Accuracy and precision of frequency-size distribution scaling parameters as a function of dynamic range of observations: example of the Gutenberg-Richter law b-value for earthquakes

submitted to *Geophys. J. Int.*

Accuracy and precision of frequency-size distribution scaling parameters as a function of dynamic range of observations: example of the Gutenberg-Richter law *b*-value for earthquakes

G.-M. Geffers[1], I.G. Main[1], M. Naylor[1]

[1] *School of Geosciences, University of Edinburgh, EH9 3FE, UK. E-mail: G.Geffers@ed.ac.uk*

Accepted date; Received; in original form

SUMMARY

Many natural hazards exhibit inverse power-law scaling of frequency and event size, or an exponential scaling of event magnitude (m) on a logarithmic scale, e.g. the Gutenberg-Richter law for earthquakes, with probability density function $p(m) \sim 10^{-bm}$. We derive an analytic expression for the bias that arises in the maximum likelihood estimate of b as a function of the dynamic range r. The theory predicts the observed evolution of the modal value of mean magnitude in multiple random samples of synthetic catalogues at different r, including the bias to high b at low r and the observed trend to an asymptotic limit with no bias. The situation is more complicated for a single sample in real catalogues due to their heterogeneity, magnitude uncertainty and the true b-value being unknown. The results explain why the likelihood of large events and the associated hazard is often underestimated in small catalogues with low dynamic range, for example in some studies of volcanic and induced seismicity.

Key words: theoretical seismology, statistical methods, statistical seismology

1 INTRODUCTION

Many natural hazards (earthquakes, floods, storms, volcanic eruptions, avalanches) exhibit a power-law relationship between frequency and some physical variable of event size, such as energy, seis-

mic moment, volume, height (Turcotte, 1997). As a consequence of the large dynamic range of these event sizes, such power-law relationships are commonly expressed in the form of an exponential relationship between frequency and a logarithmic measure of size. The classic example of this is the Gutenberg-Richter (GR) frequency-magnitude relationship (Gutenberg & Richter, 1944), which has a probability density function (pdf) of the form

$$p(m) \sim 10^{-bm} = e^{-\lambda m}, \tag{1}$$

where m is magnitude, b is the GR 'b-value' and λ is its equivalent to the base e.

The scale-free form of the GR distribution must have a finite, yet uncertain upper bound to maintain the finite flux of seismic moment or strain energy, bounded by the finite tectonic strain rate we observe in the deformation field of the Earth, or equivalents such as flux of magma, sediment supply, or rainfall and in other applications to natural hazards (for volcanic eruptions, landslides and floods respectively). Main & Burton (1984) used this finite flux constraint to derive a modified version of the GR law with an exponential taper in the pdf at large seismic moments (rather than magnitudes). Most commonly, this tapered or modified Gutenberg-Richter (MGR) law is defined concisely by a cumulative frequency distribution of the form

$$F(M_0) \sim M_0^{-\beta} \exp(-M_0/M_\theta), \tag{2}$$

where M_0 is the seismic moment and M_θ is a 'corner moment' where the cumulative frequency $F(M)$ has dropped to a value $1/e$ less than that expected by the unbounded GR model (Kagan, 1991). The term β is the equivalent of the b-value in the magnitude-frequency relation and $\beta = 2b/3$.

The most commonly-applied method of estimating the b-value is the maximum likelihood estimate (MLE) derived analytically by Aki (1965) in the form

$$\bar{b} = \frac{\log_{10} e}{\bar{m} - m_c}. \tag{3}$$

where $\bar{m}$ is the mean magnitude and m_c is the threshold for complete reporting of smaller events.

Here, we present a new theory for the convergence of the b-value obtained from equation 3, based on random sampling of the mean magnitude $\bar{m}$ with respect to the number of events n and the dynamic range r of observations for the case of an exponential FMD for a randomly-sampled

finite sample. As an intermediary step, we use a maximum likelihood solution for the modal value in the mean magnitude obtained from a random sample of n events, and obtain the same result as Ogata & Yamashina (1986), who used the expectation value in their derivation. The dynamic range is the difference between the largest observed magnitude ω and m_c and is inherently linked to, but not proportionately correlated with the sample size n, with an additional degree of freedom due to the random sampling rather than a one to one correlation assumed in the theoretical curves we show later. Dynamic range is important in its own right because it defines the useful scale of observations, and the extent to which models can be tested in competition with one another. In many applications in seismology, dynamic range is restricted in practice to a relatively narrow range, particularly for volcanic and induced seismicity. We then derive the consequent convergence of the Aki-estimated b-value, denoted b_{Aki}, as a function of n and r. We test this model against randomly sampled catalogues with known underlying parameters, and against real data. We find the theory matches the observations well, and hence can be used to estimate the associated bias to high b-values in the common situation where the number of events and the dynamic range are small. This correction is important because high b-values result in an underestimation of the extrapolated likelihood of events larger than ω. In principle, the results could be applied in a similar way to other hazards with power-law frequency-size distributions that can be expressed in the form of equation 1.

2 THEORY

The probability density function for an exponential distribution with a scale factor $\frac{1}{\lambda}$ for a variable $x \geq 0$ is

$$p(x) = \lambda e^{-\lambda x}. \tag{4}$$

The probability density function for the mean $\bar{x}$ in a sample of size n taken from an exponential distribution takes the form

$$p(\bar{x}) = \frac{\lambda^n n^n}{\Gamma(n)} x^{n-1} e^{-n\lambda x} \tag{5}$$

where $\Gamma(n) = (n-1)!$ is the gamma function. Therefore, when $n = 1$, this is the standard

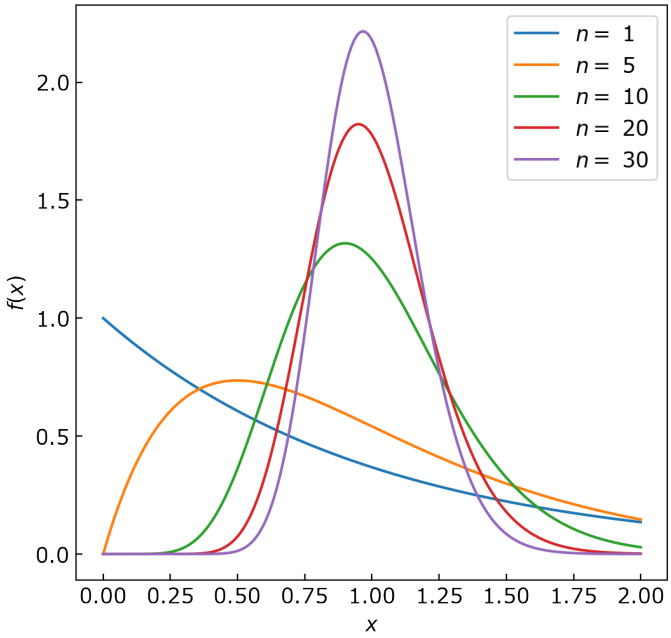

Figure 1. Plot of equation 5, where $\lambda = \frac{1}{n}$, so that $\bar{x}$ = 1.00 for an infinite sample. The peak probability (mode or maximum likelihood) occurs between $\bar{x}$ = 0 for n = 1 and $\bar{x}$ = 1 for an infinite sample.

exponential probability density function (Figure 1), where the most likely outcome (the mode) occurs at $\bar{x} = 0$. When n is large, $p(\bar{x})$ approaches a Gaussian distribution with a mode centred on the true mean $\bar{x}$. Prior to this the mode is less than the mean magnitude but greater than zero.

The maximum likelihood for the mean value of $\bar{x}$ in a finite sample is defined by the criterion $\frac{dp(\bar{x})}{dx} = 0$. This expresses the mean magnitude that is *most likely to be sampled* in a given population of trials, for example in the analysis of synthetic catalogues produced by random sampling of an underlying GR distribution. Substituting equation 5 for $p(\bar{x})$ using the product rule for differentiation then gives

$$\frac{\lambda^n n^n}{\Gamma(n)} \frac{d(\bar{x}^{n-1} e^{-n\lambda\bar{x}})}{dx} = \frac{\lambda^n n^n}{\Gamma(n)} \left[(n-1)\bar{x}^{n-2} e^{-n\lambda\bar{x}} - \bar{x}^{n-1} \lambda n e^{-n\lambda\bar{x}} \right] = 0. \quad (6)$$

The term outside the brackets is a constant, so the term in brackets must be zero. After taking out a common factor $x^{n-2}e^{-n\lambda x}$, we have

$$\bar{x}^{n-2} e^{-n\lambda\bar{x}} \left[(n-1) - \bar{x}\lambda n \right] = 0. \quad (7)$$

This implies a single maximum in $p(\bar{x})$ when

$$\bar{x} = \frac{(n-1)}{\lambda n}. \quad (8)$$

We recover $\bar{x} = 0$ at $n = 1$ and $\bar{x} = \frac{1}{\lambda} = \bar{x}_\infty$ for an infinite sample. The convergence for $\bar{x}$ then takes the form

$$\bar{x} = \bar{x}_\infty \frac{(n-1)}{n}. \tag{9}$$

Equations 9 and 3 together are equivalent to equation 7 of Ogata & Yamashina (1986), albeit derived by the maximum likelihood solution (near the mode in a finite sample) rather than the expectation value (near the mean). It confirms that we expect the most likely sample mean to start at zero in a sample of $n = 1$ and trend in a non-linear fashion asymptotically to $\bar{x} = \bar{x}_\infty$ from below.

Equation 9 specifies the convergence of the most likely mean magnitude in a finite sample of events above m_c as a function of the number of events n. However, the focus here is the dynamic range $r = \omega - m_c$ where ω is the largest magnitude in a sample. We acknowledge the inherent connection between n and the dynamic range, where often (although not exclusively), small dynamic range will also have low n and account for dynamic range in the next section.

Accounting for a finite minimum in x

If we have a finite threshold such that $x \geq x_{min}$, then equation 9 is rescaled to

$$\bar{x} - x_{min} = (\bar{x}_\infty - x_{min}) \frac{(n-1)}{n}. \tag{10}$$

In this section, we calculate a relationship between the total number of events in a sample above the threshold x_c and the dynamic range, assuming the total number of events in the largest sampling bin $n(\omega - dm, \omega) = 1$. We do this by calculating the total number of events in the sample for different threshold magnitudes 0, m_c and $\omega - dm$. In the case of a zero lower threshold, the integral of the probability density function is then

$$I_0 = \int_0^\infty \lambda e^{-\lambda x} dx = e^{-0} - e^{-\infty} = 1. \tag{11}$$

This equation proves the functional form of equation 4 is correct, i.e. unit total probability is achieved when the pre-exponential factor equals the exponent λ, so the *pdf* is specified by a single variable. In the case of a finite magnitude threshold m_c and finite sampled largest magnitude ω, the cumulative probability is

$$I_{m_c} = \int_{m_c}^{\omega} \lambda e^{-\lambda x} dx = e^{-\lambda m_c} - e^{-\lambda \omega} \tag{12}$$

and for the largest bin

$$I_{\omega-\delta m} = \int_{\omega-\delta m}^{\omega} \lambda e^{-\lambda x} dx = e^{-\lambda(\omega-\delta m)} - e^{-\lambda \omega}. \tag{13}$$

The ratio of the two cumulative probabilites I_{m_c} and $I_{\omega-dm}$ must be equal to the ratio of the numbers of events in the two samples n_{m_c} and $n_{\omega-dm}$. Given we know $n_{\omega-dm}$ = 1 and $\frac{n_\omega - m_c}{n_\omega - dm} = \frac{I_{m_c}}{I_\omega - dm}$ by proportion, we have

$$n_{m_c} = \frac{e^{-\lambda m_c} - e^{-\lambda \omega}}{e^{-\lambda(\omega-dm)} - e^{-\lambda \omega}} = \frac{e^{\lambda(\omega-m_c)} - 1}{e^{\lambda dm} - 1}. \tag{14}$$

Substituting this solution for $n(\lambda, m_c, \omega, dm)$ into equation 10, the mean magnitude then converges according to

$$\bar{m} - m_c = (\bar{m}_\infty - m_c)\left[1 - \left(\frac{e^{\lambda dm} - 1}{e^{\lambda(\omega-m_c)} - 1}\right)\right]. \tag{15}$$

The magnitude bin defining the largest single event dm is considered a free parameter since it may also depend on the uncertainty in ω and the proximity of the neighbouring bin. Equation 15 therefore has three free parameters, λ, dm and $\bar{m}_\infty$. It is easy to show from equation 15 that as ω tends to infinity, $\bar{m}$ will tend to $\bar{m}_\infty$ asymptotically from below as dynamic range increases, with a resulting increase in accuracy and decrease in finite sample bias. We can relate the left-hand side of equation 15 to the Aki b-value by restating equation 3 in the form

$$\lambda_{Aki} = \frac{1}{\bar{m} - m_c}. \tag{16}$$

Combining equations 15 and 16 then gives the following relationship

$$\frac{1}{\lambda_{Aki}} = \frac{1}{\lambda}\left[1 - \left(\frac{e^{\lambda dm} - 1}{e^{\lambda(\omega-m_c)} - 1}\right)\right]. \tag{17}$$

This formula can be used to correct for the systematic bias involved in the assumption that the mean magnitude $\bar{m}$ is a good approximation for the expectation value $\langle m \rangle$ for an infinite sample in the derivation of Aki (1965). Important to note is that the mean magnitude will be influenced to some extent through magnitude binning (Marzocchi et al., 2020), however, when bins are 0.1 or less, the bias is generally negligible (as in the current case). The bias in the estimate λ_{Aki} is then the difference between λ_{Aki} and λ. Again, it is easy to show analytically from equation 17 that

λ_{Aki} converges asymptotically to λ systematically from above, and hence b-values in data samples with small dynamic range are likely to be biased to high values.

3 METHODS AND DATA

The data used to test the hypotheses developed above include both synthetic and real earthquake magnitude data. For the synthetic data, we create 100 realisations on the real number line, of both 'perfect' GR and MGR catalogue data with the following parameters: n = 100 000, m_{min} = 0.5 and b-value = 1.0. For the MGR distribution, we take m_θ = 5.0 because this is representative of small to medium magnitude earthquakes, compared to, for example, $m_\theta \sim 8.5$ for the tectonic, global Harvard Centroid Moment Tensor (CMT) catalogue (Bell et al., 2013). The 95% confidence intervals on the randomly sampled data are obtained from the scatter of the Monte Carlo simulations, and compared to the errors obtained for the real data using equations 18 and 19 below which represent estimates of the irreducible random errors in $\bar{m}$ and b, and hence the precision of the estimate. Our estimate of confidence intervals for the synthetic data may not be ideal in the case of small samples, where a chi-square distribution might be preferable, but we preferred to apply a consistent method, recognising that the confidence intervals may not be ideally accurate in the smaller samples.

For the real earthquake catalogues, we have chosen two very different examples – The Geysers geothermal (induced) earthquake catalogue in California and the Southern California (tectonic) catalogue. Geffers et al. (2022) previously showed that The Geysers data are likely to exhibit an MGR preference at large dynamic ranges (and therefore large sample sizes), whereas in Southern California, GR is strongly preferred. This allows us also to examine the effect of the epistemic uncertainty caused by lack of knowledge of the underlying form of the distribution. In the following results section, we compare the outcomes for the synthetic GR and MGR data to those from The Geysers and Southern California data, and examine the extent to which equation 17 describes the data. The Geysers catalogue contains over 60 000 events in the complete part of the catalogue, where the magnitude of completeness m_c = 1.25 as estimated by the b-value stability method (Wiemer & Wyss, 2000; Cao & Gao, 2002). This method was also used to estimate m_c for

Southern California, given as 3.28. This leaves 9023 events remaining in the complete part of this catalogue. The largest observed magnitude in The Geysers catalogue is 5.01 and 7.30 for Southern California. This results in maximum observed dynamic ranges of 3.76 and 4.02, respectively.

To estimate the best fit free parameters in equations 15 and 17 for the synthetic data, we used the non-linear least-squares curve fit function *scipy.optimize.curve_fit* in Python. The *curve_fit* function then returns optimal values for the parameters. We find a secondary local mode that is an artefact of small samples containing fewer than 3 data points, described in more detail in the results section, so these outcomes are discarded before fitting the data.

For the case of the real catalogues, using the *curve_fit* function is problematic because it assumes that residuals are random, which can result in poor fits to the actual data. Therefore, we opted for a different approach when fitting equations 15 and 17 to the real data. The parameters $\bar{m}_\infty$ and λ_{Aki} (equivalent of b_{Aki}) were estimated instead from the value of the data point at largest dynamic range. This assumes that convergence is reached within the observed catalogue data. Having fixed these values at asymptotic convergence, we vary dm to the optimal incremental magnitude bin size which will fit equations 15 and 17 as closely as possible to the real catalogue data.

The primary sampling error in $\bar{m}$ at 95% confidence is

$$\delta\bar{m} = \pm\frac{1.96(\bar{m} - m_c)}{\sqrt{n}}. \tag{18}$$

After propagating this error in $\bar{m}$ (using equation 3) for a finite sample of n, Aki (1965) showed the equivalent uncertainty in the estimated value of b, i.e. $\hat{b}$ is

$$\delta b = \pm\frac{1.96\hat{b}}{\sqrt{n}}. \tag{19}$$

These error estimates represent the irreducible, random errors in our estimates of $\bar{m}$ and b. We consider an estimate of the mode in $\hat{b}$ to be accurate when the systematic error or bias $\hat{b} - b \leq 0.1$, and to be precise when $\delta b \leq 0.1$. In the case of the real data, the true values of $\bar{m}$ and b are not known, so we cannot define an accuracy. The equivalent (rounded) criteria for the estimates of mean magnitude are a bias $\hat{\bar{m}} - \bar{m}_\infty \leq 0.05$ and a precision $\delta m \leq \frac{0.1}{\ln(10)} \leq 0.05$. Strictly exact equivalence occurs when $\delta m \leq \frac{0.1}{\ln(10)}$ but pragmatically we choose the approximation $\frac{0.1}{2}$ here

in rounding to the nearest multiple of 0.05. In the case of the synthetic data, there are multiple catalogues, and the resulting data scatter provides a more complete range of uncertainties. In this case, the confidence intervals are estimated from the random data scatter. These reflect better the total uncertainty, including the epistemic uncertainty resulting from lack of complete knowledge in a single finite sample.

In the synthetic data, the true b-value is taken to be 1.0, consistent with many examples in the published literature where there is a good dynamic range of data (Frohlich & Davis, 1993; Kagan, 1999). However, in the real data, both the true distribution of the data and the true underlying b-value remains unknown, inevitably introducing a systematic uncertainty which we need to account for. Therefore, for the real data, we assume that convergence is reached within the data available here and that convergence to the asymptotic value of $\bar{m}_\infty$ and b is established at least approximately at largest dynamic range for both $\bar{m}$ and b. We acknowledge that this pragmatic assumption could lead to an unavoidable residual bias not accounted for in our analysis of the real data below. The estimated values of b and m_θ in the real data are obtained using equations 4 and 5 in Geffers et al. (2022), using the method of Kagan (2002) to define maximum log-likelihood functions for the distribution for a finite sample of n observations. The resulting best estimates β (equivalent of b) and M_θ (equivalent of m_θ) are obtained for the real data and are shown in Table 1.

4 SYNTHETIC AND REAL CATALOGUE RESULTS

We now present the results of the hypothesis test on both synthetic and real earthquake catalogue examples, using the theory and methods described in the previous two sections. The figures in this section all follow the same visual representation – the catalogue data are shown as blue circles, where red circles indicate discarded data ($n \leq 3$), orange circles indicate data with $n > 50$ and the solid black line represents the sampled mode from the remaining data. All of the discarded mean magnitude data with $n = 1$ in any sample plot on a straight line (shown in brown in the case of synthetic data in Figure 2) and represent a hard edge to the range of possibilities – a clear artefact of such small samples. This logic extends to the choice of $n \leq 3$ to avoid artefacts from a secondary local mode summarised in section 3. The area in light blue represents the 95%

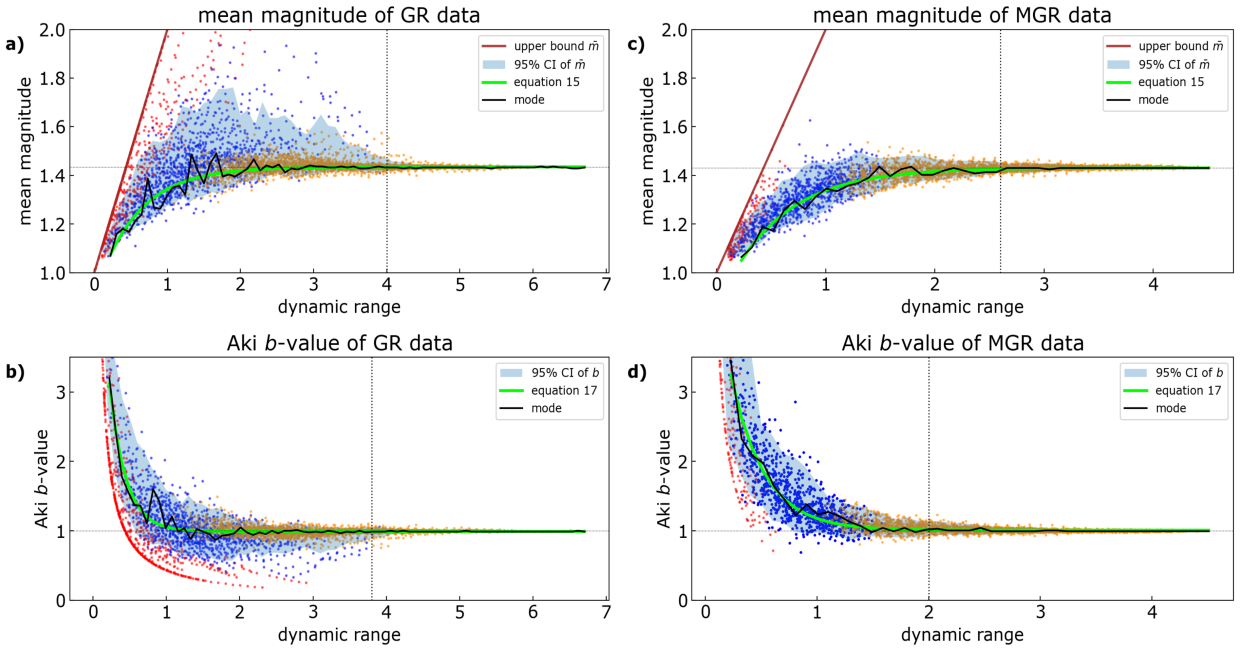

Figure 2. a) and c): mean magnitude as a function of dynamic range for GR and MGR data, respectively. **b) and d):** b-values as a function of dynamic range for GR and MGR data, respectively. The straight brown line in all plots indicates an upper bound to the mean magnitude in a sample of 1. The solid black line shows the mode for the plotted synthetic data (blue circles) and the bright green curve represents equation 15 in a) and c) and equation 17 in b) and d). The vertical dotted line represents the point at which both accuracy and precision are reached and the horizontal, dotted line represents the optimal values for convergence of $\bar{m}$ or b obtained from the curve fit. Red circles indicate discarded data where a sample $n \leq 3$. Orange circles indicate data where a sample $n > 50$.

confidence intervals of the outcomes determined by (a) the scatter in the results for the synthetic data or (b) the error estimates given in equations 18 and 19 of $\bar{m}$ and b for the single samples in the real data, also at 95% confidence. Faint, horizontal lines indicate the optimal convergence values of $\bar{m}$ and b returned from fitting equations 15 and 17.

In Figure 2a) and b), the data are randomly sampled from a GR distribution with $b = 1$. In Figure 2a), $\bar{m}$ converges from below as a function of dynamic range and there is good agreement between the mode and that predicted by equation 15 in bright green. For a dynamic range of ~ 1.4 and up, the mode already lies within $\pm$ 0.05 units of $\bar{m}_\infty$. Figure 2b) shows the corresponding Aki b-values as a function of dynamic range. The best fit (equation 17) converges extremely quickly to $b = 1.0 \pm 0.1$, similar to the mode, before a dynamic range of 1.0 is even observed. This convergence is somewhat faster than in the case of the mean magnitude because of the approximation involved

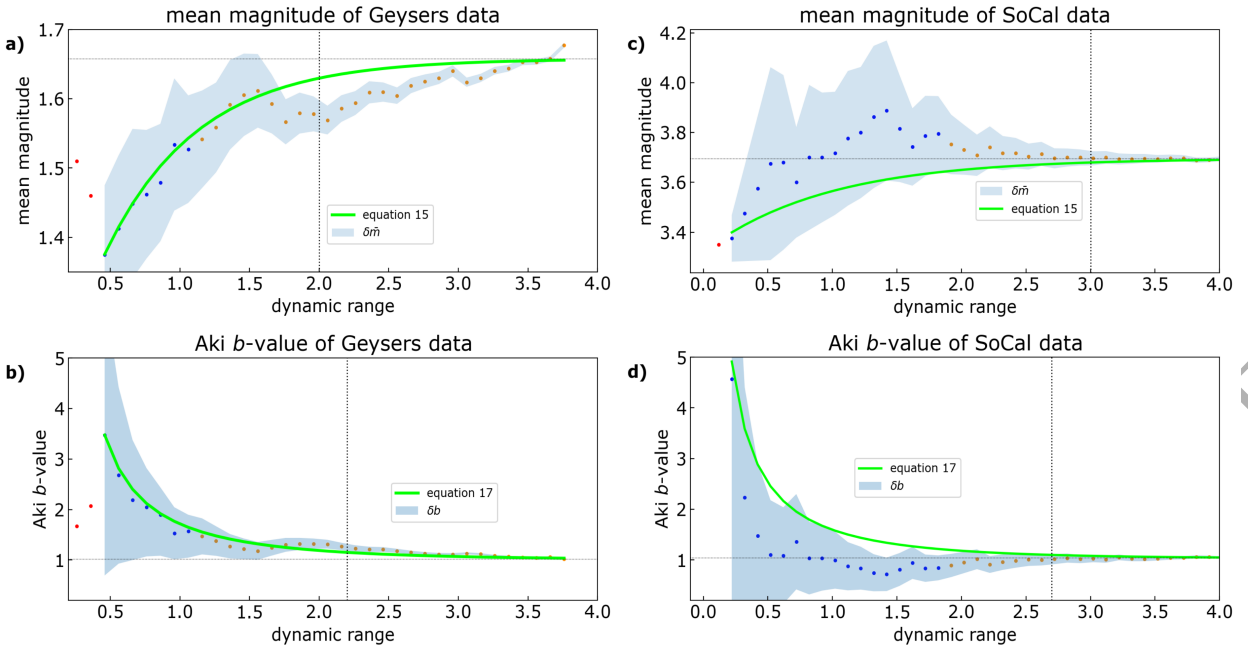

Figure 3. Presented as explained in the caption to Figure 2. **a) and c):** Mean magnitude as a function of dynamic range for The Geysers and Southern California data, respectively. **b) and d):** b-values as a function of dynamic range for The Geysers and Southern California data, respectively. Light blue shading represents the 95% confidence intervals from equations 18 and 19. The vertical dotted line represents the point at which precision is reached and the horizontal, dotted line represents the optimal values for convergence of $\bar{m}$ or b.

in rounding described above. In the case of data randomly sampled from an MGR distribution, the dynamic range observed in the synthetic data where both accuracy and precision are attained is considerably smaller than for the GR data, due to the effect of the 'corner' magnitude m_θ and the associated roll-off in frequency for larger events. In Figure 2c) and d), the data are randomly sampled from an MGR distribution with b = 1 and a corner magnitude of m_θ = 3.5. The best fit for $\bar{m}$ and b is very close to the observed modes although this time, in the case of the b-value, a dynamic range of around 1.3 is required for convergence to b = 1.0 $\pm$ 0.1 again, convergence to the benchmark value is slightly faster than in the case of the mean magnitude (Figure 2c) which requires a dynamic range of 1.6 to fall within $\pm$ 0.05 units of $\bar{m}_\infty$.

In the case of The Geysers data, the last datapoint in Figure 3a) is ignored in the fit, as appears to be an outlier, and hence may bias the fit to the other estimates. For both $\bar{m}$ and b in the case of The Geysers (Figure 3a and b), the convergence of the best fits to equations 15 and 17 respectively,

Table 1. Threshold dynamic ranges for accuracy r_a and precision r_p at 95% confidence as a function of m_c, m_∞, and the assumed or estimated b-value and corner magnitude m_θ. The bias or precision respectively are defined when they are equal to 0.05 units in $\bar{m}$ and 0.1 units in b (Figure 2). The accuracy cannot be estimated for the real data because the underlying values are not known.

Figure	**Data**	$\boldsymbol{r_a(\bar{m})}$	$\boldsymbol{r_p(\bar{m})}$	$\boldsymbol{r_a(b)}$	$\boldsymbol{r_p(b)}$	$\boldsymbol{m_c}$	estimated $\boldsymbol{\bar{m}_\infty}$	estimated $\boldsymbol{b}$-value	estimated $\boldsymbol{m_\theta}$
2a) and b)	synthetic GR	3.6	4.0	3.0	3.8	1.0	1.43	1.00	3.5
2c) and d)	synthetic MGR	1.7	2.6	1.7	2.0	1.0	1.42	1.00	3.5
3a) and b)	The Geysers	-	2.0	-	2.2	1.25	1.66	1.02	$\sim$ 4.6
3c) and d)	Southern California	-	3.0	-	2.7	3.28	3.70	1.04	$\sim$ 7.0 - 7.5

are very close to the predicted mode (specifically the maximum likelihood value) of the single sample catalogues across all dynamic ranges. Notably, in the case of $\bar{m}$, the fit overestimates the modal $\bar{m}$ between dynamic ranges of 1.5 and 3.5, most likely due to the variability of real data associated with catalogue heterogeneity. We estimate $\bar{m}_\infty$ is $\sim$ 1.66 and the associated b-value at convergence is $b \sim 1.0$. On the contrary, the best fits do a less good job in fitting to the Southern California data shown in Figure 3c) and d). This is also likely due the larger variability in the real data. The best fit parameters obtained in Figure 3 are for a) b = 1.5, dm = 0.05, b) b = 1.01, dm = 0.11, c) b = 1.05, dm = 0.01 and d) b = 1.03, dm = 0.012.

Table 1 presents a summary of the threshold dynamic ranges required for accuracy and pre-

Table 2. Convergence of the precision in the estimated b-value (δb) as a function of dynamic range for GR and MGR sampled data compared to Southern California and The Geysers data, respectively. For the synthetic data, δb is estimated from the scatter in outcomes at 95% confidence, denote δb_S and for the real data from equation 19, also representing 95% confidence, denoted δb_R. The rows in bold state the ratio of $\frac{\delta b_S}{\delta b_R}$.

Figure	**Data**	**r = 1.0**	**r = 2.0**	**r = 3.0**	**r = 4.0**
2a) and b)	synthetic GR	1.0	0.65	0.4	0.06
3b) and d)	Southern California	1.1	0.45	0.15	0.02
	ratio $\frac{\delta b_S}{\delta b_R}$	**0.9**	**1.4**	**2.7**	**3.0**
2c) and d)	synthetic MGR	0.65	0.2	0.06	-
3a) and b)	The Geysers	0.95	0.2	0.05	-
	ratio $\frac{\delta b_S}{\delta b_R}$	**0.7**	**1.0**	**1.2**	**-**

cision of $\bar{m}$ and b for both the synthetic and real datasets, including their estimated values of convergence for $\bar{m}$ and b. The increase in precision for both the synthetic (δb_S) and real (δb_R) with respect to r is shown in Table 2. In all cases, precision increases with respect to r. The synthetic data emulating the Southern California example converges more slowly than that for the real data, indicating that the total error may be underestimated by equation 19, by a factor in the range 0.9 - 3.0. For the Geysers data, convergence in δb_S and δb_R is more similar to that expected from equation 19, with a ratio near 1.0. Notably, the bias discussed here is mostly related to small r and n, and becomes less significant with increased r. While this may represent a minority of all studies of the earthquake frequency magnitude relation, these are over-represented in the case of volcanic and induced seismicity where the dynamic ranges are much smaller.

5 DISCUSSION

In our analytical theory, we have used a similar approach to that of Ogata & Yamashina (1986) who also used equations 4 and 5 to derive equation 9, as in this paper. However, there are differences. For example, we use a maximum likelihood solution for the mode of the distributions of mean magnitudes rather than the expectation value estimated from the mean value of a randomly selected set of mean magnitudes, but it is reassuring that the same result can be obtained from different methods. We then recast equation 9 in terms of dynamic range, and examine how well this equation fits results from a finite size sample with a maximum observed magnitude ω. We also consider the assumption that the mean sampled magnitude is a good approximation for the expectation value, which in a finite sample is generally not the case. The maximum observed magnitude in the cases here are those that emerge from finite samples of either the real data or the generated synthetic data, i.e. we have not limited this correction *a priori* by setting a specific upper bound before the sampling.

The analysis of the synthetic data has shown that the convergence of the mode in the mean magnitude expected in a finite sample (equation 15) in the case of an exponential FMD is in very good agreement with the observed value (Figure 2). While the trends are broadly similar, this is not as clearly the case for real data. Nevertheless, the analysis of The Geysers data seem to follow this

trend more closely than in the case of Southern California, where the picture is much more variable, leading to much larger uncertainties compared to The Geysers, most likely due to catalogue heterogeneity. Additionally, the real catalogues are already subject to larger uncertainties because the errors in both $\bar{m}$ and b for a single sample are not representative of the overall uncertainty one might expect on repeating the experiment many times (Marzocchi & Jordan (2017) provide a detailed discussion on caveats associated with handling different sources of uncertainties).

In the case of the synthetic data, we know that $b = 1.0$. However, in the case of the real catalogue data, we do not know one 'true' value of b, even though Geffers et al. (2022) have previously suggested that of The Geysers catalogue is likely to have a b-value close to 1.0. This falls within the b-values estimated in prior literature, ranging from $b \sim 0.8$ to 1.3 (Henderson et al., 1999; Convertito et al., 2012; Kwiatek et al., 2015; Leptokaropoulos et al., 2018). Similarly, $b \sim 1.0$ for Southern California (Kamer & Hiemer, 2015). These independent estimates are in agreement with the asymptotic b-values estimated in Table 1 of this study.

The improvement in accuracy of the estimated b-value in synthetic cases as a function of dynamic range (Table 1) can be attributed to the improvement in the estimated mean magnitude as sample sizes increase, captured in equation 17 and in agreement with Ogata & Yamashina (1986) who state that the bias in b is 'not small' when n is small. This also concurs with the reduced bias in the b-value with respect to dynamic range as portrayed by Figure 6 in Marzocchi et al. (2020). Table 1 highlights the fact that one can obtain answers that are accurate but not precise, precise but not accurate, or accurate and precise, depending on dynamic range and the nature of the input catalogues. Table 2 suggests that the uncertainty expressed in equation 19 may be an underestimate of the total variability in the results for the multiple realisations of the synthetic data for the same underlying distribution. Comparing the convergence of the b-value shown in Figures 2 and 3 as a function of dynamic range to previous studies (Marzocchi et al., 2020; Geffers et al., 2022), we show agreement that the bias reduces strongly if there are 3 orders of dynamic range available. In the case of synthetic GR data, we show that this decreases further with more dynamic range but it is unrealistic to expect such large dynamic ranges to be observed in any given real catalogue. The bias implied by the trend on the graphs for the real data is of a similar order of magnitude to that

of the synthetics, but this will definitely be affected by binning forced by magnitude precision in the case of the real data. Hence, it looks as if this additional bias may be smaller or similar to, but not significantly larger than, that of the finite sampling, at least for small to intermediate dynamic ranges. At larger dynamic range, the real data does show a residual systematic bias that cannot be accounted for by the ideal theory alone.

The theory for the sampled synthetic data explicitly assumes events are independent and identically-distributed (*iid*) in the case of a perfect, homogeneous catalogue of magnitudes with zero measuring uncertainty. However, in a real catalogue, the magnitude estimates are subject to uncertainty, and are measured indirectly from an evolving network with a finite number of stations with a finite sample of the radiation pattern, which in turn leads to much larger variability in the the convergence trend and often a lack of convergence to a flat asymptote. It is worth noting that the catalogues used here have not been declustered, as most declustering techniques (Gardner & Knopoff, 1974; Gerstenberger et al., 2020) do not preserve the uncorrelated magnitude distribution because they systematically remove smaller magnitude events.

There is some clear dynamic range dependent natural variability in the trends of the real data beyond that which can be explained by the theory in the case of the real catalogues, including at larger dynamic ranges (Figure 3). Our interpretation is that this is most likely due to a mixture of catalogue inhomogeneity, magnitude uncertainties and binning (Bender, 1983; Marzocchi et al., 2020) for the real data. These effects propagate into slower convergence to the asymptotic value for an infinite sample within the finite range observed. The ideal synthetic test data suffer from none of these complications. Further work is required to isolate the contributions to the trends in the dynamic range dependent variability shown on Figure 3.

Finally, we note that Yaghmaei-Sabegh & Ostadi-Asl (2021) have independently examined the issue of convergence in the MLE of b-value of finite samples, focusing only on the sample size n. Here, our results demonstrate that it is equally important to consider the effect of dynamic range.

6 CONCLUSION

The mean magnitude of an earthquake catalogue converges to an asymptotic limit from below in random samples, consistent with a hypothesis derived analytically from the expected modal values in the sampled mean magnitude in such samples. In real data, the trend is broadly similar, but in detail substantially more variable. In this case, the true b-value is unknown, the catalogue magnitudes have a finite error, and the catalogue is likely to be heterogeneous, explaining many of the uncertainties involved in fitting equations 15 and 17.

The dominant factor controlling the bias of high b-values is the convergence of the mean magnitude with respect to dynamic range, where the mode of the data matches the maximum likelihood expected in a random sample, when enough data and dynamic range are observed. In samples with a small dynamic range, the estimated mean magnitude is systematically and significantly underestimated, hence leading to an overestimate of the b-value as dynamic range reduces. While the bias is smaller than the scatter in the data, it is still likely to produce systematically high b-values on average from small catalogues with narrow dynamic range. Furthermore, we have shown that a stable estimate of $\bar{m}$ is a prerequisite for obtaining a stable b-value estimate.

Overall, this novel analysis indicated that many published studies in the literature use dynamic ranges where we would expect significant bias in the b-value, resulting in a systematic error (underestimation) in the likelihood of large events in a larger sample, and hence an underestimation of the associated hazard obtained by extrapolation to larger event sizes, as may be the case when using equation 19 on both GR or MGR data in the synthetic cases. This study highlights once again the importance of having enough data in any study involving power law scaling of a physical source size, and hence an exponential distribution in a logarithmic magnitude parameter.

It is prudent to adopt a cautious interpretation of b-values and their importance and significance in seismic hazard analysis. There is no 'one size fits all' answer to how many events or what dynamic range is required in specific catalogues for an accurate or precise estimate of b. Nonetheless, the b-value estimates become more accurate and precise as the sample size and its dynamic range increases. These conclusions are not restricted to the applications in earthquake hazard, because a range of natural hazards exhibit power-law frequency-size distributions as described in the in-

troduction, resulting in an exponential frequency-magnitude relation. Hence, the same issues of convergence, accuracy and precision in finite samples will apply and should be taken into account.

ACKNOWLEDGMENTS

We would like to thank two anonymous reviewers and K. Bayliss for their constructive comments on our manuscript, helping to improve clarity. We would like to acknowledge the support for this work by the NERC E3 Doctoral Training Partnership grant NE/L002558/1. No conflicts of interest exist. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

DATA AVAILABILITY

The Geysers catalogue was downloaded from the United States Geological Survey (USGS, https://earthquake.usgs.gov/earthquakes/search/ last accessed December 2020) and the Southern California data was accessed through the Southern California Earthquake Data Center (SCEDC, 2013, https://service.scedc.caltech.edu/eq-catalogs/date_mag_loc.php).

REFERENCES

Aki, K., 1965. Maximum likelihood estimate of b in the formula logn=a-bm and its confidence limits, *Bulletin of the Earthquake Research Institute*, **43**, 237–239.

Bell, A. F., Naylor, M., & Main, I. G., 2013. Convergence of the frequency-size distribution of global earthquakes, *Geophysical Research Letters*, **40**, 2585–2589.

Bender, B., 1983. Maximum likelihood estimation of b values for magnitude grouped data, *Bulletin of the Seismological Society of America*, **73**, 831–851.

Cao, A. & Gao, S. S., 2002. Temporal variation of seismic b -values beneath northeastern japan island arc, *Geophysical Research Letters*, **29**, 1–3.

Convertito, V., Maercklin, N., Sharma, N., & Zollo, A., 2012. From induced seismicity to direct time-dependent seismic hazard, *Bulletin of the Seismological Society of America*, **102**, 2563–2573.

Frohlich, C. & Davis, S. D., 1993. Teleseismic b values; or, much ado about 1.0, *Journal of Geophysical Research*, **98**, 631–644.

Gardner, J. K. & Knopoff, L., 1974. Is the sequence of earthquakes in southern california, with aftershocks removed, poissonian?, *Bulletin of the Seismological Society of America*, **64**, 1363–1367.

Geffers, G.-M., Main, I. G., & Naylor, M., 2022. Biases in estimating b-values from small earthquake catalogues: How high are high b-values?, *Geophysical Journal International*, **229**, 1840–1855.

Gerstenberger, M. C., Marzocchi, W., Allen, T., Pagani, M., Adams, J., Danciu, L., Field, E. H., Fujiwara, H., Luco, N., Ma, K. F., Meletti, C., & Petersen, M. D., 2020. Probabilistic seismic hazard analysis at regional and national scales: State of the art and future challenges, *Reviews of Geophysics*, **58**.

Gutenberg, B. & Richter, C. F., 1944. Frequency of earthquakes in california, *Bulletin of the Seismological Society of America*, **34**, 185–188.

Henderson, J. R., Barton, D. J., & Foulger, G. R., 1999. Fractal clustering of induced seismicity in the geysers geothermal area, california, *Geophysical Journal International*, **139**, 317–324.

Kagan, Y. Y., 1991. Seismic moment distribution, *Geophysical Journal International*, **106**, 123–134.

Kagan, Y. Y., 1999. Universality of the seismic moment-frequency relation, *Pure Applied Geophysics*, **155**, 537–573.

Kagan, Y. Y., 2002. Seismic moment distribution revisited: I. statistical results, *Geophysical Journal International*, **148**, 520–541.

Kamer, Y. & Hiemer, S., 2015. Data-driven spatial b value estimation with applications to california seismicity: To b or not to b, *Journal of Geophysical Research B: Solid Earth*.

Kwiatek, G., Martínez-Garzõn, P., Dresen, G., Bohnhoff, M., Sone, H., & Hartline, C., 2015. Effects of long-term fluid injection on induced seismicity parameters and maximum magnitude in northwestern part of the geysers geothermal field, *Journal of Geophysical Research: Solid Earth*, **120**, 7085–7101.

Leptokaropoulos, K., Staszek, M., Lasocki, S., Martinez-Garzón, P., & Kwiatek, G., 2018. Evolution of seismicity in relation to fluid injection in the north-western part of the geysers geothermal field, *Geophysical Journal International*, **212**, 1157–1166.

Main, I. & Burton, P., 1984. Information theory and the earthquake frequency-magnitude distribution, *Bulletin of the Seismological Society of America*, **74**, 1409–1426.

Marzocchi, W. & Jordan, T. H., 2017. A unified probabilistic framework for seismic hazard analysis, *Bulletin of the Seismological Society of America*, **107**, 2738–2744.

Marzocchi, W., Spassiani, I., Stallone, A., & Taroni, M., 2020. How to be fooled searching for significant variations of the b-value, *Geophysical Journal International*, **220**, 1845–1856.

Ogata, Y. & Yamashina, K., 1986. Unbiased estimate for b-value of magnitude frequency, *Journal of Physics of the Earth*, **34**, 187–194.

Turcotte, D., 1997. *Fractals and Chaos in Geology and Geophysics*, Cambridge University Press.

Wiemer, S. & Wyss, M., 2000. Minimum magnitude of completeness in earthquake catalogs: Examples from alaska, the western united states, and japan, *Bulletin of the Seismological Society of America*, **90**,

859–869.

Yaghmaei-Sabegh, S. & Ostadi-Asl, G., 2021. Bayesian estimation of b-value in gutenberg–richter relationship: a sample size reduction approach, *Natural Hazards*.