

RESEARCH

Open Access



# Global recording rates for large eruptions

Jonathan Rougier<sup>1\*</sup>, Stephen R. Sparks<sup>2</sup> and Katharine V. Cashman<sup>2</sup>

## Abstract

A non-parametric statistical approach is used to assess the global recording rate for large (M4+) stratovolcano eruptions in a modern database, LaMEVE (v3.1). This approach imposes minimal structure on the shape of the recording rate through time. We find that recording rates have declined rapidly, going backwards in time. Prior to the year 1600 they are below 50 %, and prior to 1100 they are below 20 %. Even in the recent past, e.g. the 1800s, they are likely to be less than 100 %. The assessment for very large (M5+) eruptions is more uncertain, due to the scarcity of events.

**Keywords:** Under-recording, Non-parametric, Confidence set

## Background

Under-recording is an unavoidable issue in empirical studies of large volcanic eruptions, as has been discussed by, e.g., Simkin (1993), Siebert et al. (2010, pp. 31–34) and Brown et al. (2014). As Simkin (1993, p. 436) puts it, the rise in the number of recorded eruptions follows from “more observers, in a wider geographical distribution, with better communication, and broader publication.” Statisticians refer to this type of absence as ‘missing not at random’, and it must be accounted for when making inferences from observations, in order to remove, as far as possible, systematic biases. But this under-recording poses a major challenge, because we have few quantitative beliefs about the way in which the timing, location, and magnitude of an eruption combine together to affect its missingness from the modern record.

For an initial analysis, we can make one drastic simplification at the outset, and consider global volcanism above a specified large magnitude threshold. This reduces three dimensions of missingness to just one—timing. It also simplifies the statistical modelling: first, because we avoid complicated issues regarding the modelling and accuracy of magnitude measurements; second, because the process of global volcanism is likely to be more stable than its component parts. These issues are discussed further below (Section “Our statistical non-parametric approach”).

This paper provides an empirical analysis of the global recording rate of large eruptions of stratovolcanoes; the definition of the global recording rate is discussed and clarified in Section “Definition of the recording rate”. ‘Large’ is taken to be at least 100 million tonnes of ejected matter, or M4 according to the magnitude scale of Mason et al. (2004), for which  $M := \log_{10} m - 7$ , where  $m$  is the mass erupted in kilograms.  $M$  thus provides a continuous measure of eruption magnitude based on estimates of erupted mass; more common is the use of the Volcanic Explosivity Index (VEI), which is discrete and assigned on the basis of either erupted volume or eruption intensity (average mass eruption rate).

Our key innovation is to use a non-parametric approach (Section “Our statistical non-parametric approach”), and to present our results visually in terms of a 95 % confidence set (Section “Results and discussion”). We are able to quantify recording rates through time in terms of upper bounds of 95 % confidence intervals, and show, for example, that this upper bound is below 50 % prior to the year 1600.

## Methods

### Definition of the recording rate

The notion of a ‘global recording rate’ is imprecise. We provide a definition which highlights the difference between the global recording rate and the mean recording rate across volcanoes.

Imagine a process, extensive in space and time, which indicates with a 1 or a 0 whether there is an information link from a large eruption at a given location and

\*Correspondence: j.c.rougier@bristol.ac.uk

<sup>1</sup>School of Mathematics, University of Bristol, University Walk, BS8 1TW Bristol, England

Full list of author information is available at the end of the article

date, through to today and an appearance in the database. Index the active stratovolcanoes with  $i = 1, \dots, n$ , and write  $c_i(t)$  for this linkage function at the location of volcano  $i$  and time  $t$ . If we observed this function for a given  $i$  going backwards through time, we would see it at ‘on’ (i.e. 1) for today, but then at some point in the past it would start to flicker between ‘on’ and ‘off’, indicating that a link to today is contingent on there being a suitable person present at the eruption, or traces (e.g. a near-field tephra deposit) that have survived and been discovered and interpreted. As we go back further, the function spends longer and longer amounts of time ‘off’, until finally it is permanently ‘off’. This sequence, from currently ‘on’, through flickering, to ‘off’, varies from volcano to volcano. A global recording function reflects the amalgamation of such functions for the global collection of volcanoes.

With the linkage function specified, define the global recording rate at time  $t$  as

$$\pi(t) := \frac{c_1(t) \cdot \lambda_1 + \dots + c_n(t) \cdot \lambda_n}{\lambda} \tag{1}$$

where  $\lambda_i$  is the large-eruption rate of volcano  $i$ , and  $\lambda := \lambda_1 + \dots + \lambda_n$ , the global large-eruption rate; these rates are taken to be constant in time—see the next paragraph for the generalisation. The numerator of (1) is the database’s global large-eruption rate for time  $t$ , which can be written as  $\pi(t) \cdot \lambda$ , where  $\lambda$  is today’s global large-eruption rate. Note that  $\pi(t) \rightarrow 1$  as  $t$  tends to today, because each  $c_i(t) \rightarrow 1$ . In (1), we are not assuming that large eruptions follow a Poisson process for each volcano, although we are assuming that we can combine rates in the same way that we would combine rates were the volcanoes to be following independent Poisson processes. In Section “Our statistical non-parametric approach”, we will treat the global large-eruption process as a time-homogeneous Poisson process, but we will not need to assert that the individual processes are either Poisson or stationary. This issue has been discussed by De la Cruz-Reyna (1991).

In the case where the large-eruption rates fluctuate in time, this fluctuation can be absorbed into the  $c_i(\cdot)$ ’s. Let  $\lambda_i$  be today’s large-eruption rate for volcano  $i$ , and then write

$$c_i(t) \cdot \lambda_i(t) = c_i(t) \cdot \lambda_i(t) \cdot \frac{\lambda_i}{\lambda_i} = c'_i(t) \cdot \lambda_i$$

where  $c'_i(t) := c_i(t) \cdot \lambda_i(t) / \lambda_i$ . We expect that  $c_i(t) \approx c'_i(t)$ , because this is exactly true when  $c_i(t) = 0$  or  $t$  tends to today, and approximately true elsewhere, if the fluctuations in  $\lambda_i(t)$  are not too large, in proportionate terms.

Now consider period  $j$  of width  $\Delta$  centered at time  $t_j$ . Starting with (1), take the time-average over the period to find the global large-eruption rate for period  $j$ :

$$\pi_j := \frac{\pi_{1j} \cdot \lambda_1 + \dots + \pi_{nj} \cdot \lambda_n}{\lambda} \tag{2a}$$

where

$$\pi_{ij} := \Delta^{-1} \int_{t_j - \Delta/2}^{t_j + \Delta/2} c_i(t) dt. \tag{2b}$$

The value  $\pi_{ij}$  is the recording rate of volcano  $i$  in period  $j$ . It is not a probability, but it can be interpreted as one, under additional conditions. If a large eruption of volcano  $i$  is randomly and uniformly distributed in period  $j$ , then  $\pi_{ij}$  is the probability of that eruption being recorded.

Consider (2a) in the special case where all of the large-eruption rates are approximately the same. In this case

$$\pi_j \approx \frac{\pi_{1j} + \dots + \pi_{nj}}{n}$$

and the global recording rate for period  $j$  is approximately the mean recording rate across all volcanoes. In general, though, (2a) shows that the global recording rate is biased away from the mean recording rate towards the recording rates of the most active volcanoes. So, for example, if recordings of a very active volcano start at the end of period  $j$ , then  $\pi_j$  to  $\pi_{j+1}$  will show a large increase, larger than the increase that would be expected for the addition of one extra volcano. Just monitoring  $\pi_j$  alone, it would not be possible to distinguish between, say, three newly-recorded volcanoes with modest eruption rates, and one newly-recorded volcano with a high eruption rate.

This distinction between the global recording rate and the mean recording rate is an important part of the definition and interpretation of the global recording rate, but whether it really matters is moot. Large eruptions are scarce. We do not expect to constrain our uncertainty about  $\pi_j$  to within the scale of the distinction, and from this point of view our uncertainty about  $\pi_j$  is more-or-less the same as our uncertainty about the mean recording rate, which is how we will interpret our results below.

### Review of previous approaches

Naïve estimators of the global recording rates are easy to derive. Consider a time period centred on time  $t_j$ , of width  $\Delta_j$ , which has  $x_j$  recorded large eruptions. Equating the expectation with the observation suggests that  $\Delta_j \cdot \pi_j \cdot \lambda \approx x_j$ , where  $\lambda$  is the global large-eruption rate, defined after (1), and taken to be effectively constant over the time-frame of the analysis;  $\pi_j$  is defined in (2a). Let the most recent period, period  $k$  say, have a recording rate of  $\pi_k = 1$ . Then  $\Delta_k \cdot \lambda \approx x_k$ , and substituting for  $\lambda$  gives

$$\pi_j \approx \frac{x_j}{\Delta_j} \bigg/ \frac{x_k}{\Delta_k} \quad j = 1, \dots, k. \tag{3}$$

Naïve though this is, it will turn out to be the Maximum Likelihood (ML) estimator in the model proposed in Section “Our statistical non-parametric approach”. It is,

of course, a woeful estimator, being very over-fitted: it has one parameter per observation, and no constraints.

There are several ways to proceed. One approach is to smooth the  $x_j$ 's, or equivalently the  $\pi_j$ 's. This approach was implemented in Guttorp and Thompson (1991), who discussed the critical issue of choosing a smoother and a bandwidth (their p. 579). There was no assessment of uncertainty in their estimated  $\pi_j$ 's, although this was not their main concern. We present a similar estimate below, at the start of Section "Results and discussion".

A second approach is to propose a specific functional form for  $\pi(\cdot)$ , which is controlled by a small number of unknown parameters; this imposes smoothness on  $\pi(\cdot)$  and reduces over-fitting. This approach was originally suggested in a more general context by Solow (1993), who proposed a monotonic four-parameter functional form based on the distribution function of the Beta distribution. The parameters of the proposed function are fitted to the  $x_j$ 's, which typically requires a statistical model for the number of recorded large-eruptions, if uncertainty is to be assessed. This is the parametric approach that has been used in the volcanological literature to date.

One functional form for  $\pi(\cdot)$ , used by Coles and Sparks (2006) and Deligne et al. (2010), is

$$\pi(t; t_0, b) = \left( \frac{t - t_0}{t_k - t_0} \right)^b \tag{4}$$

where  $t_0$  and  $b$  are the parameters,  $t_0 \leq t_1$ , and  $b > 0$ . This functional form encodes two basic beliefs: that the recording rate is currently 1, and that it decreases smoothly backwards in time. A very similar approach is used by Kiyosugi et al. (2015), who propose an exponential curve.

Another, used by Furlan (2010) and Mead and Magill (2014), is

$$\pi(t; \pi_0, t_s) = \begin{cases} \pi_0 & t \leq t_s \\ 1 & t > t_s \end{cases} \tag{5}$$

where  $\pi_0$  and  $t_s$  are the parameters, with  $0 \leq \pi_0 \leq 1$  and  $t_1 \leq t_s \leq t_k$ . Equation (5) defines a step function that moves abruptly from  $\pi_0$  to 1 at time  $t_s$ . This functional form encodes the same type of belief, except replacing 'smoothly backwards in time' with 'dramatically at some time previous to today'. Furlan (2010) uses a single global  $\pi_0$  and  $t_s$ , while Mead and Magill (2014) allow  $\pi_0$  and  $t_s$  to vary by region.

Mead and Magill's (2014) model is consistent with smooth functions such as (4) at the global level, as our analysis in Section "Definition of the recording rate" shows. In our terms, their model asserts that  $\pi_{ij} = \pi_0^{(i)}$  for  $t_j \leq t_s^{(i)}$ , and 1 thereafter, the superscript  $i$ 's indicating that these parameter values differ by volcano or region. If different regions have different  $t_s^{(i)}$  values, then

the total effect on the global large-eruption rate may look like a smooth curve, or a smooth curve with kinks, where a volcano or region with a high  $\lambda_i$  starts to be fully recorded. But a smooth curve would also arise in the more general case where, instead of jumping directly from  $\pi_0^{(i)}$  to 1, the  $\pi_{ij}$ 's increased more gradually.

The existence of these two different functions for  $\pi(\cdot)$  indicates a difficulty with this parametric approach: there is little guidance about the shape of the functional form. In fact there is more disagreement than we have presented here, because both functions also allowed for the effect of magnitude, but in different ways.

A third approach, adopted in this paper and new to volcanology, is to use a non-parametric representation of  $\pi(\cdot)$ , subject to some simple conditions. Because the conditions do not prescribe a tight relationship between the  $\pi_j$ 's, this option is sometimes known as a 'soft' constraint, compared to the second (parametric) approach, which would be a 'hard' constraint. A hard constraint is more effective if the parametric form of  $\pi(\cdot)$  is known, but a soft constraint is more widely applicable, and more suitable for dealing with observations affected by numerous complex factors related to recording of historical and geological events. As we show in Section "Our statistical non-parametric approach", a soft constraint can generalize both of the models given above. And as shown in Section "Results and discussion", although the uncertainty is large, the soft constraint still entails useful statements about temporal changes in the global recording rate for large eruptions.

### Our statistical non-parametric approach

The notation is the same as the previous sections. Let there be  $k$  non-overlapping periods, with known lengths  $\Delta_1, \dots, \Delta_k$ , uncertain recording rates  $\boldsymbol{\pi} := (\pi_1, \dots, \pi_k)$ , and for which the number of recorded large eruptions is  $\boldsymbol{x} := (x_1, \dots, x_k)$ . A statistical assessment of  $\boldsymbol{\pi}$ , including a quantification of uncertainty, requires a probabilistic model for  $\boldsymbol{x}$ .

We are considering global sums of large eruptions, and therefore we treat  $\boldsymbol{x}$  as arising from a homogeneous Poisson process with unknown rate denoted  $\lambda$  (units of/yr). This simple model is suggested by De la Cruz-Reyna (1991), who also noted that it would be less applicable for individual volcanoes. It is not possible to test the time-homogeneity of this model if the recording rate varies by period. Instead, homogeneity becomes the condition under which the recording rates are identified (Solow 1993, makes a similar point).

Most of the papers discussed in Section "Review of previous approaches" (Coles and Sparks 2006; Deligne

et al. 2010; Furlan 2010; Mead and Magill 2014) used a 2D Poisson process in time and magnitude, which is homogeneous in time. Because this implies a 1D Poisson process which is homogeneous in time, our model is more general. In lumping all eruptions of at least M4 together, we have ignored information in the database which these other papers have used. This is a deliberate choice on our part, due to the difficulty of incorporating the magnitude information, which we consider more than offsets the benefits of the extra information.

First, a 2D Poisson process over time and magnitude has many more possible implementations than a 1D process over time alone (which has only one). The common choice of all four papers cited above is to parameterise this process according to the peaks over threshold (POT) model for extremes, based on asymptotic theory (Pickands 1971). This requires three parameters rather than just one, and there is also the delicate issue of whether the asymptotic conditions hold, or whether some other justification might be found for the POT model.

Second, to implement this model, the papers cited above use a likelihood function defined with observed times and magnitudes, whereas we only use times. (Kiyosugi et al. 2015, use a separate recording function for each for each magnitude class using the Volcanic Explosivity Index). Thus all papers implicitly assume that magnitudes have been accurately recorded, and that, *a fortiori*, this accuracy has not changed through time. But magnitude is always imperfectly inferred. Mead and Magill (2014) and Kiyosugi et al. (2015), for example, implicitly assume that VEIs have been accurately categorized for several thousands of years, despite the very different ways in which VEIs have been assessed for eruptions occurring at different times over the Holocene and the Quaternary, respectively. Evidence of misrecording of magnitudes over the last millenium is given in Fig. 1, which shows the magnitudes ‘piling up’ on the integer values 4 and 5. Furlan (2010, sec. 5.4) shows that inferences are sensitive to misrecording of magnitudes.

So, in suppressing the magnitude information, our results are robust to the specification of the 2D Poisson process over time and magnitude, and robust to magnitude measurement errors. This is not the case for the results of the other papers. The smooth-in-time model used by Coles and Sparks (2006) and Deligne et al. (2010) is also sensitive to timing errors, while the abrupt-in-time model used by Furlan (2010) and Mead and Magill (2014) is insensitive to timing errors, as is our model (because we bin the times into intervals).

Under our model, the likelihood function for  $(\boldsymbol{\pi}, \lambda)$  is

$$\begin{aligned}
 p(\mathbf{x} \mid \boldsymbol{\pi}, \lambda) &= \prod_{j=1}^k e^{-\Delta_j \pi_j \lambda} \frac{(\Delta_j \pi_j \lambda)^{x_j}}{x_j!} \\
 &= e^{-\lambda \sum_j \Delta_j \pi_j} \cdot \lambda^s \cdot \prod_{j=1}^k \frac{(\Delta_j \pi_j)^{x_j}}{x_j!}
 \end{aligned}
 \tag{6}$$

where  $s := \sum_j x_j$ . The rate  $\lambda$  is a nuisance parameter which can be eliminated, in order to focus on  $\boldsymbol{\pi}$ . Let  $\lambda$  have a Gamma distribution with shape  $a$  and rate  $b$  (units of yr), where both these values are specified. Then integrating out  $\lambda$  gives the likelihood function for  $\boldsymbol{\pi}$ :

$$\begin{aligned}
 p(\mathbf{x} \mid \boldsymbol{\pi}; a, b) &= \int p(\mathbf{x} \mid \boldsymbol{\pi}, \lambda) p(\lambda; a, b) d\lambda \\
 &= \frac{b^a}{\Gamma(a)} \prod_j \frac{(\Delta_j \pi_j)^{x_j}}{x_j!} \cdot \left( \frac{(\sum_j \Delta_j \pi_j + b)^{s+a}}{\Gamma(s+a)} \right)^{-1} \\
 &\propto \prod_j (\pi_j)^{x_j} \cdot \left( \sum_j \Delta_j \pi_j + b \right)^{-(s+a)}
 \end{aligned}
 \tag{7}$$

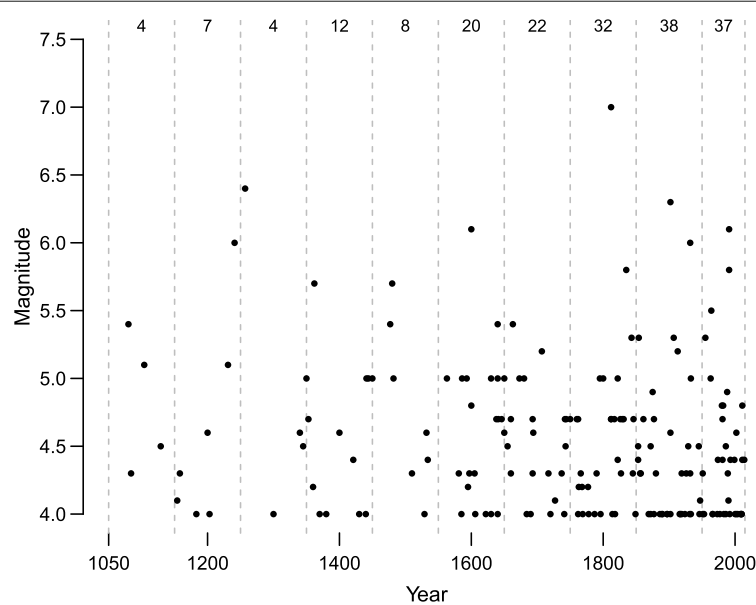
where  $\Gamma$  is the Gamma function, and the final line discards all multiplicative constants not involving  $\boldsymbol{\pi}$  see, e.g., Lunn et al. (2013, ch. 3). A similar approach was used by Solow (2001).

We need to specify  $a$  and  $b$  in the Gamma distribution for  $\lambda$ . The functional form of (7) suggests that  $a$  and  $b$  have a natural interpretation, in terms of dataset augmentation. The marginal distribution for  $\lambda$  carries the same weight as an additional period of length  $b$  yr which has  $a$  large eruptions. Informed by previous studies (e.g. Deligne et al. 2010), we assess the global rate of large eruptions at about one every two years, although acknowledge that this value is ‘contaminated’ by our exposure to the dataset  $\mathbf{x}$ . So we set  $a = 1$  and  $b = 2$  yr, with this small value for  $b$  indicating that, putting aside the information in the dataset, we have only vague beliefs about  $\lambda$ . Very similar results would follow in passing to the (improper) limit of  $a = 0$  and  $b = 0$  yr, which is completely vague. However, it is not possible to simulate  $\lambda$  from this limit, and simulation is required for code verification and for finding confidence sets.

This likelihood function in (7) has a closed-form maximum in  $\boldsymbol{\pi}$ , namely

$$\hat{\pi}_j = \frac{b}{a} \frac{x_j}{\Delta_j} \quad j = 1, \dots, k,
 \tag{8}$$

which can be verified by substituting into the first-order conditions. This expression is intuitive, because  $a/b$  is the expectation of  $\lambda$ , and so  $\Delta_j \cdot \hat{\pi}_j \cdot E(\lambda) = x_j$ , similar to the starting-point in Section “Review of previous approaches”. Replacing  $E(\lambda)$  with the estimate  $x_k/\Delta_k$  then leads to (3), although, as already discussed, this is a very poor estimator. As a Reviewer has noted, the ML estimator in (8) is insensitive to the strength of the information in the marginal distribution for  $\lambda$ , because it depends only on  $a/b$ , rather than on the value of  $b$  directly.



**Fig. 1** The dataset, of large (M4+) recorded stratovolcano eruptions, 1050–2014. The number of eruptions in century-wide intervals (excepting the most recent, which is 1950–2014) are shown at the top of the Figure

We adopt a more sophisticated approach and constrain the possible values that  $\pi$  can take, what we termed a ‘soft constraint’ in Section “Review of previous approaches”. Our basic beliefs about  $\pi$  are

1. The global large-eruption recording rate since 1980 is 1.
2. For sufficiently long periods, the global large-eruption recording rate of an earlier period is never higher than that of a later period.

The choice of 1980 is determined empirically and we use periods of 100 years, see Section “Results and discussion”. These two beliefs define the parameter space as

$$\Omega := \{ \pi : 0 \leq \pi_1 \leq \dots \leq \pi_{k-1} \leq \pi_k = 1 \}, \tag{9}$$

taking the final period to be 1980–2014. This parameter space generalizes the parametric models given in (4) and (5), since it is possible for the  $\pi_j$ ’s to decline smoothly, or abruptly, or both. The constraints in the parameter space allow observations from later periods to influence the estimates of  $\pi_j$  in earlier periods; in this way these early  $\pi_j$ ’s can ‘borrow strength’ from the later periods, to make up for the paucity of observations.

The ML estimator for  $\pi$  is now

$$\hat{\pi}(\mathbf{x}) := \max_{\pi \in \Omega} p(\mathbf{x} | \pi). \tag{10}$$

Although  $\Omega$  is hard to explore in a sequential fashion, not having a simple interior, it is easy to enumerate a large set of elements of  $\Omega$ , from which  $\hat{\pi}$  can be approximated using the maximum value from the set. This algorithm is

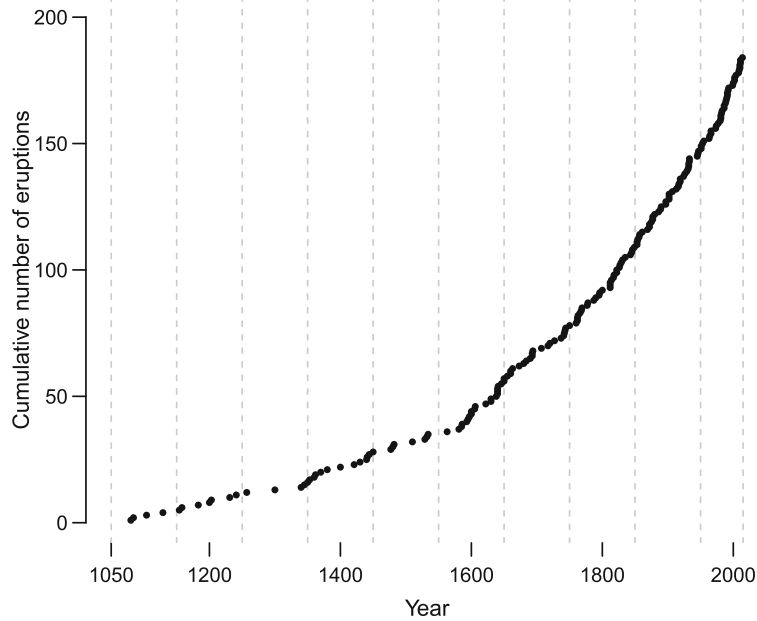
described in the Appendix, along with a method for computing a 95 % confidence set for  $\pi \in \Omega$ , which implies a 95 % confidence interval for each  $\pi_j$ .

As the Appendix demonstrates, it is quite hard to assess uncertainty about  $\pi$ . A Bayesian approach is more straightforward, but requires a prior distribution on  $\Omega$ . However, it is not possible to specify such a prior distribution without introducing additional beliefs to the two basic beliefs given above. ‘Obvious’ priors such as uniform on  $\Omega$  are actually highly constricting, since they impose a linear relationship on the prior expectations of the  $\pi_j$ ’s, namely  $E(\pi_j) = j/k$ . Thus, for a large  $k$ , the prior distribution for  $\pi_1$  is forced towards zero. We do not want to incorporate any additional beliefs about  $\pi$ , and therefore we have eschewed a Bayesian approach.

### Results and discussion

The dataset is drawn from the LaMEVE database, see Crossweller et al. (2012) and Brown et al. (2014), version 3.1, downloaded Oct 2015. Our code, including the dataset, is available from the first author, implemented in the statistical computing environment R (R Core Team 2013).

The dataset is visualised in Fig. 1, with simple summaries in 100-year intervals. Figure 1 shows that most of the under-recording has happened in the distant past and at the lower end of the magnitudes, consistent with the narrative evidence on under-recording (e.g., Siebert et al. 2010). Suppressing the magnitude information gives the cumulative eruptions sequence, shown in Fig. 2: the convex shape indicates an increasing recording rate. There

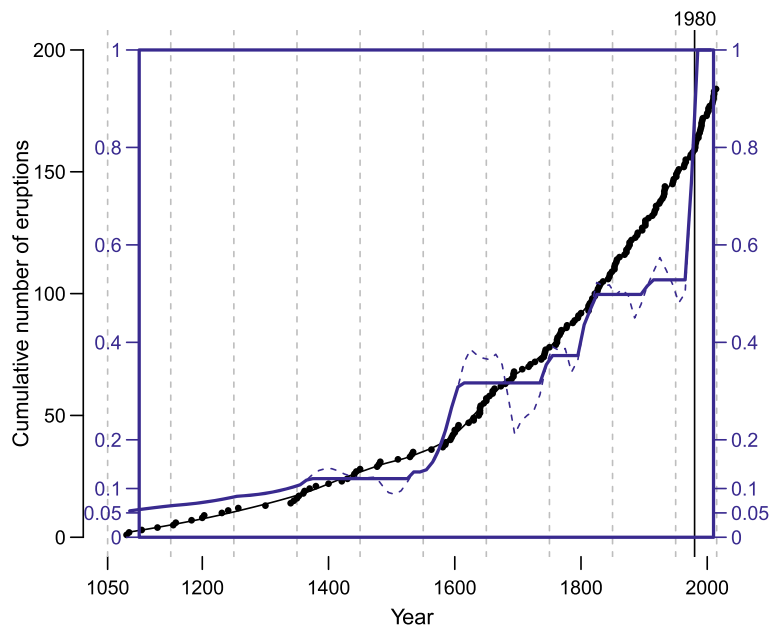


**Fig. 2** Cumulative eruption sequence, taken from Fig. 1. ‘Elbows’ between *straight lines* would indicate step changes between constant recording rates, taking the underlying eruption rate to be constant

is a distinct change in the recording rates just prior to 1600 but, otherwise, the shape of the curve looks quite smooth.

Figure 3 shows a crude point estimate based on the relative slope of a smooth line fitted through the cumulative curve shown in Fig. 2. After inspection of the recent part

of the cumulative curve we decided that the era of no under-recording could begin at 1980, and hence the average gradient of the curve for the period 1980–2014 was used to estimate the global large-eruption rate. Figure 3 includes both the raw estimate, and the estimate after

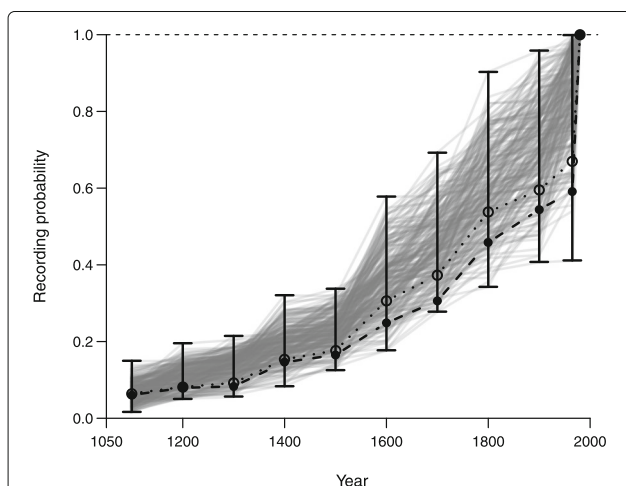


**Fig. 3** A crude point estimate of the global recording rate for large eruptions, superimposed on Fig. 2 (*dark blue* when reproduced in colour). After inspection of Fig. 2, the era of no under-recording is started at 1980. The point estimate is the ratio of the gradient of the smoothed curve to the average gradient of the smoothed curve post-1980. The raw values are shown as a *dashed line*, and the monotonically smoothed values as a *solid line*

monotonic smoothing, as suggested by Solow (1993). We regard this point estimate of the recording rate as very unreliable, being sensitive to choices such as the smoother and its bandwidth, and the way in which numerical gradients are computed. Nevertheless, its low value of about 0.06 in the year 1100 is striking.

Figure 4 visualises the 95 % confidence set for  $\pi$  for large ( $M \geq 4$ ) eruptions. Technically this is a Parallel Coordinates Plot (see, e.g., Venables and Ripley 2010, sec. 11.1); each grey line represents a point in the 95 % confidence set. The initial value of 1 and non-positive slope (going back in time) are imposed by  $\Omega$ , but the broadly exponential shape reflects the dataset, as does the low recording rate c1100. These features are consistent with the Coles and Sparks (2006) model, given in (4) above, with a value of  $b > 1$ . As anticipated, there seems to be a discontinuity around 1550, which is not a feature of the Coles and Sparks (2006) model. This discontinuity was picked up in the parametric step-change model fitted by Furlan (2010), given in (5) above, which put the step-change just prior to 1600. Furlan's (2010) dataset is similar but not identical to ours, as she uses all large eruptions, not just those for stratovolcanoes, and we have extended the dataset to include recent eruptions. But her claim that "the under-recording process largely disappears in the most recent 400 years" (Abstract, p. 113) is refuted in our analysis.

The confidence intervals for individual  $\pi_j$ 's are large, but the main message from Fig. 4 is clear: the global recording rates for large eruptions decline rapidly going back in time. Prior to 1600 they are below 50 %, and prior to 1100 they are below 20 %. Even in the recent past, e.g. the 1800s,



**Fig. 4** Visualisation of the 95 % confidence set for  $\pi \in \Omega$  for large ( $M \geq 4$ ) stratovolcano eruptions. Each grey line represents a point in the confidence set; the error bars are superimposed for convenience. The dots joined by a dashed line show the ML estimate. As an alternative point estimate, the open dots joined by a dotted line show the centroid of the 50 % confidence set

they are likely to be significantly less than 100 %. As a technical point, the error bars in Fig. 4 are margins of a 95 % confidence set, and as such their marginal coverage is *at least* 95 %. So the actual 95 % confidence intervals for each  $\pi_j$  are likely to be narrower than the error bars, and these statements are made with at least 95 % confidence.

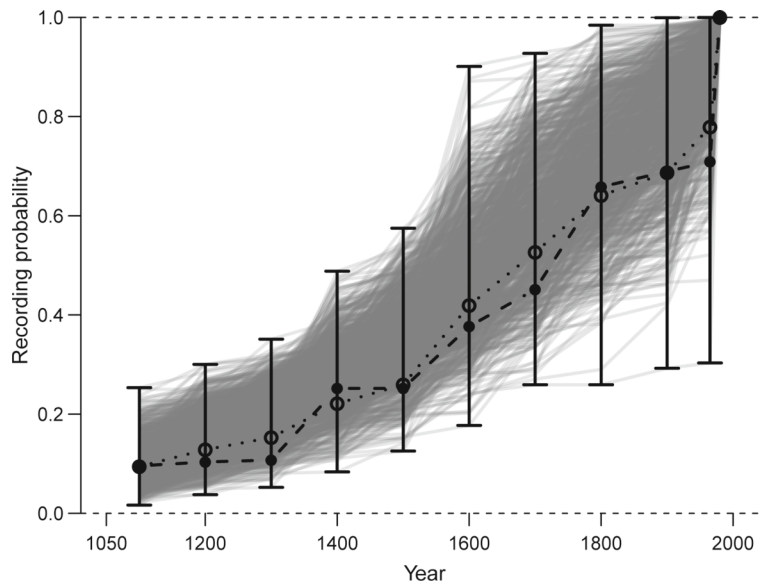
Close inspection of Fig. 1 shows that the recorded magnitudes 'pile up' at  $M = 4$  (and also at  $M = 5$ ). So possibly some of the eruptions recorded as  $M = 4$  eruptions are smaller than this, and should be excluded. To check the sensitivity of our results to this source of mis-recording, we repeated the analysis for eruptions with recorded  $M \geq 4.1$ ; the results are shown in Fig. 5. The uncertainty is larger, compared to Fig. 4, because 57 observations out of 184 have been dropped. But the qualitative conclusions are unchanged, and we are satisfied that our simple quantitative assessment of recording rates is robust to this source of mis-recording.

It is highly likely that most of the under-recording occurs for eruptions at the lower end of the large magnitude range. To test this assertion we restricted the dataset to very large magnitude eruptions ( $M \geq 5.0$ ), and redid the calculation. Figure 6 shows the result: the number of  $M5+$  eruptions in the database is low (42), and our uncertainty about the recording rates is large. We also found that using eruptions  $M \geq 5.1$  (24 eruptions) gave a similar confidence set, although the point estimates were different. According to Fig. 6, the recording rate is likely to be higher for  $M5+$  than for  $M4+$ , but we do not think the uncertainty assessment in Fig. 6 is reliable.

The confidence set figures include an additional curve, the centroid of the 50 % confidence set, as an alternative point estimate to the ML estimate. This centroid will tend to be smoother than the ML estimate, which can get closer to the edges of  $\Omega$ . But the difference between the two point estimates is mostly less than 10 percentage points for  $M4+$ , which is small in relation to the width of the confidence set. Overall, we hesitate to provide a point estimate for the recording rates from this dataset, given the volume of the 95 % confidence set.

## Conclusions

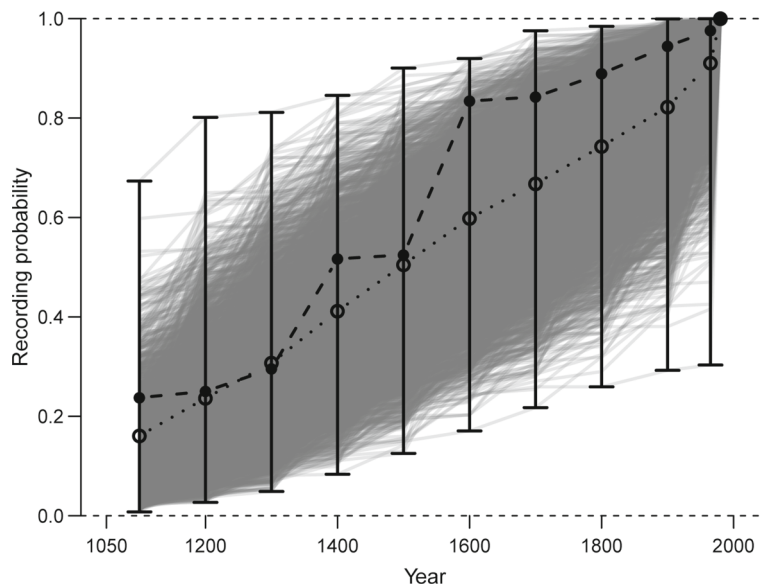
There is empirical and narrative evidence that the recording rate for volcanic eruptions was lower in the past than it is today. We would like to estimate historical under-recording rates, to better interpret the historical record, and, by extension, to better predict the future. Yet we have few beliefs about the quantitative features of the under-recording process. Hence in this paper we have adopted a non-parametric approach to estimate the recording rate as a function of time, in which the global large-eruption rate is treated as constant in time, and the only two constraints on the recording rates are that the global recording rate is now one (since 1980),



**Fig. 5** Same as Fig. 4, but for magnitudes  $M \geq 4.1$ , to investigate sensitivity to ‘piling up’ at  $M = 4$  shown in Fig. 1

and that the global recording rates are non-increasing, going backwards in time. Compared to the parametric approach used previously in the volcanological literature (Sections “Review of previous approaches” and “Our statistical non-parametric approach”), our approach has the advantage of being far more general, but may not reduce our uncertainty much below its current levels. This possibility is especially acute given our focus on large ( $M4+$ ) stratovolcano eruptions, which occur at a global rate of only about one every two years.

Fortunately, the dataset from the LaMEVE database is sufficiently rich that we can derive a clear message about historical under-recording (Section “Results and discussion”), including that the global recording rate drops below 100 % even in the recent past (e.g. in the 1800 s), is below 50 % before 1600, and is below 20 % before 1100 (all with 95 % confidence). These statements, which are formally about the global recording rate, apply, approximately, to the average recording rate across volcanoes (Section



**Fig. 6** Same as Fig. 4, but for magnitudes  $M \geq 5.0$



“Definition of the recording rate”). We also find that the sequence of global recording rates appears to combine aspects of both of the proposed parametric models, being generally exponential in shape, but also having more abrupt changes, notably around 1550. We extended our approach to very large (M5+) eruptions, but we judged our results to be unreliable.

Our results are consistent with the narrative evidence in Siebert et al. (2010, pp. 31–34) and Brown et al. (2014). Siebert et al. (2010) attribute the step-change around 1550 to the Spanish and Portuguese explorations at the end of the 15th century, which “opened Latin America and much of the western Pacific to European record-keeping” (p. 31), and also to the development of printing.

Finally the methods we have used here may also be useful in assessing under-recording of other hazards, e.g. earthquakes and floods. Like volcanoes, these infrequent hazards suffer from a sparse record of larger events, that is further thinned by low recording rates in the past. Our methods and computer code should be easily adaptable to other applications, and we would be happy to discuss this further.

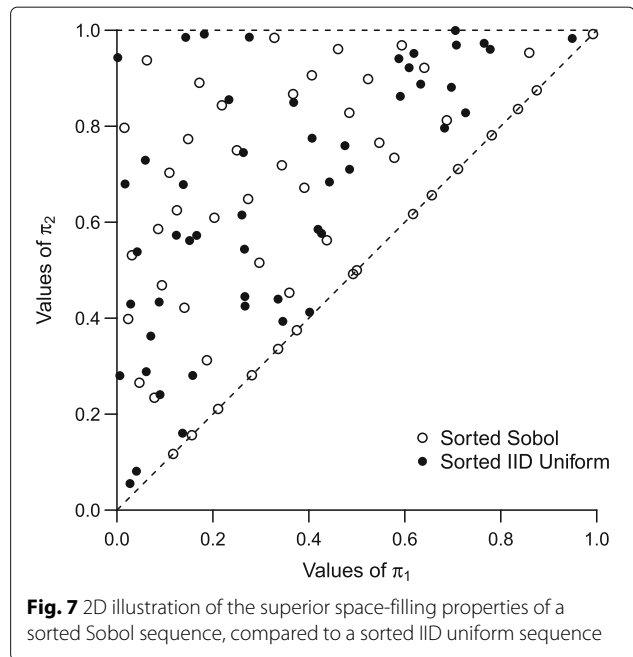
**Appendix: Computing a 95 % confidence set**

Putting aside  $\pi_k$ , which is equal to 1, the order statistics of  $k - 1$  independent uniform random quantities are uniformly distributed on  $\Omega$ , see Cox and Hinkley (1974, Appendix 2). This result provides a (stochastic) method to generate a point uniformly in  $\Omega$ : generate  $k - 1$  independent uniform random quantities, order them from smallest to largest, and then append 1.

There is a simple improvement on this method, if the intention is to generate a sequence of such points. Instead of using a sequence of random points, the points are generated according to a deterministic space-filling design on  $[0, 1]^{k-1}$ , in this case a Sobol sequence. This sequence, which does not suffer from random variability, will have better coverage of any large connected subset of  $\Omega$ , because it decreases the clumpiness of the points, relative to a random sample. Figure 7 illustrates this property in 2D.

The following algorithm enumerates a set of points occupying a particular level set of the log-likelihood function, using the space-filling approach described in the previous paragraph.

1. Generate an  $m$ -point Sobol sequence in  $k - 1$  dimensions. Order the values in each point from smallest to largest, and then append 1. This is now a deterministic  $m$ -point space-filling design on  $\Omega$ .
2. Compute the log-likelihood for each point using (7), denoted  $\ell_i$  for point  $\pi^{(i)}$ . Let  $\hat{i} = \arg \max_i \ell_i$ .
3. For some value  $c$ , keep all points  $\pi^{(i)}$  for which  $\ell_i \geq \ell_{\hat{i}} - c$ . That is, all points for which the



**Fig. 7** 2D illustration of the superior space-filling properties of a sorted Sobol sequence, compared to a sorted IID uniform sequence

log-likelihood is within  $c$  of the maximum log-likelihood. These points form a space-filling design in some subset of  $\Omega$ , denoted  $\mathcal{C}$ .

A Sobol sequence can be generated using the `sobol` function from the `randtoolbox` package in the statistical computing environment R (R Core Team 2013).

Let  $\chi_d^{-2}$  be the quantile function of the chi-squared distribution with  $d$  degrees of freedom. According to the standard asymptotics, setting

$$c \leftarrow \chi_{k-1}^{-2}(0.95)/2 \tag{11}$$

will imply that  $\mathcal{C}$  is a 95 % confidence set for  $\pi$ ; see Cox (2006, ch. 6) or van der Vaart (1998, ch. 16). But the standard asymptotics do not hold here, because the sample size is small, and the space  $\Omega$  is irregular. In a nutshell, the 95 % level set of the likelihood function is not elliptical, because it is clipped by the edges of  $\Omega$ . Therefore  $c$  must be adjusted to correct for ‘level error’, which is where the actual coverage is not equal to the nominal coverage.

The adjusted value, say  $c^*$ , is found empirically, in a simulation experiment using synthetic datasets; see DiCiccio and Efron (1996, notably sec. 7) for bootstrap methods for approximating confidence sets. Set  $\hat{\pi} := \pi^{(\hat{i})}$ , and generate  $n$  independent samples  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$  from the model with  $\pi \leftarrow \hat{\pi}$ . This involves, for each sample

1. Simulate  $\lambda \sim \text{Gamma}(a, b)$ ,
2. For  $i = 1, \dots, k$ , simulate  $x_i \sim \text{Poisson}(\Delta_i \cdot \hat{\pi}_i \cdot \lambda)$ .

For each sample, evaluate the log-likelihood for all points  $\pi^{(1)}, \dots, \pi^{(m)}$ , denote this  $\ell_{ij} := \log p(\mathbf{x}^{(j)} | \pi^{(i)})$ .

Also identify the maximum log-likelihood value,  $\hat{\ell}_j := \max_i \ell_{ij}$ . The value  $c^*$  is chosen so that 95 % of the sample contain the 'true' value  $\hat{\pi}$ . In other words,  $c^*$  solves

$$n^{-1} \sum_{j=1}^n \mathbb{1} \left\{ \ell_{ij} \geq \hat{\ell}_j - c^* \right\} = 0.95$$

where  $\mathbb{1}\{\cdot\}$  is the indicator function. A simple rearrangement shows that  $c^*$  is the 95th percentile of the histogram of  $\{\hat{\ell}_j - \ell_{ij}\}_{j=1}^m$ .

Denote the resulting level set based on  $c^*$  as  $\mathcal{C}^*$ . This empirical approach ensures that the coverage of  $\mathcal{C}^*$  is very close to 95 % at  $\pi = \hat{\pi}$ , and, one hopes, approximately 95 % in the region of  $\Omega$  around  $\hat{\pi}$ . Thus the resulting  $\mathcal{C}^*$  is approximately an exact 95 % confidence set.

In our calculation we used  $m \leftarrow 10^4$  and  $n \leftarrow 1001$ , and the 95 % confidence set cut-off at  $\hat{\pi}$  was found to be  $c^* = 4.61$ . By way of contrast, the asymptotic cut-off from (11) is  $c = 9.15$ . In order to check the coverage of this cut-off at plausible values other than  $\hat{\pi}$ , we chose ten other points at random from the 95 % confidence set, and evaluated the coverage of the confidence set at these  $\pi$ 's using the above values of  $m$ ,  $n$ , and  $c^*$ ; the coverages were 94, 92, 94, 92, 94, 93, 95, 93, 95, 92 %, which is reassuring.

#### Acknowledgements

This work was supported by the Natural Environment Research Council (NERC) funded Consortium on Risk in the Environment: Diagnostics, Integration, Benchmarking, Learning and Elicitation (CREDIBLE); grant number NE/J017450/1. Sparks acknowledges funding from the European Research Council in the VOLDIES project which led to the development of the LaMEVE database. Cashman acknowledges support from the AXA Research Fund. We would like to thank two anonymous Reviewers who helped improve the description of our approach and how it differs from past parametric approaches.

#### Authors' contributions

JR carried out the statistical analysis, SS and KC provided volcanological expertise. All authors read and approved the final manuscript.

#### Competing interests

The authors declare they have no competing interests.

#### Author details

<sup>1</sup>School of Mathematics, University of Bristol, University Walk, BS8 1TW Bristol, England. <sup>2</sup>School of Earth Sciences, University of Bristol, Bristol, England.

Received: 22 May 2015 Accepted: 18 September 2016

Published online: 10 October 2016

#### References

- Brown S, et al. Characterisation of the Quaternary eruption record: Analysis of the Large Magnitude Explosive Volcanic Eruptions (LaMEVE) database. *J Appl Volcanol.* 2014;3(5):. <http://www.appliedvolc.com/content/3/1/5>.
- Coles S, Sparks R. In: Mader H, Coles S, Connor C, Connor L, editors. Extreme value methods for modelling historical series of large volcanic magnitudes. Geological Society of London: Special Publication of IAVCEI; 2006, pp. 47–56.
- Cox D. Principles of Statistical Inference. Oxford University Press. Oxford: UK; 2006.
- Cox D, Hinkley D. Statistics Theoretical. Chapman and Hall. London: UK; 1974.

- Croweller H, et al. Global database on Large Magnitude Explosive Volcanic Eruptions (LaMEVE). *J Appl Volcanol.* 2012;1(4):. <http://www.appliedvolc.com/content/1/1/4>.
- De la Cruz-Reyna S. Poisson-distributed patterns of explosive eruptive activity. *Bull Volcanol.* 1991;54:57–67.
- Deligne N, Coles S, Sparks R. Recurrence rates of large explosive volcanic eruptions. *J Geophys Res.* 2010;115(B06):203.
- DiCiccio T, Efron B. Bootstrap confidence intervals. *Stat Sci.* 1996;11(3): 189–212. with discussion and rejoinder, 212–228.
- Furlan C. Extreme value methods for modelling historical series of large volcanic magnitudes. *Stat Modell.* 2010;10(2):113–32.
- Guttorp P, Thompson M. Estimating second-order properties of volcanicity from historical data. *J Am Stat Assoc.* 1991;86:578–83.
- Kiyosugi K, Connor C, Sparks R, Croweller H, Brown S, Siebert L, Wang T, Takarada S. How many explosive eruptions are missing from the geologic record? Analysis of the quaternary record of large magnitude explosive eruptions in Japan. *J Appl Volcanol.* 2015;4(1):. doi:10.1186/s13617-015-0035-9.
- Lunn D, Jackson C, Best N, Thomas A, Spiegelhalter D. The BUGS Book: A Practical introduction to Bayesian Analysis. Boca Raton FL, USA: CRC Press; 2013.
- Mason B, Pyle D, Oppenheimer C. The size and frequency of the largest explosive eruptions on earth. *Bull Volcanol.* 2004;66(8):735–48.
- Mead S, Magill C. Determining change points in data completeness for the Holocene eruption record. *Bull Volcanol.* 2014;76(14).
- Pickands J. The two-dimensional Poisson process and extremal processes. *J Appl Probab.* 1971;8:745–56.
- R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria; 2013. <http://www.R-project.org/>.
- Siebert L, Simkin T, Kimberly P. Volcanoes of the World. Berkeley and Los Angeles CA, USA: University of California Press; 2010.
- Simkin T. Terrestrial volcanism in space and time. *Ann Rev Earth Planetary Sci.* 1993;21:427–52.
- Solow A. Estimating record inclusion probability. *Am Stat.* 1993;47(3):206–8.
- Solow, A. An Empirical Bayes analysis of volcanic eruptions. *Math Geology.* 2001;33(1):95–102.
- van der Vaart A. Asymptotic Statistics. Cambridge, UK: Cambridge University Press; 1998.
- Venables W, Ripley B. Modern Applied Statistics with S, 4th edn. USA: Springer, New York NY; 2010.

Submit your manuscript to a SpringerOpen® journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Immediate publication on acceptance
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](http://springeropen.com)