International Handbook of Earthquake and Engineering Seismology, Vol.81B*, chapter 82. International Association of Seismology and Physics of Earth's Interior.

Ogata, Y. and Zhuang, J. (2006). Space-time ETAS models and an improved extension. *Tectonophysics*, 413(1-2):13–23.

Omi, T., Ogata, Y., Hirata, Y., and Aihara, K. (2013). Forecasting large aftershocks within one day after the main shock. *Scientific Reports*, 3:2218.

Omi, T., Ogata, Y., Hirata, Y., and Aihara, K. (2014). Estimating the ETAS model from an early aftershock sequence. *Geophysical Research Letters*, 41(3):850–857.

Omi, T., Ogata, Y., Hirata, Y., and Aihara, K. (2015). Intermediate-term forecasting of aftershocks from an early aftershock sequence: Bayesian and ensemble forecasting approaches. *Journal of Geophysical Research: Solid Earth*, 120(4):2561–2578.

Passarelli, L., Sandri, L., Bonazzi, A., and Marzocchi, W. (2010). Bayesian hierarchical time predictable model for eruption occurrence: an application to kilauea volcano. *Geophysical Journal International*, 181(3):1525–1538.

Peng, Z., Vidale, J. E., Ishii, M., and Helmstetter, A. (2007). Seismicity rate immediately before and after main shock rupture from high-frequency waveforms in Japan. *Journal of Geophysical Research: Solid Earth*, 112(B3):B03306.

Pyle, D. M. (2015). Sizes of volcanic eruptions. In Sigurdsson, H., editor, *The Encyclopedia of Volcanoes (Second Edition)*, chapter 13, pages 257 – 264. Academic Press, Amsterdam, second edition edition.

Sawazaki, K. and Enescu, B. (2014). Imaging the high-frequency energy radiation process of a main shock and its early aftershock sequence: The case of the 2008 Iwate-Miyagi Nairiku earthquake, Japan. *Journal of Geophysical Research: Solid Earth*, 119(6):4729–4746.

Schoenberg, F. P. (2003). Multidimensional residual analysis of point process models for earthquake occurrences. *J. Am. Stat. Assoc.*, 98:789–795(7).

Schoenberg, F. P., Chang, C., Keeley, J., Pompa, J., Woods, J., and H., X. (2007). A critical assessment of the burning index in Los Angeles County, California. *International Journal of Wildland Fire*, 16:473?483.

Siebert, L. and Simkin, T. (2002). Volcanoes of the world: an illustrated catalog of holocene volcanoes and their eruptions, smithsonian institution, global volcanism program digital information series, gvp-3.

Utsu, T., Ogata, Y., and Matsu'ura, R. S. (1995). The centenary of the Omori formula for a decay law of aftershock activity. *Journal of Physics of the Earth*, 43(1):1–33.

Vere-Jones, D. (1970). Stochastic models for earthquake occurrence. *J. Roy. Stat. Soc. Series B (Methodological)*, 32(1):1–62 (with discussion).

Wang, T. and Bebbington, M. (2012). Estimating the likelihood of an eruption from a volcano with missing onsets in its record. *Journal of Volcanology and Geothermal Research*, 243–244:14–23.

Wang, T. and Bebbington, M. (2013). Robust estimation for the weibull process applied to eruption records. *Mathematical Geosciences*, 45(7):851–872.

Werner, M. J., Helmstetter, A., Jackson, D. D., and Kagan, Y. Y. (2011). High-resolution long-term and short-term earthquake forecasts for California. *Bulletin of the Seismological Society of America*, 101(4):1630–1648.

Yamamoto, T. (2015). Cumulative volume step-diagrams for eruptive magmas from major quaternary volcanoes in japan. Technical Report GSJ Open-File Report, No.613, Geological Survey of Japan, AIST.

Yukutake, Y., Tanada, T., Honda, R., Harada, M., Ito, H., and Yoshida, A. (2010). Fine fracture structures in the geothermal region of Hakone volcano, revealed by well-resolved earthquake hypocenters and focal mechanisms. *Tectonophysics*, 489:104–118.

Zhuang, J. (2011). Next-day earthquake forecasts by using the ETAS model. *Earth, Planet, and Space*, 63:207–216.

Zhuang, J., Ogata, Y., and Vere-Jones, D. (2002). Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, 97(3):369–380.

Zhuang, J., Ogata, Y., and Vere-Jones, D. (2004). Analyzing earthquake clustering features by using stochastic reconstruction. *Journal of Geophysical Research*, 109(3):B05301.

Zhuang, J., Ogata, Y., and Wang, T. (2017). Data completeness of the Kumamoto earthquake sequence in the JMA catalog and its influence on the estimation of the ETAS parameters. *Earth, Planets and Space*, 69(1):36.

Zipkin, J. R., Schoenberg, F. P., Coronges, K., and Bertozzi, A. L. (2015). Point-process models of social network interactions: Parameter estimation and missing data recovery. *European Journal of Applied Mathematics*, FirstView:1–28.