Advances in Data Analysis and Classification
https://doi.org/10.1007/s11634-025-00632-x

REGULAR ARTICLE

# Modelling time-inhomogeneous incomplete records of point processes using variants of hidden Markov models

Amina Shahzadi[1,2] · Ting Wang[1] · Matthew Parry[1] · Mark Bebbington[3]

Received: 25 December 2023 / Revised: 10 October 2024 / Accepted: 1 March 2025


## Abstract
Many point processes such as earthquakes or volcanic eruptions have incomplete records with the degree of incompleteness varying over time. For these point processes, the number of missing events between each pair of consecutively observed events can be a random variable that may depend on time, effecting the estimation of parameters or hazard. Such incomplete point processes can be modelled by compound renewal processes where the sum of renewal processes is a random variable because of random variable number of missing events. We propose shifted compound Poisson-Gamma and time-dependent shifted compound Poisson-Gamma renewal processes. Since the number of missing events can be regarded as an unobserved process, the proposed renewal processes are introduced to use in the framework of different types of homogeneous and inhomogeneous hidden Markov models to model the time-dependent variable number of missing events between each pair of consecutively observed events of incomplete point processes. Simulation experiments are employed to check the performance of proposed renewal processes with hidden Markov models. We apply the proposed models to the large magnitude explosive volcanic eruptions

✉ Amina Shahzadi
aminashahzadi@gmail.com

Ting Wang
ting.wang@otago.ac.nz

Matthew Parry
matthew.parry@otago.ac.nz

Mark Bebbington
m.bebbington@massey.ac.nz

1 Department of Mathematics and Statistics, The University ofOtago, Dunedin, Otago, New Zealand

2 Department of Statistics, Government College University, Lahore, Punjab, Pakistan

3 Volcanic Risk Solutions, Massey University, Private Bag 11222, Palmerston North 4442, New Zealand

Published online: 23 April 2025

database to analyze the time-dependent incompleteness and demonstrate how we estimate the completeness of the record and the future hazard rate.

**Keywords** Point process · Time-inhomogeneous incompleteness · Shifted compound Poisson-gamma renewal process · Inhomogeneous hidden Markov models · The LaMEVE database · Hazard

**Mathematics Subject Classification** 62M05

## 1 Introduction

Point process modelling is widely used in many social and natural sciences. Examples include epidemiology (Quesada et al. 2017), criminology (Mohler et al. 2011; Mohler 2013), telecommunications (Decreusefond and Moyal 2013; Zipkin et al. 2016), geology (Bebbington and Lai 1996; Wang and Bebbington 2012; Bebbington 2013), seismology (Veneziano et al. 1987; Ogata 1998; Wang and Bebbington 2013), hydrology (Kaczmarska et al. 2014) and many others. In many applications, point process data have missing events due to discontinuous methods of data collection, e.g., missing crime events in criminal records because of transmission of data from hard copy to digitized form (Tucker et al. 2019). The missingness of events is also a problem common to most point process records of geophysical events. For example, in volcanic and earthquake records, smaller events are less likely to be recorded than larger events (Furlan 2010; Brown et al. 2014; Veneziano et al. 1987; Ogata and Katsura 1993). For such incomplete records, point process modelling may provide biased estimation of hazard or parameters.

Rapid advances in science and technology have substantially increased the amount of data in different fields during the last few decades, resulting in less incompleteness of point process records in recent times. However, the older part of the historical records is substantially incomplete for many reasons, resulting in time-inhomogeneity in the completeness of point process records. Take the volcanic eruption records for example. The records are constructed based on available geological and historical records. The older part of the volcanic eruption records is mostly based on the geological records. Small eruptions leave behind minimal traces in the geological record and are easily removed by erosion (Coles et al. 2006; Furlan 2010). Also, many geological records were destroyed when the ice age ended at the start of the Holocene (Deligne et al. 2010). Consequently, missigness of events in the older part of the volcanic eruption records becomes more pronounced. Various authors have reported this time-inhomogeneous under-recording in volcanological literature (Guttorp and Thompson 1991; Furlan 2010; Siebert et al. 2010; Deligne et al. 2010; Crosweller et al. 2012; Brown et al. 2014; Kiyosugi et al. 2015; Wang et al. 2020).

The number of missing events between each pair of consecutively observed events for point processes may be considered as a random variable. In many real-world applications, the random variable number of missing events tends to decrease over time, where the older part is likely to have more missing events. As the missingness of events is unknown and thus a hidden process, it may be advantageous to model such unobserved missingness in the framework of hidden Markov models (HMMs). For example, Wang and Bebbington (2012) modelled the incompleteness of volcanic eruption records using HMMs in combination with renewal processes. Because of the time homogeneity in the transition matrix, HMMs may be used to model incompleteness of point process records if the missingness probability does not depend on time, and if events are missing independently. In the case of Wang and Bebbington (2012), the records were from sedimentary cores, and so HMMs could be used. However, when the record is a composite of observations from different sources based on different observational and recording techniques (Furlan 2010; Deligne et al. 2010; Kiyosugi et al. 2015), these assumptions cannot necessarily be made. In such conditions, to model time-dependent missingness in point process records, Shahzadi et al. (2023) introduced inhomogeneous hidden semi-Markov models (IHSMMs), an extension of hidden semi-Markov model (HSMM) (Ferguson 1980; Guédon and Cocozza-Thivent 1989; Bulla 2006; Barbu and Limnios 2008). In an inhomogeneous HSMM (IHSMM), the explicit state duration of the underlying semi-Markov chain depends on time. In Wang and Bebbington (2012) and Shahzadi et al. (2023), the authors assumed that each state of hidden process represents an arbitrary number of missing events between each pair of consecutively observed events in a point process record.

Since the number of missing events between each pair of consecutively observed events can be a random variable which may also depend on time, we propose shifted compound Poisson-gamma (SCPG) and time-dependent SCPG (TSCPG) renewal processes as the observed processes in the structure of hidden Markov type models. In SCPG and TSCPG renewal processes, the number of missing events between each pair of consecutively observed events is (1) a shifted Poisson random variable, and (2) a time-dependent shifted Poisson random variable, respectively. When using HMMs for time-inhomogeneous and incomplete point process records, the transition probability from a state with missing events to a state with no missing event may not remain constant over time. As a result, the sojourn time within the state of completeness may increase, as recent segments of the records have fewer missing events than older segments. We propose an inhomogeneous hidden Markov models (IHMMs) with time-varying transition probabilities defined by multinomial logistic functions (Diebold et al. 1994; Filardo 1994; Hughes and Guttorp 1994, 1999) to model the hidden process. The novelty (advance on Wang and Bebbington (2012) and Shahzadi et al. (2023)) is that we are transferring complexity from the HMM part to the observed distribution part, hopefully resulting in a more efficient and interpretable model. We propose the observed part in HMM type models to follow SCPG and TSCPG renewal processes.

In Sect. 2, we introduce the observed (or sometimes known as the emission) processes for the hidden Markov type models incorporating SCPG and TSCPG renewal processes. In Sect. 3, we introduce the different variants of hidden Markov models. The ultimate goal of the proposed models is to estimate the time-dependent completeness level or recording rate of a real-world point process and to obtain better hazard estimates. In Sect. 5, we test the consistency of parameter estimators of the proposed models via simulation studies. In Sect. 6, the proposed models are applied to the Large Magnitude Explosive Volcanic Eruptions (LaMEVE) record during the last

10000 years. In this case, we use a Poisson process for the ground process. Moderately large global volcanic eruption events can be considered a superposition of many independent processes from different geographical regions and thus, if completely observed, could be modelled using a Poisson process (Guttorp and Thompson 1991; Cruz-Reyna 1991, 1993). This solves our identifiability problem, allowing us to model the missing data explicitly. Section 7 concludes the paper.

## 2 Shifted compound Poisson-gamma renewal processes

Suppose we have a point process with event occurrence times $H_0 < H_1 < \ldots < H_t < \ldots$ at discrete-time points $0,\ 1,\ \ldots,\ t,\ \ldots$. The interevent times are given by $X_t = H_t - H_{t-1},\quad t = 1,\ 2,\ \ldots,$ with $H_0 = 0$ by convention. When the interevent times $X_t$ are independent and identically distributed, the point process is called a renewal process. The hazard rate or point process conditional intensity is defined as:

$$\boldsymbol{\lambda}(h_t) = \frac{f(h_t - h_s)}{1 - F(h_t - h_s)}, \qquad h_t > h_s, \tag{1}$$

where $h_s$ is the occurrence time of the most recent event before time $h_t$, $F(\cdot)$ is the cumulative distribution function (CDF) of $X_n$, and $f(\cdot)$ is the corresponding probability density function (PDF). The time $h_t - h_s$ is called the elapsed time that determines the likelihood of the next event since the occurrence of the last event. Previous events contribute through the parameter estimates. If we assume that the interevent times of a point process follow a gamma distribution with the PDF

$$f(x_t) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x_t^{\alpha-1} \exp(-x_t/\beta), \quad x_t > 0, \tag{2}$$

where, $\alpha$ and $\beta$ are shape and scale parameters, respectively, and $\Gamma$ is the gamma function defined as $\Gamma(s) = \int_0^{\infty} w^{s-1}e^{-w}dw,\ s > 0$, then the point process forms a gamma renewal process. When $\alpha = 1$ in (2), we have a homogeneous Poisson process. The gamma distribution has significance because of its properties and numerous applications in a wide variety of fields. It generalizes the exponential distribution and possesses an intrinsic property of additivity. The gamma distribution also provides the probability distribution for the length of time required for the $k$th arrival in a Poisson process as the sum of $k$ exponential random variables is a gamma random variable.

As discussed in the Introduction, the number of missing events between each pair of consecutively observed events in a point process record, for example, volcanic eruption records, is not known. We assume the number of such missing events to follow the shifted Poisson distribution. The rationale behind using the shifted Poisson distribution is that we will consider one state with no missing events (the state of completeness), and the others with variable number of missing events (the states of incompleteness) via the SCPG distribution in different types of HMMs. Suppose there are $N$ missing events between event occurrence times $H_{t-1}$ and $H_t$. Then the observed interevent time $X_t$ is an aggregate of $N + 1$ interevent times $Y_1, \ldots, Y_{N+1}$ and can be written as the sum

$$X_t = \sum_{k=1}^{N+1} Y_k, \tag{3}$$

where $Y_k$ are independent and identical gamma variables with the PDF as in (2). Now assume that $N$ is a random variable and follows a shifted Poisson distribution with parameter $\lambda$,

$$\Pr(N = k) = \frac{\lambda^{k-1}\exp(-\lambda)}{(k-1)!}, \qquad k = 1,\ 2,\ \dots . \tag{4}$$

When $Y_k$ are independent of each other and of $N$, the interevent time $X_t$ in (3) follows a compound of gamma and shifted Poisson distributions and is called an SCPG distribution. We call the resulting point process an SCPG renewal process. It follows that if there is no missing event, then the interevent time $X_t$ has a gamma distribution, else it will follow an SCPG distribution. Thus, for the interevent times $X_t$ in (3), the conditional distribution of $X_t$ given $N = k$ follows a gamma distribution with parameters $((k+1)\alpha,\ \beta)$. The marginal PDF of $X_t$ is then obtained as

$$\begin{aligned} f(x_t) &= \sum_{k=1}^{\infty} \Pr(N = k) g(x_t;\ (k+1)\alpha,\ \beta \mid N = k) \\ &= \exp(-\lambda - x_t/\beta)\ \frac{x_t^{2\alpha-1}}{\beta^{2\alpha}} \sum_{k=0}^{\infty} \frac{1}{\Gamma(k\alpha + 2\alpha)\, k!} \left(\frac{\lambda x_t^{\alpha}}{\beta^{\alpha}}\right)^k, \end{aligned} \tag{5}$$

where, $x_t > 0$. The above SCPG distribution corresponds to the continuous and positive part of a mixed compound Poisson-gamma (MCPG) distribution of Tweedie family of distributions when the power parameter $p$ of the family is $1 < p < 2$ (Tweedie 1984; Smyth and Jorgensen 2002). In particular, an SCPG distribution is a distribution for the sum of $N + 1$ independent and identical gamma variables with $N$ being a shifted Poisson distributed variable. The PDF in (5) has an implicit form and can be evaluated in a variety of ways. We consider the asymptotic expansion of Wright's generalized Bessel function proposed by Paris (2010, 2017) and Paris and Vinogradov (2016) and the direct evaluation method proposed by Dunn and Smyth (2001, 2005) to evaluate the SCPG PDF. These methods and a simulation study for an SCPG distribution with the PDF evaluated by the two methods are detailed in the Supplementary Material.

Note that when $\alpha = 1$, the PDF in (2) reduces to an exponential distribution, and the distribution of $X_t$ in (5) is still an SCPG distribution, but with two parameters $(\beta,\ \lambda)$. If $Y_k$ in (3) are independently and identically distributed Weibull random variables, then $X_t$ in (3) will follow a shifted compound Poisson-Weibull distribution. Thus, different interevent time distributions on the positive real line for $Y_k$ produce different classes of shifted compound Poisson renewal processes.

The SCPG distribution in (5) models the interevent times following a gamma distribution with a shifted Poisson distributed number of missing events between each pair of consecutively observed events in an incompletely observed point process record. The incompleteness of a record may depend on time when, for example, the older part of the record has more missing events than the recent part. Consequently, the shifted Poisson distributed number of missing events between each observed pair in an SCPG renewal process may depend on time.

To model time-dependent missing events, we propose that the expected number of missing events represented by $\lambda$ in (5) depends on time. In order to formulate $\lambda$ as a function of time, we use the generalized logistic function and denote it by $\lambda(n)$ at time $n$. In a renewal process, the likelihood of the next event depends on the elapsed time since the last event occurred. Therefore, $\lambda(t)$ can be defined to be a function of the last occurrence time $h_{t-1}$. It follows that

$$\lambda(t) = \frac{M}{1 + \exp(R\, h_{t-1} - P)} + K, \tag{6}$$

where $M,\ R,\ K > 0$ and $-\infty < P < \infty$ are the parameters. The parameters $M$ and $K$ determine the limiting values of $\lambda(t)$ such that $\lambda(t)$ decreases from $M + K$ to $K$. The parameter $R$ represents the rate of change in $\lambda(t)$, and the parameter $P$ controls the point of inflection. Note that $\lambda(t)$ is a decreasing function of the last occurrence time as the mean number of missing events may decrease from the older part of a record to the recent part. The shifted Poisson distribution with parameter $\lambda(t)$ becomes a time-dependent shifted Poisson distribution and is given by

$$\Pr(N_t = k_t) = \frac{e^{-\lambda(t)}[\lambda(t)]^{k_t-1}}{(k_t - 1)!}, \qquad k_t = 1,\ 2,\ \ldots \tag{7}$$

Replacing a shifted Poisson distribution with a time-dependent shifted Poisson distribution in (5), we have a TSCPG distribution with PDF

$$f(x_t) = \exp(-\lambda(t) - x_t/\beta)\ \frac{x_t^{2\alpha-1}}{\beta^{2\alpha}} \sum_{k_t=0}^{\infty} \frac{1}{\Gamma(k_t\alpha + 2\alpha)\, k_t!} \left( \frac{\lambda(t) x_t^{\alpha}}{\beta^{\alpha}} \right)^{k_t}, \tag{8}$$

where, $x_t > 0$. A TSCPG distribution can be used to model the interevent times of a partially observed gamma renewal process with time-dependent Poisson distributed number of missing events between each pair of consecutively observed events. The PDF for a TSCPG distribution can be evaluated using the methods to evaluate an SCPG distribution discussed in the Supplementary Material. However, $\lambda$ needs to be replaced by $\lambda(t)$ wherever it appears in those methods.

## 3 Variants of hidden Markov models

In this section, we introduce SCPG and TSCPG renewal processes as the observed processes in different types of HMMs for modelling point processes with time-dependent

**Table 1** An $m-$ state HMM type model

| | State definition | Observed process |
|---|---|---|
| State 1 | The state of completeness represents no missing events between each pair of consecutively observed events | the interevent times $X_t = H_t - H_{t-1}, \quad t = 0, 1, 2, \ldots$, of the observed process in State 1 is assumed to follow a gamma distribution given by (2) |
| State $j,\ j = 2, \ldots, m$ | (a) The state of incompleteness represents the shifted Poisson-distributed number of missing events between each pair of consecutively observed events | the interevent times $X_t = H_t - H_{t-1}, \quad t = 0, 1, 2, \ldots$, of the state-dependent observed process in State $j$ is an SCPG renewal process given by (5) with parameters $\lambda_j, \alpha, \beta$ |
| State $j,\ j = 2, \ldots, m$ | (b) The state of incompleteness represents the time-dependent shifted Poisson-distributed number of missing events between each pair of consecutively observed events | the interevent times $X_t = H_t - H_{t-1}, \quad t = 0, 1, 2, \ldots$, of the state-dependent observed process in State $j$ is a TSCPG renewal process given by (8) with parameters $\lambda_j(t), \alpha, \beta$ |

incompleteness. The occurrence times, interevent times, and counting process are all equivalent representation of a point process. We use a discrete-time hidden Markov type model for modelling interevent times of a point process, since a discrete-time Markov chain detects recurring temporal patterns and determines change points (Allen and Wang 2024). We define an $m-$state hidden Markov type model in Table 1. If States $2, \ldots, m$ follow definition (a), we have an $m-$state discrete-time hidden Markov type model with SCPG renewal process as the observed process; if States $2, \ldots, m$ follow definition (b), there is an $m-$state discrete-time hidden Markov type model with TSCPG renewal process as the observed process. With these state definitions and the corresponding observed processes, we define the variants of HMMs in the following subsections.

### 3.1 Homogeneous and inhomogeneous hidden Markov models

Homogeneous HMMs become inhomogeneous when the transition probabilities are time-varying (Hughes and Guttorp 1994)[1] and the underlying first-order Markov chain becomes an inhomogeneous first-order Markov chain. Special cases of IHMMs can be obtained using different parameterizations of the transition probabilities which preserve the row stochastic property of transition probability matrix at each time (Hughes and Guttorp 1994; Diebold et al. 1994; Filardo 1994; Filardo and Gordon 1998; Otranto 2008).

[1] Note that the authors originally used the name 'nonhomogeneous hidden Markov models (NHMMs)' for HMMs with time-varying transition probabilities. In this paper, we prefer the term 'inhomogeneous' for HMMs with time-varying transition probabilities.

We propose a special case of IHMMs by defining the time-varying transition probabilities using multinomial logistic functions for an $m-$state IHMM. In a multinomial logistic function, one of $m$ categories is taken as a reference category. For our particular case of IHMMs with the aforementioned state definitions and state-dependent observed processes, we consider State 1 as a reference state, since State 1 is the state of completeness. Based on the renewal process characteristic, the transition probabilities at time $t$ are assumed to be a function of the occurrence time at $t-1$, i.e., $h_{t-1}$. Thus, denoting the state of the underlying Markov chain at time $t$ by $S_t$, we define the time-varying transition probabilities $\Pr(S_t = j \mid S_{t-1} = i,\ H_{t-1} = h_{t-1})$ of the underlying inhomogeneous Markov chain using the following parameterization

$$\begin{aligned} a_{i1}(t) &= \frac{1}{1 + \sum_{k=2}^{m} \exp(-r_{ik}h_{t-1} + w_{ik})}, \qquad i = 1, \ldots, m, \\ a_{ij}(t) &= \frac{\exp(-r_{ij}h_{t-1} + w_{ij})}{1 + \sum_{k=2}^{m} \exp(-r_{ik}h_{t-1} + w_{ik})}, \qquad i = 1, \ldots, m,\ \ j = 2, \ldots, m, \end{aligned} \tag{9}$$

where $r_{ij} > 0$ measures the rate of change and $-\infty < w_{ij} < \infty$ controls the points of inflection. The number of parameters to be estimated for an $m \times m$ transition probability matrix is $2m(m-1)$. The above parameterization of time-varying transition probabilities ensures that the transition probabilities from States $j = 2, \ldots, m$ to State 1 are increasing functions of event occurrence times, whereas the transition probabilities from States $i = 1, 2, \ldots, m$ to States $j = 2, \ldots, m$ can be increasing or decreasing, to reflect the fact that most real-world time-dependent point process records are more likely to be in the state of completeness (State 1) than in a state of incompleteness (States $2, \ldots, m$) in recent times. It also ensures that the transition probability matrix is row stochastic, that is, $\sum_{j=1}^{m} a_{ij}(t) = 1$ for $i = 1, \ldots, m$ and $t > 1$.

The likelihood function of the IHMM defined above can be obtained from the usual formula of the likelihood function of an HMM by replacing the transition probabilities with the time-varying transition probabilities (Zucchini and MacDonald 2009; Wang 2010). That is,

$$\begin{aligned} \mathcal{L}(\Theta) = \Pr(X_{1:T} \mid H_{1:T};\ \Theta) &= \sum_{all\ S} \Pr(X_{1:T},\ \ S_{1:T} \mid H_{1:T};\ \Theta) \\ &= \sum_{all\ S} \Pr(X_1,\ \mid S_1;\ \Theta) \Pr(S_1 \mid \Theta) \prod_{t=2}^{T} \Pr(X_t \mid S_t;\ \Theta) \Pr(S_t \mid S_{t-1},\ H_{t-1};\ \Theta) \\ &= \sum_{s_1, \ldots, s_T = 1}^{m} \pi_{s_1} f_{s_1}(x_1) \prod_{t=2}^{T} a_{s_{t-1} s_t}(t) f_{s_t}(x_t), \end{aligned} \tag{10}$$

where, $\Theta = \{\pi_j,\ r_{ij},\ w_{ij}\ K_j,\ M_j,\ P_j,\ R_j,\ \alpha,\ \beta;\ i,\ j = 1, \ldots, m\}$ is the set of model parameters, with $\pi_j$ being the initial probabilities of the underlying inhomogeneous first-order Markov chain; and $f_{s_t}(x_t)$ is the SCPG or time-dependent SCPG distribution with parameters depending on which state $s_t$ is in.

The likelihood function can be evaluated recursively using the forward algorithm. The forward algorithm is defined by

$$\alpha_t(j) = \Pr(X_1,\ \ldots,\ X_t,\ S_t = j \mid H_{t-1};\ \Theta),$$

which can be calculated inductively as

$$\alpha_1(j) = \Pr(X_1,\ S_1 = j \mid \Theta) = \pi_j f_j(x_1)$$
$$\alpha_t(j) = \sum_{i=1}^{m} \alpha_{t-1}(i)\, a_{ij}(t)\, f_j(x_t), \quad t = 2,\ \ldots,\ T,\ \ j = 1,\ \ldots,\ m.$$

Whence, the likelihood function is

$$L(\Theta) = \sum_{j=1}^{m} \Pr(X_{1:T},\ \ S_T = j \mid H_{1:T};\ \Theta) = \sum_{j=1}^{m} \alpha_T(j). \tag{11}$$

In order to avoid the numerical issues while evaluating the likelihood function in (11) using the forward algorithm in an IHMM, we use the scaling procedure in a similar manner as for HMMs. Also, we can easily employ the backward algorithm, Viterbi path and computing the smoothed probabilities of an HMM to an IHMM by replacing the transition probabilities with time-varying transition probabilities discussed in Zucchini and MacDonald (2009).

Using the definitions of states in Table 1 in an HMM and IHMM, we have the following four models. Note that in an $m-$state model, the definition of State 1 remains same.

1. An $m-$state HMM with an SCPG renewal process with State $j,\ j = 2, \ldots, m$, following definition (a) in Table 1.
2. An $m-$state HMM with a TSCPG renewal process with State $j,\ j = 2, \ldots, m$, following definition (b) in Table 1.
3. An $m-$state IHMM with an SCPG renewal process with State $j,\ j = 2, \ldots, m$, following definition (a) in Table 1.
4. An $m-$state IHMM with a TSCPG renewal process with State $j,\ j = 2, \ldots, m$, following definition (b) in Table 1.

### 3.2 Homogeneous and inhomogeneous hidden semi-Markov models

A homogeneous HSMM is an extension of a homogeneous HMM when the state durations are explicitly modelled using a discrete distribution, which transforms a Markov chain into a semi-Markov chain. A semi-Markov chain satisfies a restricted memoryless property: it is memoryless for transitions between distinct states but not memoryless while staying in a particular state (Barbu and Limnios 2008).

HSMMs may not be suitable when modelling nonstationary point process. For that reason, Shahzadi et al. (2023) proposed a general class of IHSMMs. When the explicit state duration depends on time, an HSMM is called an inhomogeneous hidden semi-Markov model (IHSMM) and the underlying semi-Markov chain becomes an inhomogeneous semi-Markov chain. In order to model the time-dependent number of missing events between each pair of consecutively observed events in a point process record, the authors also introduced a special case of IHSMMs with the gamma renewal process as the observed process and the time-varying shifted Poisson distributed state duration. In an $m$-state model, State 1 was defined as the state of completeness and States $2, \ldots, m$ were defined as the states of incompleteness to represent the fixed numbers of missing events between each pair of consecutively observed events. This may not be apt when the number of missing events has a large range, as in a long historical record. Also, as mentioned in Sect. 2, the number of missing events between each pair of consecutively observed events in a point process record may depend on time.

Employing the state definitions in Table 1, the SCPG and TSCPG renewal processes can be incorporated into the IHSMMs with state duration following a time-dependent shifted Poisson distribution (Shahzadi et al. 2023). Assuming that record completeness increases over time, for point process records with time-dependent missingness, we use the following distributions for the state duration:

$$p_j(d, t) = \exp(-\mu_j(t)) \frac{[\mu_j(t)]^{d-1}}{(d-1)!}, \qquad d = 1, 2, \ldots, \tag{12}$$

and hence the process sojourns longer (shorter) in states with complete (incomplete) observations, where $\mu_j(t) > 0, \;\; j = 1, 2, \ldots, m$, can be interpreted as the expected sojourn time in state $j$ depending on time $t$. For State 1, the state of completeness, $\mu_1(t)$ is defined as the increasing function of time given by

$$\mu_1(t) = \frac{A_1}{1 + \exp(-C_1\, h_{t-1} + B_1)} + D_1, \tag{13}$$

and for State $j$, $j = 2, \ldots, m$, $\mu_j(t)$ is defined by

$$\mu_j(t) = \frac{A_j}{1 + \exp(C_j\, h_{t-1} - B_j)} + D_j, \tag{14}$$

where, $A_j, \; C_j, \; D_j \; > 0$ and $-\infty < B_j < \infty$ for $j = 1, 2, \ldots, m$ are the parameters. Note that the time-dependent expected state durations for the states of incompleteness in (14) in the IHSMM have the same formulation as we defined for the time-dependent expected number of missing events in a TSCPG renewal process, but with different symbols for the parameters.

The detailed formulation of an $m-$state IHSMMs regarding likelihood function, forward-backward algorithm, viterbi paths and simulation algorithm can be found in Shahzadi et al. (2023) which can be easily reduced to HSMMs when the state duration is time-independent. Using the definitions of states in Table 1 in an HSMM and IHSMM, we have the following four models. Note that in an $m-$state model, the definition of State 1 remains the same. Also, explicit state durations of underlying semi-Markov chain in an HSMM follows shifted Poisson distributions, and time-dependent explicit state durations of underlying inhomogeneous semi-Markov chain in an IHSMM follows time-dependent shifted Poisson distributions.

1. An $m-$state HSMM with an SCPG renewal process with State $j$, $j = 2, \ldots, m$, following definition (a) in Table 1.
2. An $m-$state HSMM with a TSCPG renewal process with State $j$, $j = 2, \ldots, m$, following definition (b) in Table 1.
3. An $m-$state IHSMM with an SCPG renewal process with State $j$, $j = 2, \ldots, m$, following definition (a) in Table 1.
4. An $m-$state IHSMM with a TSCPG renewal process with State $j$, $j = 2, \ldots, m$, following definition (b) in Table 1.

## 4 Fitting of hidden Markov type models

Fitting of a hidden Markov type model can be carried out either using the direct numerical optimization (DNO) method or the expectation maximization (EM algorithm. The estimation of HMMs is an incomplete data problem as the underlying Markov chain is an unobserved process. Therefore, the primary method is to fit an HMM by the EM algorithm (Dempster et al. 1977; Baum et al. 1970). The E-step of the EM algorithm generally involves integrals and is computationally complex. The M-step involves derivatives and its computational complexity depends on the state-dependent probability distribution. The solution is not unique if closed-form solutions are not available, for example, in the case of the SCPG or TSCPG state-dependent distributions, then numerical maximization by a Newton-type algorithm or some variant has to be carried out. Also, EM algorithm is sensitive to the initial values of parameters and the convergence rate of the EM algorithm is very slow for complex models (Dempster et al. 1977; Wu 1983; Campillo and Gland 1989; Jordan and Xu 1996; McLachlan and Krishnan 2008.

DNO algorithms can be used as alternatives to the EM algorithm for the maximum likelihood estimation of parameters in HMMs (Zucchini and MacDonald 2009; Bulla and Berzel 2008). These algorithms have seen significant growth due to their speed of convergence relative to the EM algorithm (Campillo and Gland 1989) and their availability in the form of packages in the free statistical software, e.g., R (R Core Team 2017). In this study, we fit various hidden Markov type models defined in Subsections 3.1 and 3.2 using DNO of log-likelihood functions available in R (R Core Team 2017).

## 5 Simulation study

In Sect. 3, we defined homogeneous and inhomogeneous hidden Markov type models each with $m$ states following state definitions in Table 1 and SCPG and TSCPG as state-dependent distributions, producing eight models listed in Table 2.

**Table 2** Variants of HMM with SCPG and TSCPG renewal processes

| No | Type of HMM | Observed process |
|---|---|---|
| 1 | HMM | SCPG |
| 2 | IHMM | |
| 3 | HSMM | |
| 4 | IHSMM | |
| 5 | HMM | TSCPG |
| 6 | IHMM | |
| 7 | HSMM | |
| 8 | IHSMM | |

For each of eight models listed in Table 2, we performed simulation studies to inspect the behaviour of parameter estimators. We present here the simulation study results of Model 6 only. The results for the remaining models can be found in the Supplementary Material (Section S.5).

We consider a 2−state model for an IHMM with TSCPG renewal process. State 1 represents no missing event between each pair of consecutively observed events in a point process record, and the observed interevent times in State 1 follow a gamma distribution. State 2 represents time-dependent Poisson distributed number of missing events between each pair of consecutively observed events in a point process record, and the observed interevent times in State 2 follow a TSCPG distribution. The parameter values are catalogued in Table 3. We generate 100 simulations each of sample sizes 200, 500, 1000 and 2000. For maximum likelihood estimation, the direct numerical minimization of the negative log-likelihood function of the model is performed in R (R Core Team 2017). The maximum likelihood estimates (MLEs) of parameters are plotted in Fig. 1. Boxplots of MLEs are constructed for those parameters which are

**Table 3** The 2−state IHMM with TSCPG renewal process

| **IHMM** | Parameters | | | State 1 | State 2 |
|---|---|---|---|---|---|
| IMC | $\pi_j$ | | | 0.40 | 0.60 |
| | $a_{ij}(t)$ | $r_{ij}$ | State 1 | – | 0.16 |
| | | | State 2 | – | 0.24 |
| | | $w_{ij}$ | State 1 | – | 10 |
| | | | State 2 | – | 15 |
| TSCPG process | $\alpha$ | | | 2 | 2 |
| | $\beta$ | | | 0.25 | 0.25 |
| | $\lambda_j(t)$ | $M_j$ | | – | 7 |
| | | $P_j$ | | – | 10 |
| | | $R_j$ | | – | 0.30 |
| | | $K_j$ | | – | 1 |

independent of time (Fig. 1c), whereas time-dependent parameter estimates from the 100 simulations are simply plotted against the occurrence times (Fig. 1a and b).

From Fig. 1, we observe that the interquartile range of boxplots for the MLEs reduces with increasing sample size and medians of the parameter estimates from different models converge to their respective true parameter values, which demonstrates the consistency of the parameter estimators.

From the simulation experiments of other models in Table 2, the proposed SCPG and TSCPG renewal processes produce satisfactory simulation results with different types of HMMs (see Section S.5 in the Supplementary Material).

## 6 Application to the LaMEVE database

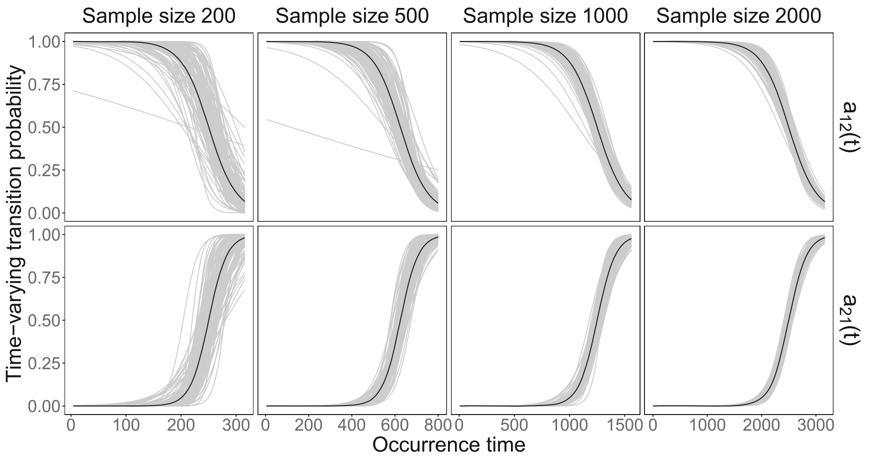
The aim of this section is to present a case study, in which we investigate the time-dependent incompleteness of a global volcanic eruption record using the proposed SCPG and TSCPG renewal processes in combination with different types of HMMs.

### 6.1 The LaMEVE database

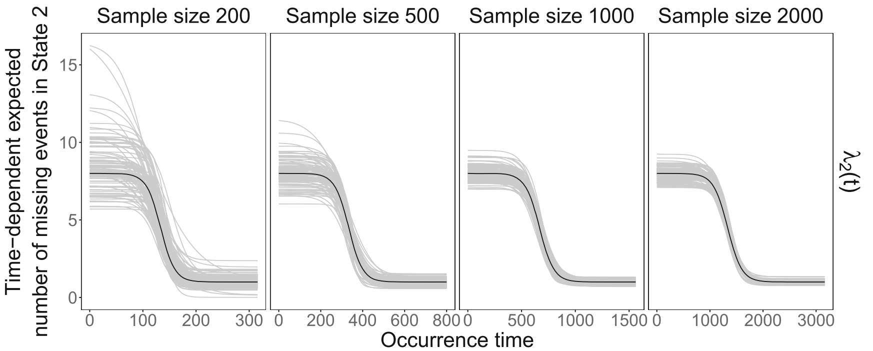
Crosweller et al. (2012) created a global database of Quarternary Large Magnitude Explosive Volcanic Eruptions (LaMEVE). The main objective of developing this database was to provide the necessary information on global explosive volcanism needed to facilitate the estimation of hazard and risk from large volcanic eruptions and to evaluate the effects of volcanism on society and environment (Crosweller et al. 2012). This database was developed in the Volcanic Global Risk Indentification and Analysis Project (VOGRIPA) framework and is made publicly available at https://www.bgs.ac.uk/.

A global magnitude scale has been established by Pyle (2015) which is equivalent to the logarithm of the eruption volume. The LaMEVE database includes all global volcanic eruptions of magnitude 4 or more which are dated with known source volcanos, and it spans over the last 2.5 million years, which accounts for the Quaternary period of geological timescales. Brown et al. (2014) identified the substantial gaps in the recording of eruption events from global to local scales. In general, the under-recording of eruption events was found to depend on time, becoming less pronounced from AD 1450 to the present day. The incompleteness of the record gets worse going back in time for small magnitude eruptions, which is consistent with the findings elsewhere, for example, Simkin (1993), Furlan (2010) and Deligne et al. (2010).

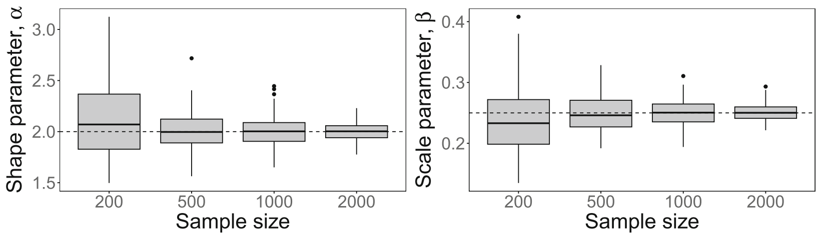
In this section, we investigate and model the time-dependent incompleteness of the LaMEVE database using our proposed models in order to estimate the time-dependent record completeness and hazard rate. We extracted the volcanic eruptions of the Holocene period, the current geological epoch which began about 10,000 years ago and continues through today. The event occurrence times have been recorded in year before present (YBP) in the record, where YBP means the year before AD 1950 (e.g. 600 YBP = AD 1950 − AD 1350). The eruption occurrence times and their magnitudes are plotted in Fig. 2, which appears to suggest that the incompleteness of




(a) MLEs of time-varying transition probabilities of IHMM. The solid black curve represents the true parameter curve.




(b) MLEs of time-dependent expected number of missing events for TSCPG distribution in State 2. The solid black curve represents the true parameter curve.




(c) MLEs of shape and scale parameters of TSCPG distribution in State 2. The dashed line represents the true parameter value.

**Fig. 1** MLEs of parameters of the 2−state IHMM with the TSCPG renewal process in Table 3 from 100 simulations of different sample sizes

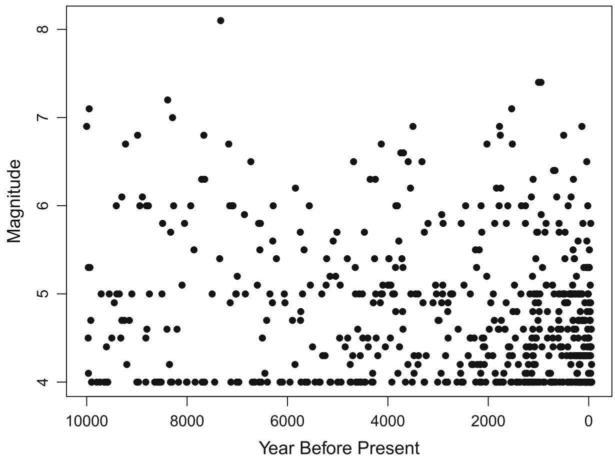



**Fig. 2** Global volcanic eruptions in the Holocene period of the LaMEVE database with magnitude $\geq 4$

the record increases, particularly, for small eruptions with magnitude $< 6$ when we go back in time.

Note that the occurrence pattern of moderately large global volcanic eruption events should exhibit Poissonian behaviour, being a superposition of many independent processes (Guttorp and Thompson 1991; Cruz-Reyna 1991, 1993). Because of this Poissonian behaviour, the interevent times from global volcanic eruption onsets follow an exponential distribution. Therefore, when we fit the models, we assume that the shape parameter $\alpha$ of gamma, SCPG and TSCPG distributions is 1. Hence, the number of parameters to be estimated is reduced by one.

### 6.2 Data analysis

We consider HMMs, IHMMs, IHSMMs and IHSMMs with 2, 3, 4, 5 and 6 hidden states for the global volcanic eruption record shown in Fig. 2. In these models, State 1 represents no missing events between each pair of consecutively observed events in the record, and the observed interevent times in State 1 follow a gamma distribution. Other states represent a variable number of missing events between each pair of consecutively observed events in the record, and the observed interevent times in these states follow different SCPG distributions. When the variable number of missing events in States 2 to 6 depends on time, the observed interevent times are assumed to follow a TSCPG distribution.

There have been studies on how to choose the information criteria for HMMs (Pohle et al. 2017; Buckby et al. 2023). Buckby et al. (2023) suggested to use the selection criteria based on sample size, emission probability distribution and long/short sojourn times in each state. Pohle et al. (2017) proposed a pragmatic approach to select the number of states by using selection criteria as guidance, inspecting Viterbi path and then checking the model using residual analysis. Based on their studies, to select the best fit model from a set of candidate models, we use the Akaike

Information Criterion (AIC) (Akaike 1974) and Viterbi path combined with residual analysis of the point process (Berman 1983; Ogata 1988; Wang et al. 2012). The residual analysis states that the rescaled times $\tau_k$'s defined by the transformation $\tau_k = \Lambda(h_k) = \int_0^{h_k} \lambda(h)\, dh,\ k = 1, \ldots, T$ form a Poisson process with unit rate if the model is a good fit to the data ( Daley and Vere-Jones 2003, Theorem 7.4.I), where $\boldsymbol{\lambda}(h)$ is the estimated condition intensity function (1). For this purpose, we can plot the cumulative number of events versus transformed times $\tau_k$, with 95% and 99% confidence intervals of the Kolmogrov-Smirnov (KS) statistic. Also, $E_k = \tau_k - \tau_{k-1}$ are independent exponential random variables with a unit mean or $U_k = 1 - \exp(-E_k)$ are independent uniform random variables on $[0,\ 1)$, if the fitted model is valid.

The various types of HMM models are fitted to the data with a time unit of decades using an unconstrained minimization of the negative log-likelihood functions. The fitted models along with the number of parameters ($k$), maximum log-likelihood (MLL) and AIC values are listed in Table 4. The 5−state HMM with TSCPG has the smallest AIC. The residual process of this model in Fig. S.12 in the Supplementary Material falls within the confidence intervals but deviates further away from a stationary Poisson process with unit rate compared to the 4−state IHMM with TSCPG renewal process.

**Table 4** No. of parameters ($k$), MLL and AIC values of fitted HMMs

| Model | | SCPG RP | | | TSCPG RP | | |
|---|---|---|---|---|---|---|---|
| | | $k$ | MLL | AIC | $k$ | MLL | AIC |
| HMM | 2−state | 5 | −804.0 | 1618.1 | 8 | −779.6 | 1575.1 |
| | 3−state | 11 | −754.5 | 1530.9 | 17 | −702.6 | 1439.2 |
| | 4−state | 19 | −727.4 | 1492.8 | 28 | −668.9 | 1393.4 |
| | 5−state | 29 | −706.0 | 1470.0 | 41 | −646.9 | 1375.7 |
| | 6−state | 41 | −696.1 | 1474.2 | 56 | −643.1 | 1398.2 |
| IHMM | 2−state | 7 | −764.6 | 1543.1 | 10 | −747.0 | 1514.0 |
| | 3−state | 17 | −714.4 | 1462.9 | 23 | −686.5 | 1419.1 |
| | 4−state | 31 | −680.5 | 1423.0 | 40 | −655.9 | 1391.9 |
| | 5−state | 49 | −660.8 | 1419.7 | 61 | −640.6 | 1403.1 |
| | 6−state | 71 | −648.4 | 1438.7 | 86 | −626.3 | 1424.5 |
| HSMM | 2−state | 5 | −803.6 | 1617.3 | 8 | −778.1 | 1572.1 |
| | 3−state | 11 | −772.0 | 1566.0 | 17 | −714.9 | 1463.8 |
| | 4−state | 19 | −737.1 | 1512.1 | 28 | −678.4 | 1412.8 |
| | 5−state | 29 | −705.9 | 1469.8 | 41 | −664.2 | 1410.4 |
| | 6−state | 41 | −699.6 | 1481.3 | 56 | −657.0 | 1426.0 |
| IHSMM | 2−state | 11 | −755.4 | 1532.8 | 14 | −737.9 | 1503.8 |
| | 3−state | 20 | −729.2 | 1498.5 | 26 | −688.5 | 1429.0 |
| | 4−state | 31 | −708.5 | 1479.1 | 40 | −659.8 | 1399.7 |
| | 5−state | 44 | −690.7 | 1469.3 | 56 | −647.5 | 1407.0 |
| | 6−state | 59 | −687.8 | 1493.7 | 74 | −645.0 | 1437.9 |

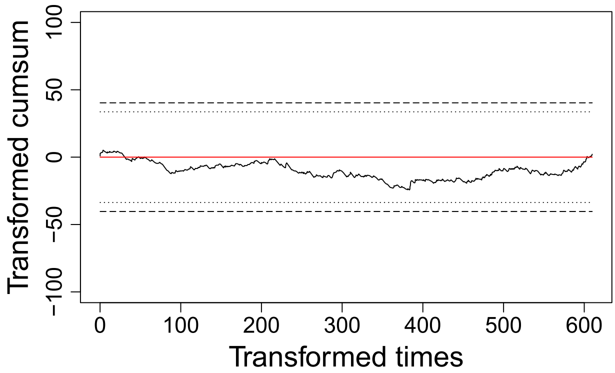



**Fig. 3** The deviated cumulative number of events in the residual process from the stationary process versus the transformed times for the 4−state IHMM with TSCPG renewal process for the global volcanic eruption record from the LaMEVE database. The solid red line is the theoretical curve under the null hypothesis of stationary process. The dotted and dashed lines represent the two-sided 95% and 99% confidence limits of the KS statistic, respectively

However, the Viterbi path of this model (Fig. S.11 in the Supplementary Material) suggests that while the record may be complete for some periods before 420 YBP (AD 1530), it is entirely in state of incompleteness during the most recent 500 years, i.e., from 420 YBP to present, with the most likely state sequence (Viterbi path) being in only States 3, 4 or 5 since 420 YBP. The empirical evidence showed that the most recent part of the global eruption record is almost complete for large eruptions in the most recent 500 years (Furlan 2010; Deligne et al. 2010). The information criterion does not guarantee that the chosen model describes the main features of the data (Pohle et al. 2017). We therefore consider this model inappropriate.

Further, for the 4−state IHMM with TSCPG renewal process that has the second smallest AIC, the cumulative number of events versus transformed times $\tau_k$ in Fig. 3 with 95% and 99% confidence intervals of the KS statistic shows no significant deviation from the theoretical curve. We can see that $U_k$ shows no particular pattern of association with $U_{k+1}$ (Fig. 4). This is also confirmed by a t-test for the correlation coefficient with a $p-$value greater than 0.05.

### 6.3 Estimates of the 4−state IHMM with TSCPG renewal Process

From the data analysis, we arrive at the conclusion of choosing the 4−state IHMM with TSCPG renewal process as one of the best candidates for the unknown true model of the given global volcanic eruption record. State 1 defines no missing event and States 2, 3 and 4 represent time-dependent Poisson distributed number of missing events between each pair of consecutively observed events in the record. The MLEs of the parameters of this model via unconstrained optimization are listed in Table 5, involving the estimates of the parameters in the time-varying transition probabilities defined by (9) and the estimates of parameters for the time-dependent Poisson mean number of missing events defined by (6).

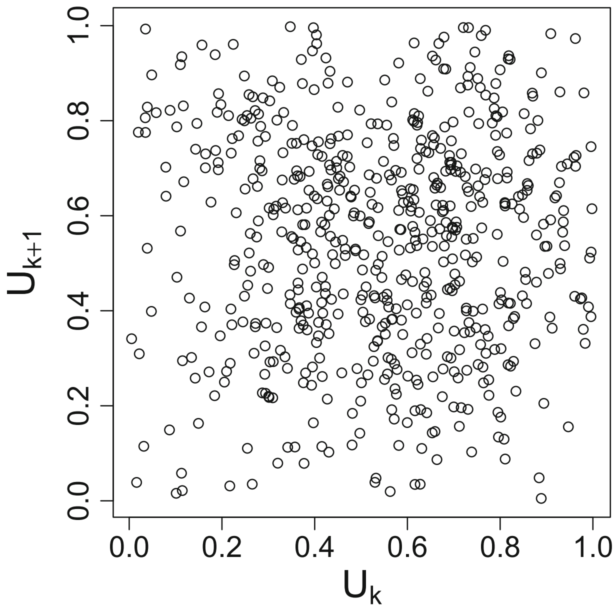



**Fig. 4** Scatter plot of $U_{k+1}$ versus $U_k$ from the 4−state IHMM with TSCPG model fitted to the global volcanic eruption record from the LaMEVE database

In Table 5, the values in brackets are the corresponding bootstrap standard errors of estimates of the model, reflecting the variability and uncertainty associated with estimating parameters from bootstrap samples.

Using the fitted model in Table 5, the estimated time-dependent expected number of missing events in the states of incompleteness (States 2, 3 and 4) and the estimated time-varying transition probabilities are shown in Figs. 5 and 7, respectively.

From Fig. 5, we see that the expected number of missing events per visit in State 2 remains at 16.8 before 3000 YBP and decreases between 3000 YBP and 1000 YBP until it approaches zero missing events during the last 1000 years. State 3 has the highest expected number of missing events at 44 prior to 7000 YBP. In this state, the expected number of missing events decreases during 7000 YBP to 3000 YBP at a relatively slower rate than that in State 2. In the last 3000 YBP, the expected number of missing events in State 3 stays at 6. State 4 has the least expected number of missing events when compared to States 2 and 3 during the last 10000 YBP, decreasing slowly and approaching zero in recent times.

It can be useful, recalling that our model does not use eruption magnitude, to examine the missing states in terms of the observed magnitudes. We plot the empirical CDFs of eruption magnitudes observed in each state according to the Viterbi path of the 4−state IHMM with TSCPG renewal process in Fig. 6. Also, the proportion of eruption magnitudes in each of two categories 4 − 5.5 (lower magnitude) and 5.5+ (higher magnitude) in each of four states is listed in Table 6. State 1 is assumed to be complete, and hence the magnitude distribution to be that of the underlying process. State 3 has the largest expected number of missing events, covering the lowest

**Table 5** Estimates of the 4−state IHMM with TSCPG renewal process

| States | $j$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Initial distribution | $\hat{\pi}_j$ | 0.0000 | 0.0000 | 1.0000 | 1.0000 |
| Time-varying transition probability parameters | $\hat{r}_{ij}$ | | | | |
| $i$ | 1 | – | 1.8085 *(0.0217)* | 0.3431 *(0.0034)* | 0.1886 *(0.0011)* |
| | 2 | – | 0.0001 *(0.0002)* | 0.0009 *(0.0000)* | 0.0029 *(0.0002)* |
| | 3 | – | 0.0103 *(0.0035)* | 0.0386 *(0.0001)* | 0.0000 *(0.0000)* |
| | 4 | – | 0.0021 *(0.0001)* | 0.0000 *(0.0000)* | 0.0041 *(0.0001)* |
| | $\hat{w}_{ij}$ | | | | |
| $i$ | 1 | – | 570.4872 *(0.0064)* | 230.7568 *(0.4425)* | 179.6826 *(1.1118)* |
| | 2 | – | 12.5208 *(0.1418)* | -7.2895 *(0.0128)* | 13.4648 *(0.1428)* |
| | 3 | – | 3.9239 *(0.0999)* | -8.0738 *(0.0015)* | 1.0563 *(0.0204)* |
| | 4 | – | 2.2551 *(0.0511)* | 1.3429 *(0.0470)* | 3.2938 *(0.0511)* |
| Gamma and TSCPG distributions | $\alpha$ | | | 1 | |
| | $\hat{\beta}$ | | | 0.2590 *(0.0006)* | |
| | $\hat{\lambda}_j(t)$ | | | | |
| | $\hat{M}_j$ | – | 16.7803 *(0.0526)* | 38.0721 *(0.2200)* | 4.1344 *(0.0276)* |
| | $\hat{P}_j$ | – | 26.1777 *(0.4621)* | 10.4260 *(0.5040)* | 4.5845 *(0.6994)* |
| | $\hat{R}_j$ | – | 0.0337 *(0.0006)* | 0.0231 *(0.0011)* | 0.0062 *(0.0010)* |
| | $\hat{K}_j$ | – | 0.0000 *(0.0000)* | 6.2585 *(0.0374)* | 0.0001 *(0.0003)* |

The values in brackets are bootstrap standard errors for the estimates

percentage of lower and higher magnitude events (9.3% and 2.45%). State 4 has the smallest expected number of missing events over the entire period, with the coverage of 22.06% of lower magnitude events, similar to State 1. State 4 has significantly larger magnitudes than all other states. Lower magnitude events are more prevalent in State 2 (29.90 %), and hence it represents a more hybrid situation where there is magnitude dependent missingness, but less severely. Where these States 3 and 4 differ is in the TSCPG distribution (Table 5), with State 4 having far fewer missed events per visit (Fig. 5).

We observe the different behaviour of transition probabilities over time in Fig. 7. The first row of time-varying transition probability matrix shows an interesting pattern of transitions from State 1 to different states over the entire period. Before 8000 YBP, the preferred transitions is to State 2; then the record is more likely to transit from State 1 to State 3 between 8000 YBP and 6000 YBP. A visual examination of Fig. 2

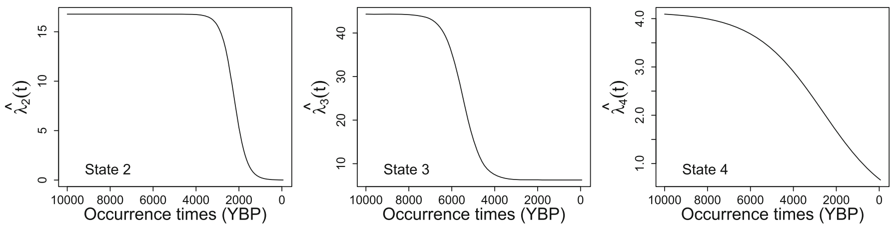



**Fig. 5** Estimated time-dependent expected number of missing events in States 2, 3 and 4 of the 4−state IHMM with TSCPG renewal process

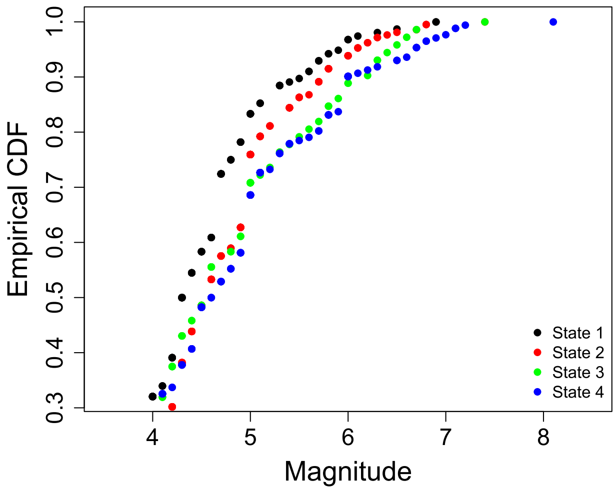



**Fig. 6** Empirical CDF of magnitudes falling in each state of the 4−state IHMM with TSCPG renewal process

**Table 6** Proportion of events according to eruption magnitude falling in each of four states

| State | Magnitude | Percentage |
|---|---|---|
| State 1 | 4 − 5.5 | 22.88 |
| | 5.5+ | 2.61 |
| State 2 | 4 − 5.5 | 29.90 |
| | 5.5+ | 4.74 |
| State 3 | 4 − 5.5 | 9.31 |
| | 5.5+ | 2.45 |
| State 4 | 4 − 5.5 | 22.06 |
| | 5.5+ | 6.05 |

suggests that the period from 8000 YBP to 6000 YBP has a high number of missing events with magnitude $< 6$. For that period, the record prefers State 3 as State 3 has the highest expected number of missing events before 6000 YBP (Fig. 9). After 6000 YBP, State 1 transits to State 4 until 700 YBP with highest probability, indicating that with greater likelihood there are fewer missing events. This is also evident from Table 6 that States 1 and 4 has about the same percentage of lower magnitude events. Finally, there is an abrupt decrease in the transition probability from State 1 to State 4 towards zero in the most recent past, corresponding to the rapid decrease in estimated underreporting during the last 400 (Furlan 2010) to 800 (Deligne et al. 2010) years.

The increasing self-transition probability in State 2 in Fig. 7 is consistent with an increasing level of completeness as the number of missing events in State 2 in recent times is close to zero (Fig. 5). The transition probability from State 2 to State 4 exhibits a complementary decreasing tendency over time. The decreasing and increasing trends in the transition probabilities from State 3 to States 2 and 4, respectively, clearly

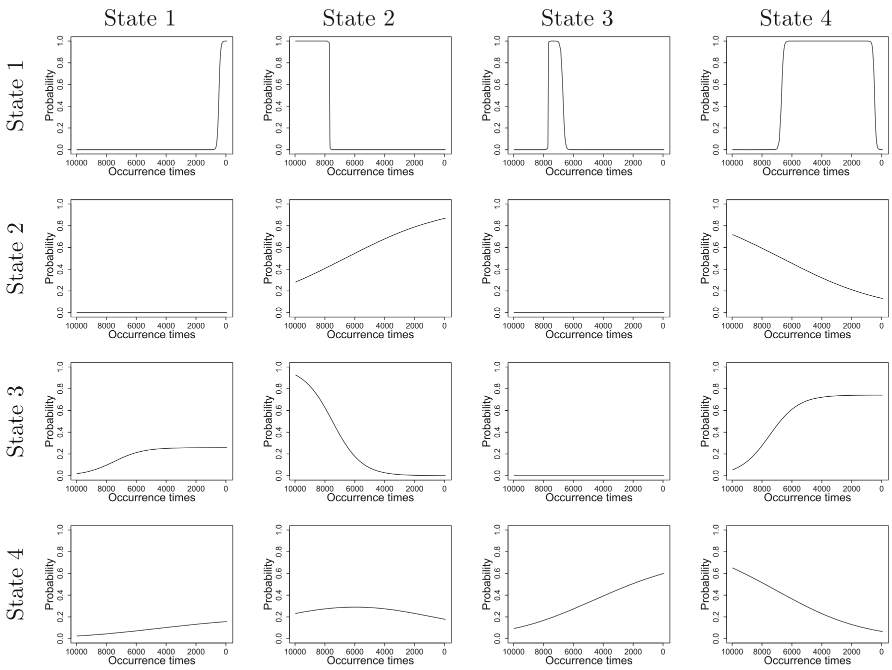



**Fig. 7** Estimated time-varying transition probabilities from the selected 4−state IHMM with TSCPG renewal process

represent that the process is switching from a high number of missing events state to a small number of missing events state over time, demonstrating the record being less incomplete in its recent past. Finally, the self-transitions in State 4 may signify a decreasing tendency in the number of missing eruptions of smaller magnitudes.

In the IHMM, the inherent geometric state duration probability distribution in state $j$, $p_j(d, t)$, with self-transition probability parameter $a_{jj}(t)$ is of the form

$$p_j(d,\ t) = \{a_{jj}(t)\}^{d-1}\{1 - a_{jj}(t)\}, \tag{15}$$

which represents the probability of $d$ consecutive observations in state $j$. Therefore, the expected sojourn time in state $j$ is $1/(1 - a_{jj}(t))$. We plot the implicit geometric mean sojourn times in each state of the 4−state IHMM in Fig. 8. The sojourn time in State 1 increases abruptly in the last few hundred years. State 3 has duration 1 over the entire period as there are no self-transitions in this state. State 2 has a longer sojourn time than State 4, though the expected number of missing events in State 2 is trending to zero as we move towards the present.

The Viterbi path for the 4−state IHMM with TSCPG in Fig. 9 also sketches the overall picture of staying in each state. The visualization of the Viterbi path for the model indicates the completeness of the global volcanic record during the last few hundred years. The 4−state IHMM estimates that the record is complete after 211 YBP (AD 1739).

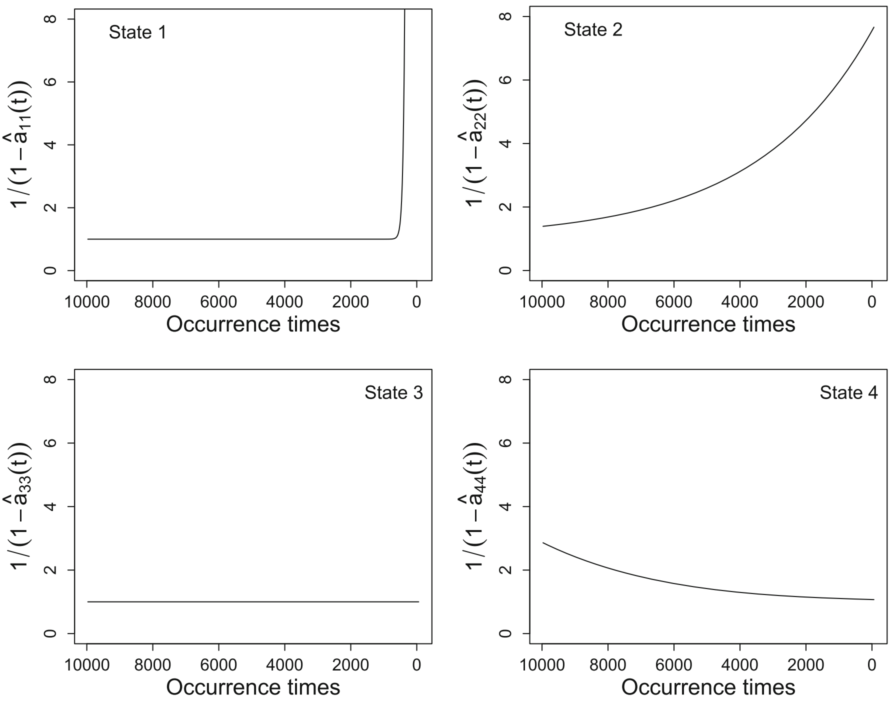



**Fig. 8** Estimated time-dependent expected implicit state durations in each state of the 4−state IHMM with TSCPG renewal process

Estimating the hazard rate is one of the main objectives of the proposed models. The hazard rate $\lambda(h_t)$ defined in (1) can be estimated using the estimates of the gamma distribution in State 1 from the selected 4−state IHMM with TSCPG renewal process, which is shown in Table 7. To make a comparison, we also calculate the hazard rate by merely fitting a gamma distribution with $\alpha = 1$ to the observed interevet times without considering any missing data. The interevent times in State 1 follows an exponential distribution for the global volcanic eruption record, so the estimated hazard rate is constant over time. The 4−state IHMM model produces higher future hazard estimates of approximately 4 events per decade which are more than six times of that estimated by a gamma renewal process fitted on interevent times since 10000 YBP. Since the 4−state IHMM estimates the recurrence time intervals to be in State 1 since 211 YBP, the gamma renewal process fitted on this part of data gives the hazard estimate slightly greater than that obtained by the 4−state IHMM, which includes a small chance of transition to a lower rate regime.

One of the objectives of the proposed models is to estimate the time-dependent proportion of completeness of a point process record. The time-dependent proportion

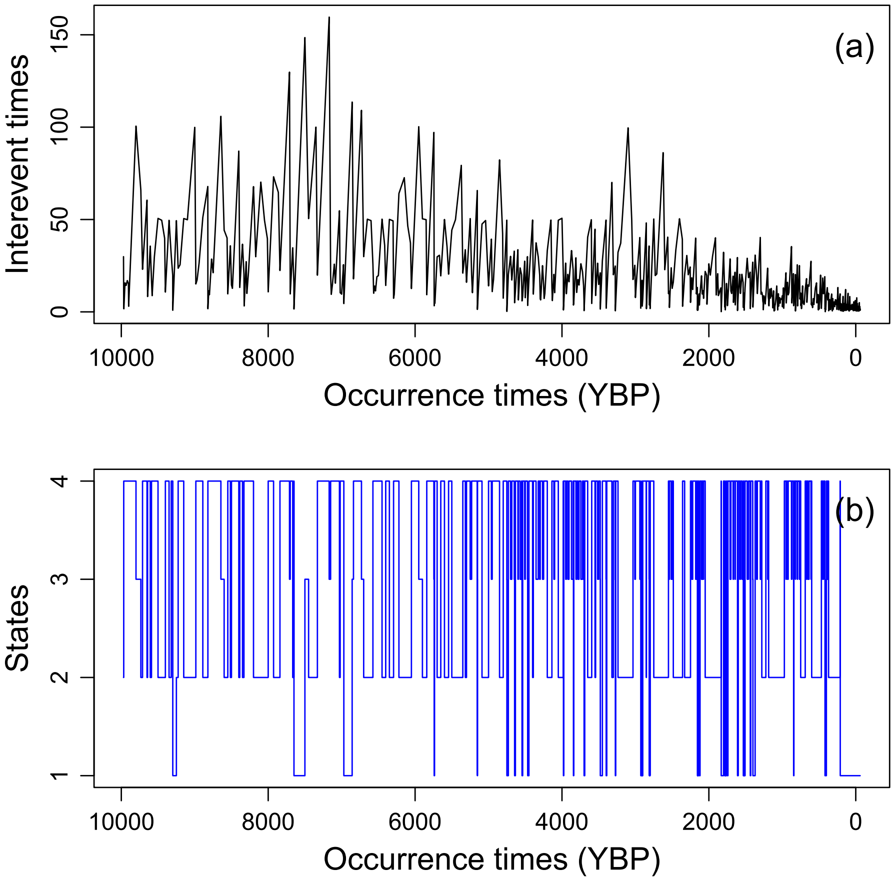



**Fig. 9** (**a**) The observed interevent times, and (**b**) The Viterbi path for the 4−state IHMM with TSCPG renewal process

**Table 7** Hazard rate estimate

| Model | $\alpha$ | $\hat{\beta}$ | $\hat{\lambda}(h_t)$ per decade |
|---|---|---|---|
| 4−State IHMM with TSCPG renewal process fitted on interevent times since 10000 YBP | 1 | 0.26 | 3.85 |
| Gamma (exponential) distribution fitted on interevent times since 10000 YBP | 1 | 1.64 | 0.61 |
| Gamma (exponential) distribution fitted on interevent times since 211 YBP | 1 | 0.23 | 4.32 |

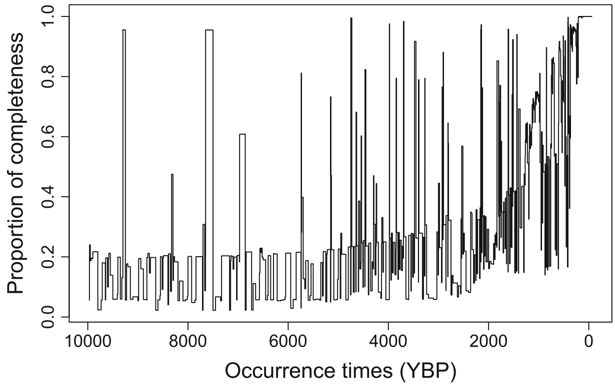



**Fig. 10** Estimated time-dependent proportion of completeness from the selected 4−state IHMM with TSCPG renewal process

of completeness can be estimated by

$$\hat{p}(t) = \hat{\gamma}_1(t) + \sum_{j=2}^{m} \hat{\gamma}_t(j)/(\hat{\lambda}_j(t) + 1), \tag{16}$$

where, $\hat{\gamma}_j(t)$ is the estimated probability of being in state $j$ at time $t$ and $\hat{\lambda}_j(t)$ are the estimated time-dependent expected number of missing events in state $j$. The estimated proportion of completeness of the record from the 4−state IHMM with TSCPG renewal process is shown in Fig. 10, which demonstrates that the proportion of completeness is mostly less than or equal to 20% before 5000 YBP. Deligne et al. (2010) analyzed a global volcanic eruption record in the Holocene period and found that before 2000 years ago there is 20% chance of recording an eruption event with magnitude $\geq$ 6. Brown et al. (2014) concluded that the LaMEVE database includes less than 10% of eruptions with magnitude $4 - 4.9$, less than 20% of eruptions with magnitude $5 - 5.9$ and less than 25% of eruptions with magnitude $6 - 6.9$ before 5000 YBP. The estimated proportions of completeness prior to 5000 YBP in Fig. 10 reflects the above findings. At some time points, the proportion of completeness from the IHMM is very high before 5000 YBP, which may indicate a high proportion of large eruption events since they are less likely to be missed. This may be better explained by a model that includes the magnitude effect, which is an avenue for future research. The high proportion of completeness in very early times may also be an effect of non-random sampling in recording events using different geological techniques in different locations and environments.

In general, the estimated time-dependent completeness of the global volcanic eruption record in Holocene period indicates that the record is incomplete from the start of the Holocene until 211 YBP and above. After 5000 YBP, there are many fluctuations in the proportion of completeness which may represent regional biases. The completeness of the record increases markedly during the last 2000 years. The considerably increasing proportion of completeness from 1250 YBP to 750 YBP given by the

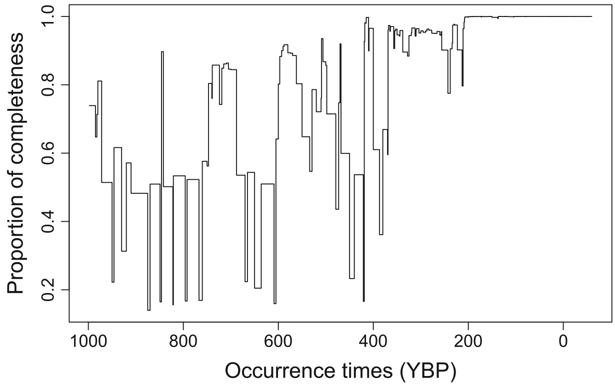



**Fig. 11** Estimated time-dependent proportion of completeness from the selected 4−state IHMM with TSCPG renewal process during the last 1000 YBP

4−state IHMM with TSCPG renewal process aligns with the dominant contribution of Japanese volcanic eruption events to the global database during this period (Kiyosugi et al. 2015). Using a nonparametric approach, Rougier et al. (2016) estimated the global recording rate of large eruptions of magnitude $\geq 4$, concluding that the proportion of recorded events is below 20% before 850 YBP and below 50% before 350 YBP. The proportion of completeness in Fig. 11 reflects these proportions at the same time points. The incompleteness largely disappears during the last few hundred years (Fig. 11), approaching 100% in line with the empirical findings in previous studies (Furlan 2010; Coles et al. 2006; Deligne et al. 2010; Rougier et al. 2016).

## 7 Conclusion

We have proposed the use of SCPG and TSCPG renewal processes in combination with different types of HMMs to investigate the time-dependent incompleteness of point processes. The simulation experiment confirms the consistency of parameter estimators. The proposed models were applied to study the time-dependent incompleteness behavior of a global volcanic eruption record. The AIC was used to choose the best model, and then residual analysis (Berman 1983; Ogata 1988; Wang et al. 2012) was performed to check how well this model captures the main features of the data.

The application of the proposed models to a global volcanic eruption record revealed the necessity of utilizing the inhomogeneous models when the data exhibits time-dependent missingness. We applied the proposed models with up to 6 hidden states to a catalogue of global volcanic eruptions in the Holocene epoch extracted from the LaMEVE database (Crosweller et al. 2012). The 4−state IHMM with TSCPG renewal process was selected as the best candidate for the unknown true model for the record. The model produced residual processes consistent with a stationary Poisson process with unit rate (Figs. 3 and 4). The hazard rate estimates from the 4−state IHMM with TSCPG renewal process came out to be about 4 events per decade.

Further improvement of the model may be achieved by using a different formulation of time-dependent transition probabilities based on covariates such as magnitudes in the IHMMs. Finally, we suggest future work focuses on including eruption magnitude data to construct a more transparent structure of time-inhomogeneous incompleteness of the record.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11634-025-00632-x.

**Acknowledgements** We acknowledge the New Zealand eScience Infrastructure (NeSI) for granting access to the Mahuika high performance computing cluster. We especially acknowledge the contribution of the late Prof. Richard B. Paris in facilitating the Mathematica code for calculating the coefficients used in the asymptotic expansion of Wright's generalized function.

**Author contributions** All authors contributed to the study conception and design. Material preparation and data analysis were performed by Amina Shahzadi and supervised by Ting Wang, Matthew Parry and Mark Bebbington. The first draft of the manuscript was written by Amina Shahzadi and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

**Funding** Open access funding provided by The University of Otago. Amina Shahzadi is a recipient of a Doctoral Scholarship from the University of Otago, New Zealand. This research is also sponsored by The Royal Society of New Zealand Te Aprangi Marsden Fund (UOO1419). Ting Wang and Mark Bebbington are supported by the Resilience to Nature's Challenges Volcano Programme (GNS-RNC047).

**Data availability** Data will be provided.

**Code availability** The R code will be provided if required.

## Declarations

**Conflict of interest** We do not have any Conflict of interest in the publication of this work.

**Ethics approval** Not applicable.

**Informed consent** For this type of study formal consent is not required.

**Consent to participate** Not applicable.

**Consent for publication** All authors give explicit consent for the submission of this study.



## References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723. https://doi.org/10.1109/TAC.1974.1100705

Allen J, Wang T (2024) Hidden Markov models for low-frequency earthquake recurrence. Amer Statist 78(1):100–110. https://doi.org/10.1080/00031305.2023.2282631
Barbu V, Limnios N (2008) Semi-Markov Chains and Hidden Semi-Markov Models Toward Applications: Their Use in Reliability and DNA Analysis, 1st edn. Springer, New York . https://doi.org/10.1007/978-0-387-73173-5
Baum LE, Pretrie T, Soules G, Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chain. Ann Math Stat 41(1):164–177
Bebbington MS (2013) Models for temporal volcanic hazard. Stat Volcanol 1:1–24. https://doi.org/10.5038/2163-338X.1.1
Bebbington MS, Lai CD (1996) Statistical analysis of New Zealand volcanic occurrence data. J Volcanol Geoth Res 74(1):101–110. https://doi.org/10.1016/S0377-0273(96)00050-9
Berman M (1983) Comment on Likelihood analysis of point processes and its applications to seismological data by Ogata. Bull Int Stat Inst 50:412–418
Brown SK, Crosweller HS, Stephen R, Sparks J, Cottrell E, Deligne NI, Guerrero NO, Hobbs L, Kiyosugi K, Loughlin SC, Siebert L, Takarada S (2014) Characterisation of the Quaternary eruption record: analysis of the large magnitude explosive volcanic eruptions (LaMEVE) database. J Appl Volcan 3(1):5. https://doi.org/10.1186/2191-5040-3-5
Buckby J, Wang T, Fletcher D, Zhuang J, Takeo A, Obara K (2023) Finding the number of latent states in hidden Markov models using information criteria. Environ Ecol Stat. https://doi.org/10.1007/s10651-023-00584-5
Bulla J (2006) Application of hidden Markov models and hidden semi-Markov models to financial time series. Georg-August-Universitat Gottingen, PhD-Thesis
Bulla J, Berzel A (2008) Computational issues in parameter estimation for stationary hidden Markov models. Computation Stat 23:1–18
Campillo F, Gland FL (1989) MLE for partially observed diffusion: Direct maximization vs. the EM algorithm. Stoch Proc Appl 33:245–274
Coles Stuart G., Sparks R. Stephen J. (2014) Extreme value methods for modelling historical series of large volcanic magnitudes. In: Mader Heidy M., Coles Stuart G., Connor Charles B., Connor Laura J. (eds) Statistics in Volcanology. The Geological Society of London on behalf of The International Association of Volcanology and Chemistry of the Earth's Interior, pp 47–56. https://doi.org/10.1144/IAVCEI001.5
Crosweller HS, Arora B, Brown SK, Cottrell E, Deligne NE, Guerrero NO, Hobbs L, Kiyosugi K, Loughlin SC, Lowndes J, Nayembil M, Siebert L, Sparks R, Takarada S, Venzke E (2012) Global database on large magnitude explosive volcanic eruptions LaMEVE. J Appl Volcan 1:4. https://doi.org/10.1186/2191-5040-1-4
Cruz-Reyna S (1993) Random patterns of occurrence of explosive eruptions at Colima volcano Mexico. J Volcanol Geoth Res 55(1):51–68. https://doi.org/10.1016/0377-0273(93)90089-A
Cruz-Reyna S (1991) Poisson-distributed patterns of explosive eruptive activity. B Volcanol 54(1):57–67. https://doi.org/10.1016/0377-0273(93)90089-A
Daley DJ, Vere-Jones D (2003) Introduction to the Theory of Point Processes. Springer, New York. https://doi.org/10.1007/b97277
Decreusefond L, Moyal P (2013) Stochastic modeling and analysis of telecom networks. Wiley, USA. https://doi.org/10.1002/9781118562628
Deligne NI, Coles SG, Sparks R (2010) Recurrence rates of large explosive volcanic eruptions. J Geophys Res 115:06203. https://doi.org/10.1029/2009JB006554
Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B Stat Methodol 39(1):1–38
Diebold FX, Lee J-H, Weinbach GC (1994) Regime switching with time-varying transition probabilities. In: Diebold, F.X., Rudebusch, G.D. (eds.) Business Cycles: Durations, Dynamics and Forecasting, pp. 283–302. Princeton Univeristy Press, London . Chap. 10
Dunn PK, Smyth GK (2001) Tweedie family densities: methods of evaluation. In: Proceedings of the 16th international workshop on statistical modelling, Odense, Denmark, pp. 2–6
Dunn PK, Smyth GK (2005) Series evaluation of Tweedie exponential dispersion model densities. Stat Comput 15(4):267–280. https://doi.org/10.1007/s11222-005-4070-y
Ferguson JD (1980) Variable duration models for speech. Proceedings Symposium on the Application of Hidden Markov Models to Text and Speech, Princeton, NJ, 143–179

Filardo AJ (1994) Business-cycle phases and their transitional dynamics. J Bus Econ Stat 12(3):299–308. https://doi.org/10.1080/07350015.1994.10524545
Filardo AJ, Gordon SF (1998) Business cycle durations. J Econometrics 85(1):99–123. https://doi.org/10.1016/S0304-4076(97)00096-1
Furlan C (2010) Extreme value methods for modelling historical series of large volcanic magnitudes. Stat Model 10(2):113–132. https://doi.org/10.1177/1471082X0801000201
Guédon Y, Cocozza-Thivent C (1989) Use of the Derin's algorithm in hidden semi-Markov models for automatic speech recognition. Proceedings - IEEE international conference on acoustics, speech and signal processing, 282–285 https://doi.org/10.1109/ICASSP.1989.266420
Guttorp P, Thompson ML (1991) Estimating second-order parameters of volcanicity from historical data. J Am Stat Assoc 86(415):578–583. https://doi.org/10.1080/01621459.1991.10475082
Hughes JP, Guttorp P (1994) A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. Water Resour Res 30(5):1535–1546. https://doi.org/10.1029/93WR02983
Hughes JP, Guttorp P (1999) A non-homogeneous hidden Markov model for precipitation occurrence. Appl Stat 48(1):15–30. https://doi.org/10.1111/1467-9876.00136
Jordan M, Xu L (1996) Convergence results for the EM approach to mixtures of experts architectures. Neural Netw 8:1409–1431
Kaczmarska J, Isham V, Onof C (2014) Point process models for fine-resolution rainfall. Hydrol Sci J 59(11):1972–1991. https://doi.org/10.1080/02626667.2014.925558
Kiyosugi K, Connor C, Sparks R, Crosweller HS, Brown SK, Siebert L, Wang T, 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. J Appl Volcan 4(1):17. https://doi.org/10.1186/s13617-015-0035-9
McLachlan GJ, Krishnan K (2008) The EM Algorithms and Extensions, 2nd edn. John Willey & Sons Inc, The Wiley Network
Mohler G (2013) Modelling and estimation of multi-source clustering in crime and security data. Ann Appl Stat 7(3):1525–1539. https://doi.org/10.48550/arXiv.1311.7279
Mohler GO, Short MB, Brantingham PJ, Schoenberg FP, Tita GE (2011) Self-exciting point process modeling of crime. J Amer Statist Assoc 106(493):100–108. https://doi.org/10.1198/jasa.2011.ap09546
Ogata Y (1988) Statistical models for earthquake occurrence and residual analysis for point processes. J Am Stat Assoc 83:9–27. https://doi.org/10.1080/01621459.1988.10478560
Ogata Y (1998) Space-time point process models for earthquake occurrences. Ann Inst Math Sci 50(2):379–402. https://doi.org/10.1023/A:1003403601725
Ogata Y, Katsura K (1993) Analysis of temporal and spatial heterogeneity of magnitude frequency distribution inferred from earthquake catalogues. Geophys J Int 113(3):727–738. https://doi.org/10.1111/j.1365-246X.1993.tb04663.x
Otranto E (2008) A time varying hidden Markov model with latent information. Stat Model 8(4):347–366. https://doi.org/10.1177/1471082X0800800402
Paris RB (2010) Exponentially small expansions in the asymptotics of the Wright function. J Comput Appl Math 234(2):488–504. https://doi.org/10.1016/j.cam.2009.12.040
Paris RB (2017) The asymptotics of the generalized Bessel function. Mathematica Aeterna 7(4):381–406. https://doi.org/10.48550/arXiv.1711.03006
Paris RB, Vinogradov V (2016) Asymptotic and structural properties of special cases of the Wright function arising in probability theory. Lith Math J 56(3):377–409. https://doi.org/10.1007/s10986-016-9324-1
Pohle J, Langrock R, van Beest FM, Schmidt NM (2017) Selecting the number of states in hidden Markov models: pragmatic solutions, illustrated using animal movement. JABES 22:270–293. https://doi.org/10.1007/s13253-017-0283-8
Pyle DM (2015) Sizes of Volcanic Eruptions. The Encyclopedia of Volcanoes. Elsevier, pp 257–264. https://doi.org/10.1016/B978-0-12-385938-9.00013-4
Quesada JA, Melchor I, Nolasco A (2017) Point process methods in epidemiology: application to the analysis of human immunodeficiency virus/acquired immunodeficiency syndrome mortality in urban areas. Geospat Health 12(1):506. https://doi.org/10.4081/gh.2017.506
R Core Team (2017) R: The R Project for Statistical Computing. R Foundation for Statistical Computing, Auckland, New Zealand. R Foundation for Statistical Computing. http://www.R-project.org/
Rougier J, Sparks SR, Cashman KV (2016) Global recording rates for large eruptions. J Appl Volcan 5(1):11. https://doi.org/10.1186/s13617-016-0051-4

Shahzadi A, Wang T, Parry M, Bebbington M (2023) Inhomogeneous hidden semi-Markov models for incompletely observed point processes. Ann Inst Statist Math. https://doi.org/10.1007/s10463-022-00843-5

Siebert L, Simkin T, Kimberly P, Institution S (2010) Volcanoes of the World, 3rd edn. University of California Press, London

Simkin T (1993) Terrestrial volcanism in space and time. Annu Rev Earth Pl Sc 21(1):427–452. https://doi.org/10.1146/annurev.ea.21.050193.002235

Smyth GK, Jorgensen B (2002) Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling. ASTIN Bulletin 32(1):143–157. https://doi.org/10.2143/AST.32.1.1020

Tucker JD, Shand L, Lewis JR (2019) Handling missing data in self-exciting point process models. Spat Stat 29:160–176. https://doi.org/10.1016/j.spasta.2018.12.004

Tweedie MCK (1984) An index which distinguishes between some important exponential families. In: Statistics: Applications and New Directions. Proceedings of the Indian statistical institute golden jubilee international conference, pp. 579–604

Veneziano D, Van Dyck J (1987) Statistical analysis of earthquake catalogs for seismic Hazard. In: Lin YK, Minai R (eds) Stochastic approaches in earthquake engineering. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 385–427. https://doi.org/10.1007/978-3-642-83252-9_20

Wang T (2010) Statistical Models for Earthquakes Incorporating Ancillary Data. PhD Thesis, Massey University, Palmerston North, New Zealand

Wang T, Bebbington MS (2012) Estimating the likelihood of an eruption from a volcano with missing onsets in its records. J Volcanol Geoth Res 243:14–23. https://doi.org/10.1016/j.jvolgeores.2012.06.032

Wang T, Bebbington MS (2013) Identifying anomalous signals in GPS data using HMMs: an increased likelihood of earthquakes. Comput Stat Data Anal 58:27–44. https://doi.org/10.1016/j.csda.2011.09.019

Wang T, Bebbington MS, Harte D (2012) Markov-modulated Hawkes process with stepwise decay. Ann Inst Stat Math 64:521–544. https://doi.org/10.1007/s10463-010-0320-7

Wang T, Schofield M, Bebbington MS, Kiyosugim K (2020) Bayesian modelling of marked point processes with incomplete records: volcanic eruptions. J R Stat Soc Ser C Appl Stat 69:109–130. https://doi.org/10.1111/rssc.12380

Wu C (1983) On the convergence property of the EM algorithm. Ann Statist 11(1):95–103

Zipkin JR, Schoenberg FP, Coronges K, Bertozzi AL (2016) Point-process models of social network interactions: Parameter estimation and missing data recovery. Eur J Appl Math 27(3):502–529. https://doi.org/10.1017/S0956792515000492

Zucchini W, MacDonald IL (2009) Hidden Markov Models for Time Series: An Introduction Using R. Chapman and Hall/CRC, New York. https://doi.org/10.1201/9781420010893