Correcting Under-Recording of Eruptions in Historical Volcano Data

Oliver Stoner

[1] University of Exeter, United Kingdom

E-mail for correspondence: ors203@exeter.ac.uk

Abstract: Historical volcano data can suffer from under-recording of eruption occurrence, which can vary with time and magnitude. A Bayesian hierarchical framework is employed, to model simultaneously the true eruption rate and the under-recording mechanism, in order to obtain a more reliable inference on the relationship between eruption magnitude and frequency.

Keywords: Hierarchical; Bayesian; Natural Hazards.

1 Introduction

The LaMEVE dataset is a record of historic eruptions, with each entry including an estimated eruption year and an estimate of the magnitude of the eruption. As in *Rougier et al.*, this is defined by:

$$\text{magnitude} = \log_{10}(\text{erupted mass in kg}) - 7$$

Unfortunately, the recording of volcano eruptions is not complete; some entries rely on historical records, while many rely on geological analyses, where the likelihood of an eruption leaving a discoverable trace depends on the location, time and magnitude of the eruption (*Rougier et al.*).
This means that any inference on the temporal profile of eruption rates which assumes complete recording is likely biased. It is therefore desirable to quantify the under-recording, such that the frequency of eruptions can be more reliably investigated.

2 Methodology

A framework for correcting under-recording in count data has been recently developed and presented in *Stoner et al.*. In this framework, the recorded

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018.

count z is modelled as a Binomial quantity, where the number of trials is an unobserved Poisson quantity y, representing the true value of the count that has been incompletely recorded. The true count generating process is then modelled through the mean of the Poisson quantity λ, while the under-recording mechanism can be modelled through the Binomial probability π, analogous to the recording probability, to mitigate the bias involved in inference on both processes. The basic structure of the model is hence given by:

$$z \mid y \sim \text{Binomial}(\pi, y); \qquad y \sim \text{Poisson}(\lambda) \tag{1}$$

In order to use this framework, the data must be aggregated into counts of eruptions over a chosen time interval and to achieve this the data were aggregated over 1000 intervals of 100 years.

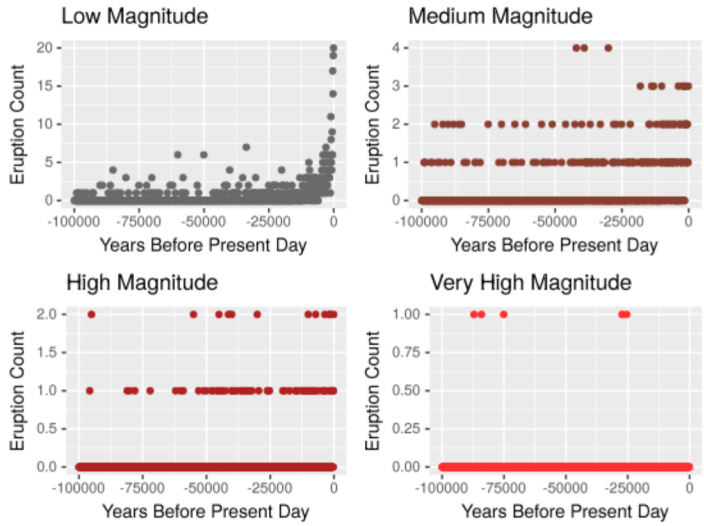

FIGURE 1. Total eruption counts for each of the last 1000 centuries, separated by magnitude.

As discussed in *Rougier et al.*, a large portion of eruption magnitude estimates have been rounded to the nearest integer, which could introduce issues if eruption magnitude is treated as a continuous variable, such as potential overestimation of the rate of near-integer value magnitudes and underestimation elsewhere. To overcome this problem, we follow *Rougier et al.* by classifying the data into four bins based on magnitude: Low [4.5,5.5), Medium [5.5,6.5), High [6.5,7.5) and Very High [7.5,8.5). The resulting set of eruption counts can be seen in Figure 1. Ignoring under-recording, the data appear to suggest the rate of eruptions has been dramatically increasing in recent centuries.

For an eruption in century $t \in T = 1, 2, ..., 1000$, where $T = 0$ represents the 21st century, and of magnitude $m \in M = \{\text{Low, Medium, High, Very High}\}$ the model is structured as follows:

$$z_{t,m} \mid y_{t,m} \sim \text{Binomial}(\pi_{t,m}, y_{t,m}) \quad (2)$$

$$\log\left(\frac{\pi_{t,m}}{1-\pi_{t,m}}\right) = \beta_{0,m} + \sum_{k=1}^{3} \beta_{k,m} w_{t,m}^{k} \quad (3)$$

$$y_{t,m} \sim \text{Poisson}(\lambda_m) \quad (4)$$

$$\log(\lambda_m) = \alpha_0 + \alpha_1 x_m + \epsilon_m \quad (5)$$

$$\epsilon_m \sim \text{Normal}(0, \sigma^2) \quad (6)$$

Here $w_{t,m}$ is the transformed century t ($w_{t,m} = \log(t+1)$ for $m =$ Low, $w_{t,m} = t/1000$ otherwise), and x_m is defined by the midpoint of the magnitude bin, minus the mean of the midpoints. The change in the recording probability $\pi_{t,m}$ is characterised by a different cubic polynomial for each magnitude bin in (3). A log-linear relationship between the eruption rate and magnitude is introduced in (5), such that information is pooled from the different bins into parameters α_0 and α_1. This is intended to aid in estimating the rate of Very High eruptions, of which there are very few observations. Additional flexibility is introduced by allowing the eruption rate for each bin to deviate from this line according to a Normal distribution, to allow for potential biases in the estimation of eruption magnitude.

A key concept in *Stoner et al.* is that the presence of under-recording means the information provided by the data is only partial, and must be supplemented to ensure parameter identifiability between λ_m and $\pi_{t,m}$. The simplest way to achieve this is to specify an informative prior distribution for $\beta_{0,m}$. Noting that the linear predictor of $\pi_{t,m}$ reduces to $\beta_{0,m}$ when $t = 0$, a Normal distribution with mean 4 and precision 2 was specified for each $\beta_{0,m}$, to represent a hypothetical belief that present day recording of eruptions is near complete. Finally, the number of non-zero observations for the Very High bin was determined to be too low to provide any meaningful inference on its change in recording probability over time, so the recording probability was fixed at 1 for this bin.

3 Results

The estimated change in the recording probabilities, for the first three magnitudinal bins, can be seen in Figure 2. All three curves show near monotonic decreasing trends going backwards in time, with a pattern of increasing recording probability for higher magnitudes generally holding.

Figure 3 shows the posterior mean estimates of the eruption rate for each magnitudinal bin, with associated 95% credible intervals. The solid black line represents the median predicted relationship between magnitude and rate, as defined in (5).

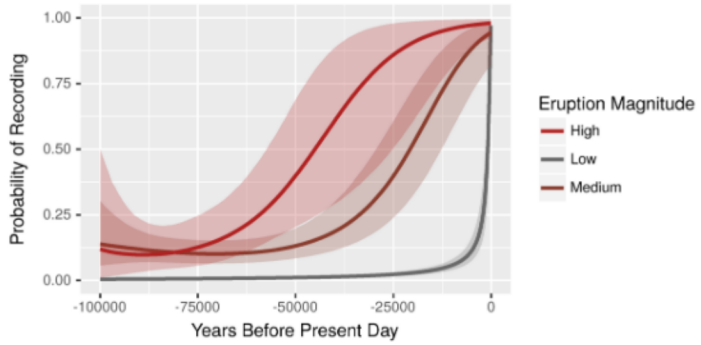

FIGURE 2. Posterior median estimates of the effect of time on the probability of recording a volcano eruption, for the first three magnitude bins, with associated 95% credible intervals.

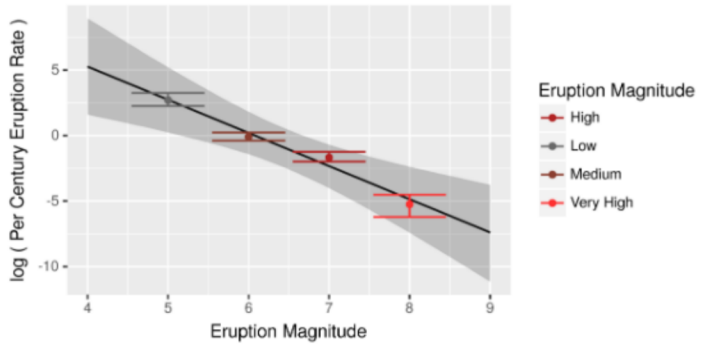

FIGURE 3. Plot showing the estimated eruption rate (on the log-scale) for the different magnitudinal bins, with associated 95% credible intervals. The solid black line shows the posterior median estimate of $\alpha_0 + \alpha_1 x_m$, with associated 95% credible interval.

Finally, the return period R for an eruption in a given magnitude classification m can be calculated as:

$$R_m = \frac{1}{\lambda_m}$$

In Table 1, it can be noted that the return period estimates for a Very High eruption is similar to the estimate in *Rougier et al.* for the return period of a volcano exceeding magnitude 8, which has a median of 17000 years, with associated 95% credible interval (5200,48000), though it is not yet clear if the two quantities are directly comparable.

TABLE 1. Return period (years) approximate predictive quantiles for an eruption in each of the four magnitudinal bins.

Magnitude	Lower 95%	Median	Upper 95%
Low	4.4	6.2	8.3
Medium	77	110	150
High	360	530	730
Very High	9100	19000	50000

4 Summary and Conclusions

In this article the challenges posed by under-recording were explored in the context of historic volcano eruptions. A dataset of volcanic eruptions was aggregated both by century and into four bins, based on eruption magnitude, which resulted in a dataset of counts. A general framework for correcting under-recording in count data was borrowed from *Stoner et al.*, the flexibility and generality of which meant that little adaption was necessary to design and implement an appropriate model for this problem. By accounting for the relationship between eruption magnitude and time and the under-recording mechanism, a more reliable inference for the relationship between eruption magnitude and rate of occurrence was made possible. This inference relies on the assumption that the rate of eruptions was constant in time over the period analysed, and the results could also be sensitive to the way in which the eruptions were aggregated both into centuries and into four magnitudinal bins. Both of these sensitivities are worthy of future investigation, though it can be noted that these results were similar to those in *Rougier et al.*, a study of the same data set with a substantially different approach.

Acknowledgments: Special Thanks to Professor Jonathan Rougier for introducing the author to the dataset and his existing work on the subject.

References

Rougier, J., Sparks, R.S.J., Cashman, K.V., and Brown, S.K. (2017). The global magnitude-frequency relationship for large explosive volcanic eruptions. *Earth and Planetary Science Letters*, **482**, 621 – 629.

Stoner, O., Economou, T., and Drummond, G. (Under Review). A Hierarchical Framework for Correcting Under-Reporting in Count Data. *Journal of the American Statistical Association.*