JGR Solid Earth

RESEARCH ARTICLE
10.1029/2020JB021242

A Normalized Distance Test for Co-Determining the Completeness Magnitude and *b*-Value of Earthquake Catalogs

A. M. Lombardi[1]

[1]Istituto Nazionale di Geofisica e Vulcanologia, Roma, Italy

Key Points:

- Commonly used methods provide biased estimations of the main parameters of the Gutenberg-Richter Law
- A new method is proposed, to estimate the *b*-value and the completeness magnitude, based on statistical hypothesis testing
- Unreliable methods may fail to reveal the actual *b* and completeness magnitude, bringing to wrong conclusions on *b*-value variations

Supporting Information:

- Supporting Information S1

Correspondence to:
A. M. Lombardi,
annamaria.lombardi@ingv.it

Citation:
Lombardi, A. M. (2021). A normalized distance test for co-determining the completeness magnitude and *b*-value of earthquake catalogs. *Journal of Geophysical Research: Solid Earth*, *126*, e2020JB021242. https://doi.org/10.1029/2020JB021242

Received 27 OCT 2020
Accepted 13 FEB 2021

Abstract The spatial-temporal variations of *b*-value in the Gutenberg-Richter's relation are understood as stress meter and premonitory indicator of strong events. The characterization of the Gutenberg-Richter's relation is complicated by the difficulty of identifying the completeness magnitude *Mc,* since its inaccurate estimate may lead to biased conclusions on *b*-value. Therefore, many methods have been proposed to estimate *b*-values and their changes, but no agreement has been reached on which is the most appropriate algorithm. Here a principled statistical framework is presented for discerning and quantifying variations in the Gutenberg-Richter Law on empirical data. A new approach, called Normalized Distance (ND) test, is proposed, combining maximum-likelihood estimates with a goodness-of-fit test and bootstrap resampling. The proposed method is evaluated on synthetic data and compared with the most used approaches. The main result is that the ND test provides better estimates of *Mc* and *b*-value respect to all other methods. Finally, the same algorithms applied to the most recent Italian sequence provide contradictory results, indicating that detected *b*-value fluctuations may be apparent.

Plain Language Summary The number of earthquakes in a seismic region increases about tenfold, on average, at each one-degree step of the magnitude scale. In any case, the ratio of small to large earthquakes, represented by a parameter usually called *b*-value, may show significant variations in space and time. The study of these changes has attracted great attention among scientists, for a long time now, as a way of characterizing seismic potential in different zones. In particular, transient decreases of *b*-values have been observed during seismic sequences, before strong earthquakes, opening possible prospects to short-term earthquake prediction. Despite the great effort invested, a full agreement on the choice of the most appropriate algorithm for detecting *b*-value changes has not been still reached. This study shows as the most common algorithms may fail to reveal actual *b* values variations, bringing to wrong conclusions on their predictive skill. Moreover, it defines the framework in which a rigorous *b*-value analysis must be conducted and proposes a new detection method, checked on both simulations and the most recent Italian sequence, which significantly improves the performance of the other algorithms, casting doubts on the inference made by them.

1. Introduction

The quantitative study of the earthquake magnitude distribution started with the publication of what is traditionally referred as "Gutenberg-Richter Law" (GRL; Gutenberg & Richter, 1942; Ishimoto & Iida, 1939), stating the log-linearity of the empirical complementary cumulative distribution of magnitudes. In mathematical terms, the GRL may be described by the equation

$$\log_{10}\left[N\left(M\right)\right] = a - b \cdot M \tag{1}$$

where *a* (productivity) and *b* (slope) are constants and *N*(*M*) is the number of events with magnitude equal to or above *M*. The parameter *b* represents the ratio of small to large earthquakes: a low *b*-value describes a data set with a large proportion of larger magnitudes, and vice versa.

Following the publication of the GRL, a great amount of studies was carried out, showing a great attention of scientists, which continues to this day. Many of these studies concern some technical problems about the estimation of the *b*-value, given by the rounding and the uncertainties of magnitudes, the size and the incompleteness of catalogs and the methodologies applied (Bender, 1983; Kamer & Hiemer, 2015; Marzocchi

& Sandri, 2003; Mignan, 2012, 2019; Mignan & Woessner, 2012; Shi & Bolt, 1982; Tinti & Mulargia, 1987). Other studies focused on more interpretative issues such as: the temporal and spatial variations of *b*-values; the use of *b*-value as stress meters, helping to image asperities; the changes of *b* with focal depth or tectonic frameworks; the measurement or real-time monitoring of *b*-values for hazard purposes; the *b*-value decrease in critical sub-regions prior to mainshocks (Gulia & Wiemer, 2019; Rundle et al., 2000; Schorlemmer et al., 2005; Wiemer & Wyss, 2000).

The technical and interpretative aspects of *b*-value estimations are both relevant and, of course, strictly linked. The use of *b*-values (and of its variations) to measure, describe, predict or interpret seismicity requires a great attention on technical details that, if neglected, would lead to misinterpretation and ambiguity.

By a statistical point of view, Equation 1 leads to a well-defined hypothesis, marked by H_0 in the following, that is that magnitudes follow an exponential distribution with completeness magnitude *Mc* (Aki, 1965; Utsu, 1965). Under the hypothesis H_0, a correct estimate of *Mc* is a crucial part of *b*-value estimation (Hainzl, 2016), whereas a *pseudo* log-linearity of $N(M)$ (Equation 1) might be find also on incomplete data.

Most of the scientific literature, following the publication of GRL, relies on ambiguous hypotheses. The *b*-value and its uncertainty, given a prefixed *Mc*, are commonly computed using a Maximum Likelihood assessment from an exponential distribution (marked by *ML*exp in the following; Aki, 1965; Utsu, 1965). Instead, *Mc* estimation is often done by catalog- or network-based methodologies (Mignan & Woessner, 2012), not assuming a specific probability distribution for magnitudes (Cao & Gao, 2002; Mignan et al., 2011; Schorlemmer & Woessner, 2008; Wiemer & Wyss, 2000).

A representative, although not exhaustive, list of catalog-based techniques to assess *Mc* includes the Maximum Curvature Method (MAXC; Wiemer & Wyss, 2000), the Goodness of Fit test (GF; Wiemer & Wyss, 2000), the *Mc* by *b*-value stability (MBS; Cao & Gao, 2002), the method proposed by Ogata and Katsura (1993; OK93) and its following revision, the Mc from the Entire Magnitude Range method (EMR; Woessner & Wiemer, 2005), the Median-based analysis of the segment slope (MBASS; Amorèse, 2007) and the Non Linear Index test (NLI; Tormann et al., 2014). Whereas GF, MBS, OK93, EMR and NLI are parametric methods, that is based on fitting the GRL on data, MAXC and MBASS are non-parametric techniques, based on the evaluation of changes in the slope of cumulative distribution of magnitudes (Mignan & Woessner, 2012). Only OK93 and EMR, among the parametric techniques, formalize a probability density function (PDF) for magnitudes below *Mc*.

Recently, test-based methods were proposed to fit generic power law distributions (Clauset et al., 2009; Corral et al., 2011) and applied to earthquake magnitudes (Corral & González, 2019). These approaches combine the maximum-likelihood fitting methods with goodness-of-fit tests, based on the Kolmogorov-Smirnov statistic (Gibbons & Chakraborty, 2003).

The present study seeks to clarify the scientific rationale to adopt for *Mc* and *b*-value estimation, based on statistical hypothesis test. The first part discusses and compares the most common published method, by their application to simulated magnitudes, to highlights their limits. The second part proposes a new method for the GRL estimation, fully placed in the outlined statistical framework.

2. The Mathematical Background for *b*-Value Estimation

By a strictly statistical point of view, the empirical GRL (Equation 1) leads to the well-defined hypothesis that magnitudes follow an exponential distribution, with probability density function

$$f_{\exp}(M) = b \cdot ln(10) \cdot e^{-b \cdot ln(10) \cdot (M - Mc)} \tag{2}$$

where, *Mc* is the minimum magnitude of a complete data set.

The instrumental earthquake magnitudes are usually discretized, since they do not have higher precision than the first decimal place (Bender, 1983; Marzocchi & Sandri, 2003; Tinti & Mulargia, 1987). The most widely used estimator for *b*-value is *ML*exp, the Maximum Likelihood (ML) assessment for an exponential distribution (Aki, 1965; Utsu, 1965), which for rounded magnitudes is given by

$$\tilde{b}(Mc) = \frac{1}{ln(10)\left(\bar{M} - Mc + \delta M / 2\right)} \tag{3}$$

where, $\bar{M}$ is the average of magnitudes above Mc and δM is the magnitude bin size (Bender, 1983; Marzocchi & Sandri, 2003). The associated error, $\sigma_{\tilde{b}}$, is given by (Shi & Bolt, 1982)

$$\sigma_{\tilde{b}} = ln(10) \cdot \left[\tilde{b}(Mc)\right]^2 \sqrt{\frac{\sum_{M_i \geq Mc}\left(M_i - \bar{M}\right)^2}{N \cdot (N-1)}}. \tag{4}$$

In any case, a full account of the discrete nature of rounded magnitude measurements is achieved only by replacing the continuous exponential with the discrete geometric distribution

$$f_{\text{geo}}(M_i \mid p) = p(1-p)^i \; i = 0,1,... \tag{5}$$

where $M_i = M_0 + i \cdot \delta M$, $M_0 = Mc + \delta M / 2$, and $p = 1 - \exp[b \cdot ln(10) \cdot \delta M]$ (Bender, 1983; Marzocchi & Sandri, 2003; Tinti & Mulargia, 1987). Without loss of generality, $M_0 = 0$ and $\delta M = 0.1$, in the following.

The geometric ML estimator (MLgeo) of b, $\hat{b}$, and the associated error, $\sigma_{\hat{b}}$, on N rounded magnitudes, are given by

$$\hat{b} = \frac{1}{ln(10) \cdot \delta M} ln(1 - \hat{p})$$

$$\sigma_{\hat{b}} = \frac{\hat{p}}{ln(10) \cdot \delta M \cdot \sqrt{N(1-\hat{p})}} \tag{6}$$

where $\bar{M}$ is the average of M_i and $\hat{p} = \frac{\delta M}{\bar{M} - +\delta M}$ (Bender, 1983; Marzocchi & Sandri, 2003; Tinti & Mulargia, 1987). The percent variation of $\hat{b}$, with respect to b, $\Delta b = \hat{b} / b \cdot 100$, is a function of N and represents the aleatory, irreducible, uncertainty on b-value estimation.

3. The Existing Approaches for *Mc* Estimation

The biggest challenge in fitting the GRL on a magnitude data set is to identify the range over which it holds, more than a b-value estimation. Not all earthquakes are detected from a seismic network, due to multifold reasons. Therefore, the completeness magnitude Mc may be defined as the lowest magnitude at which 100% of the earthquakes are detected (Rydelek & Sacks, 1989).

The most used procedures to define the GRL parameters require the estimation of b, by the MLexp method, as a function of ascending Mc, and the application of suitable criteria, for choosing Mc (and then b). As will be shown below, these criteria are, in large part, unjustified and not concordant with the hypothesis H_0, that, if assumed, must be coherently integrated in the whole procedure.

In this section, an exhaustive list of previously published procedures is applied to several classes of simulated data, under the hypothesis of equivalence between H_0 and the empirical GRL (Equation 1). Simulated samples, marked by D_{inc} in the following, have different size N and mimic incomplete magnitude databases. They are given by the PDF

$$g\left(M_i / p, \mu, \sigma\right) = f_{\text{geo}}\left(M_i / p\right) \cdot F_{\mu,\sigma}\left(M_i\right) \tag{7}$$

where $b = 1$ and $F_{\mu,\sigma}$ is the truncated Gaussian cumulative function, with parameters $\mu = 0.4$ and $\sigma = 0.4$ and lower bound -0.05 (Ogata & Katsura, 1993; Ringdall, 1975). The incomplete portion of a real magni-

tude database may have very different shapes (Mignan, 2012, 2019) and the model formalized by Equation 7 is just one possible example. D_{inc} samples are simulated by randomly removing a subset of data from complete datasets, hereinafter marked by D_c and simulated by a geometric distribution with $b = 1$. Specifically, a magnitude M_i is removed by the original data set if $r > F_{\mu,\sigma}(M_i)$, where r is a uniform random value. The size of D_c simulations is $N' = N/[1-F_{\mu,\sigma}(\mu+2\sigma)]$, with N varying from $5{\cdot}10^1$ to 10^4, so that the expected number of events above $\mu+2\sigma$ is equal to N. The original complete data sets D_c are retained, as a reference for following stages of this research.

Most of published algorithms to estimate *Mc* belong to one of following classes of methodologies:

- The catalog-based methods that check the validity of GRL on data (parametric catalog-based methods; Cao & Gao, 2002; Tormann et al., 2014; Wiemer & Wyss, 2000; Woessner & Wiemer, 2005) or evaluate changes in the empirical cumulative distribution (nonparametric catalog based methods; Amorèse, 2007; Wiemer & Wyss, 2000)
- The network-based methods that use the seismic network distribution and its changes to measure the detection magnitude threshold (Gomberg, 1991; Schorlemmer & Woessner, 2008)
- The test-based methods that fit a generic power law distributions in a more rigorous statistical testing framework (Clauset et al., 2009; Corral et al., 2011)
- The rate-dependent methods that estimate *Mc* basing on the expected short-term variations of earthquake rate (Hainzl, 2016)

In this research, the network-based and the nonparametric catalog-based techniques are not taken into account, since they cannot be easily carried out and discussed in the statistical testing framework, considered here as the most correct way to formulate the problem. Moreover, the method proposed by Hainzl (2016) is not discussed here, since it is focused on variations of *Mc* during the seismic sequences and requires the spatio-temporal modeling of aftershock rates.

3.1. Catalog-Based Methods

An exhaustive list of catalog-based methods is applied on D_c and D_{inc} datasets, to check the efficiency of each algorithm. Specifically, the methods involved in this investigation are the Goodness of Fit test (GF; Wiemer & Wyss, 2000), the *Mc* by *b*-value stability (MBS; Cao & Gao, 2002; Woessner & Wiemer, 2005) and the Non Linear Index test (NLI; Tormann et al., 2014). The Entire Magnitude Range method (EMR; Woessner & Wiemer, 2005) and its early version proposed by Ogata and Katsura (1993), are not included, since they involve a specific distribution for magnitudes below *Mc,* contrary to statements that propose that incomplete magnitudes may have different shapes (Mignan, 2012, 2019) and a strict modeling of them may be problematic (Kagan, 2002).

The correctness of GF, MBS and NLI procedures is checked on D_c and D_{inc} databases. Specifically, *b*-values are estimated, as a function of ascending cutoff magnitude, by the *ML*exp method ($\tilde{b}$, Equation 3), which is the estimator chosen by the authors of these methods (however, *ML*exp and *ML*geo estimations are close, for binning of $\delta M = 0.1$; Bender, 1983; Marzocchi & Sandri, 2003; Tinti & Mulargia, 1987). Therefore, *Mc* (and then *b*) is selected by applying the criteria defined by each method, which are summarized in Appendix A, for the sake of completeness and clarity.

Results for D_{inc} simulations are shown in Figures 1 and 2, whereas results for D_c simulations are shown in the Supporting Information, for a comparison (Figures S1 and S2). The main result for D_{inc} data sets is a joint underestimation of *b* (up to 60%–70% of the real value) and *Mc*, by GF and, even more, by NLI methods, also for large *N* values (Figures 1a–1d). The failure of the GF test is greater, the closer *R* value is to 90% (see Appendix A; Figure 1a), its lowest acceptable value (Wiemer & Wyss, 2000). Besides, values of *R* close to 100% are reachable only for large sample sizes *N* (Figure 1a). Both GF and NLI provide for D_c simulations *b*-values that are fully in agreement with the expected variability, but values of *Mc* well above 0, also for samples with thousands of data, due to (wrong) rejection of the GRL for lower threshold magnitudes (Figures S1a–S1d).

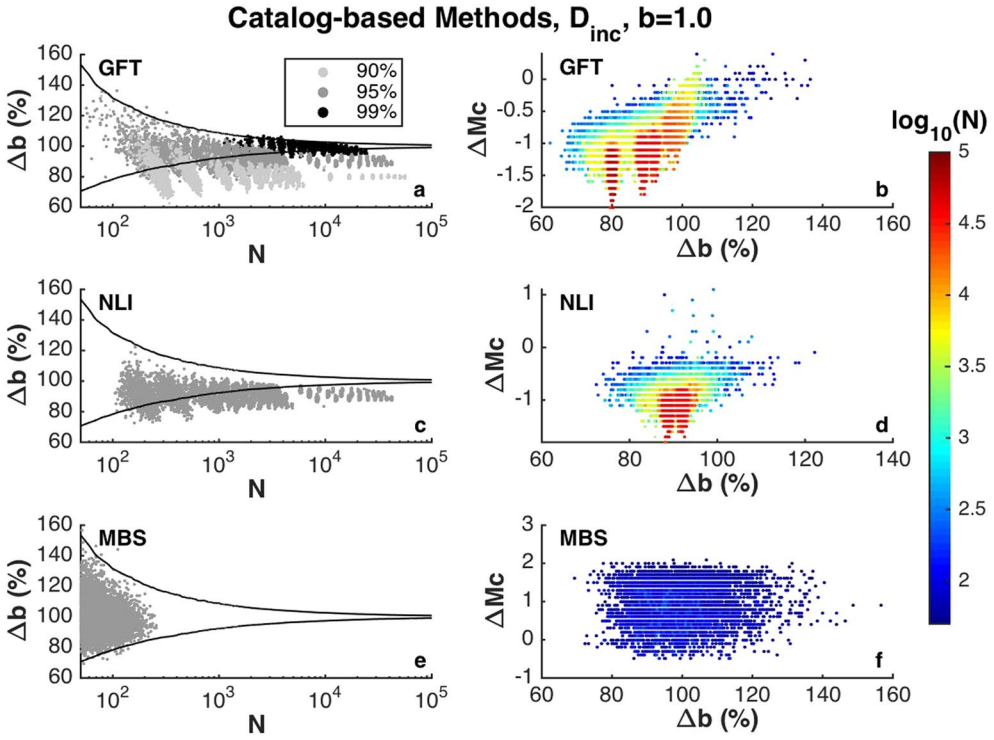

Figure 1. Results of the catalog-based tests for $b = 1.0$ and D_{inc} databases. (a) Plot of $\Delta b = \tilde{b} / b * 100$ versus N for GF test. Solid black lines mark the 99% confidence bounds of aleatory uncertainty of Δb. (b) Plot of ΔMc versus Δb, for the GF test. (c) The same as (a) but for the NLI test. (d) The same as (b) but for the NLI test. (e) The same as (a) but for the MBS method. (f) The same as (b) but for the MBS method. GF, Goodness of Fit; MBS, Mc by b-value stability; NLI, Non Linear Index test.

The MBS method provides correct b-values for both D_c and D_{inc} datasets, but strongly overestimates Mc, with a consequent strong reduction of sample size (Figures 1e and 1f and Figures S1e and S1f). This means that the b-value uncertainty is too big, with respect to Shi and Bolt's (1982) criteria, so that the b-value is

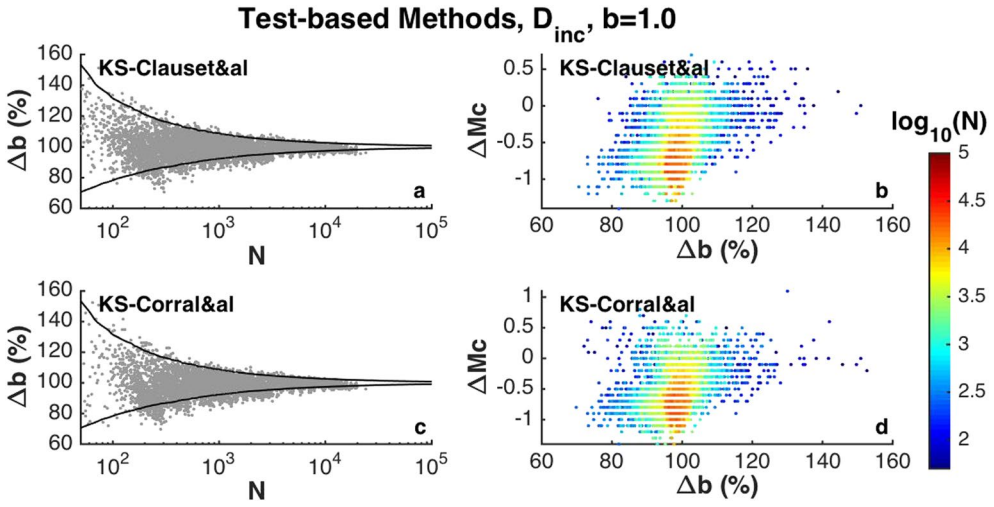

Figure 2. Results of the test-based tests for $b = 1.0$ and D_{inc} databases. (a) Plot of $\Delta b = \tilde{b}/b*100$ versus N for KS-Clauset&al test. Solid black lines mark the 99% confidence bounds of aleatory uncertainty of Δb. (b) Plot of ΔMc versus Δb, for KS-Clauset&al test. (c) The same as (a) but for the KS-Corral&al test. (d) The same as (b) but for the KS-Corral&al test.

considered unstable. In this regard, it has been recognized that the performance of the MBS method strongly depends on the choice of some details (Mignan, 2012; Woessner & Wiemer, 2005; Zhou et al., 2018). Specifically, the length of the magnitude window, ΔM, in which searching for the b-value stability (see Appendix A), has a strong influence on results, even if it is generally set to 0.5, for a magnitude binning of 0.1.

The GF and NLI tests are aimed to check the log-linearity of empirical complementary cumulative distribution of magnitudes (Equation 1). The MBS is instead based on the stability of b-values above Mc. None of these methods is fully placed in a statistical hypothesis-testing framework. A more rigorous approach, in this sense, is provided by the test-based methods.

3.2. Test-Based Methods

Under the hypothesis that the GRL is equivalent to H_0, the natural tools to check a distribution are the goodness of fit tests (Gibbons & Chakraborty, 2003). Among them, the Kolmogorov–Smirnov (KS) test quantifies the maximum absolute distance $D_{\mathrm{KS}} = \max_x \left| F_T(x) - F_E(x) \right|$, between the empirical F_E and the theoretical F_T cumulative distribution functions (Gibbons & Chakraborty, 2003). The KS test was designed for continuous distributions, but it may be easily adapted to discrete distributions (Conover, 1999) and when parameters of theoretical distribution, under testing, are estimated from data, rather than being known a priori (Lilliefors, 1969; Marzocchi et al., 2020).

The KS distance underlies both methods proposed by Clauset et al. (2009) and Corral et al. (2011), to estimate a generic power law function. A summary of these two methods is provided in Appendix B. Both of them are applied here supposing a geometric distribution for magnitudes, to formalize the problem in the most correct way.

The methods proposed by Clauset et al. (2009; hereinafter marked by KS-Clauset&al) and Corral and colleagues (Corral et al., 2011; Corral & González, 2019; marked by KS-Corral&al in the following) provided similar results. Both had a significantly better performance on D_{inc} datasets (Figure 2) with respect to the catalog-based methods, and, in particular, the b-value estimations $\hat{b}$ (Equation 6) are in much better agreement with the expected distribution. In any case, cases of underestimation of Mc (and, therefore, of b) for D_{inc} data are still more than expected (Figures 2a and 2c), especially for small N, due to weak points of both tests (see Appendix B for details).

4. The Normalized Distance (ND) Test

All methods applied in previous sections estimate b-values as a function of Mc and, then, apply a criterion (different among the methods) to select Mc. In any case, the magnitude range on which the GRL holds, and therefore the b-value, may be highly uncertain (Corral & González, 2019; Mignan, 2012). Moreover, the use of KS-distance could be a problem, due to its dependence on sample size (Gibbons & Chakraborty, 2003). In this section, a new test-based method is proposed to address both these questions.

The problem of the dependence of KS-distance on N is solved in the following way. First, 10^4 sets of synthetic geometric variables are simulated, varying b and the sample size N; then for each of them the KS statistic $D(N,b)$ is computed by refitting b. For each b-value, the statistic $D(N,b)$ turns out to be proportional to $N^{-0.5}$. Therefore, the statistic $W(b) = \sqrt{N} \cdot D(N,b)$ has the useful property to be independent on N and a theoretical cumulative distribution $F_w(w|b)$ may be computed, as function of b, but independent of N, by simulated data (see Figure S3).

The problem of the uncertainty of Mc is, instead, solved by repeating the fitting procedure on bootstrap samples S_B from the original data S (Corral & González, 2019; Woessner & Wiemer, 2005). Generally, the mean value and the standard deviation of the bootstrapped Mc determinations are adopted as suitable estimates of Mc and of its uncertainty (Woessner & Wiemer, 2005). Here the bootstrap resampling is used to assess the degree of completeness of S, as a function of ascending magnitude. In this view, the empirical cumulative distribution of bootstrapped Mc gives the probability that S is complete; therefore the most proper estimator of Mc, at a significance level α, is the (1-α)th percentile of this distribution. Then, a rigorous method to

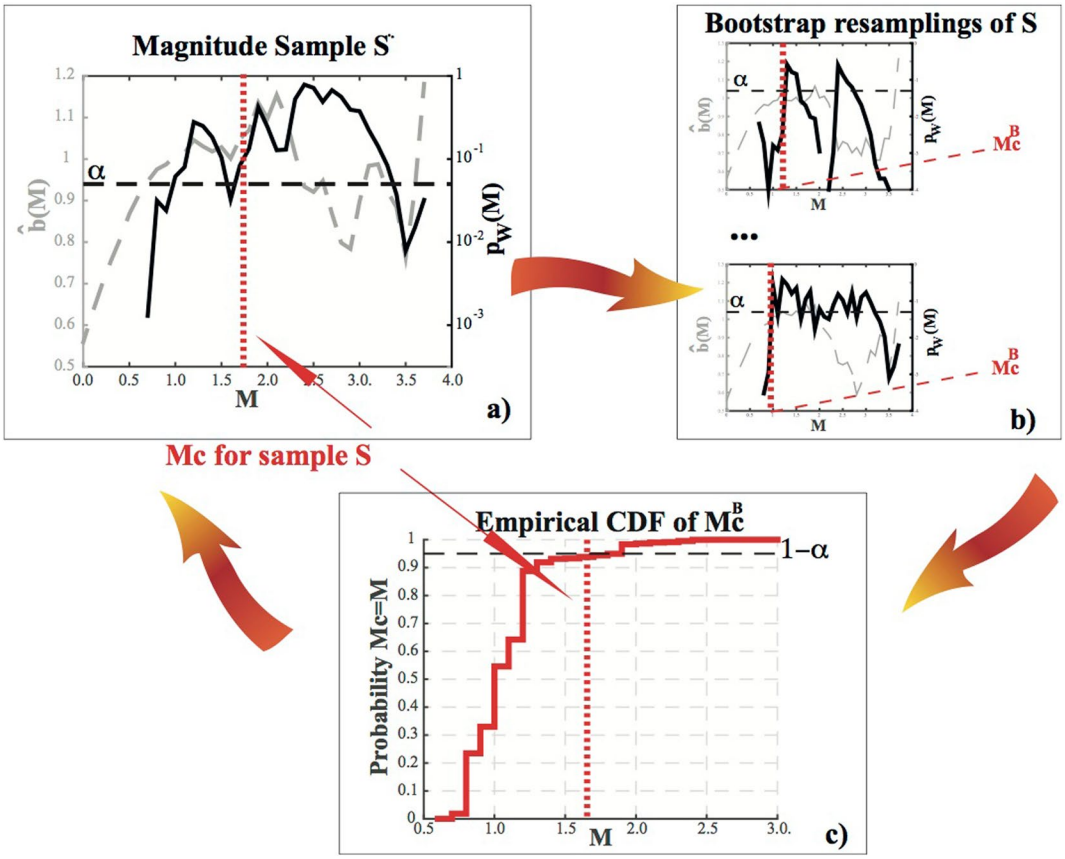

Figure 3. Overview of key steps of the ND test. (a) Estimation of b and computation of p_W, as a function of M, for the database S. (b) Estimation of M_c^B for N_B bootstrap resampling of S: for each of them, M_c^B is the lower M having $p_W(M) > \alpha$. (c) Estimation of Mc for the sample S: it is estimated as the $(1-\alpha)$% percentile of the empirical distribution of bootstrapped Mc.

estimate the GRL on a magnitude sample S, called in the following ND test, at a significance level α, can be structured as following (Figure 3):

- MLgeo estimators $\hat{b}(M)$ are computed, as function of ascending M, on the sample S; then for each M the W-statistic (W_M) and the probability $p_W(M) = 1 - F_w[W_M \mid \hat{b}(M)]$ to exceed it are computed (Figure 3a)
- The previous step is repeated for N_B bootstrap resampling $\{S_B^i,\ i = 1,...,N_B\}$ of the sample S and the completeness magnitude M_c^B is computed, for each of them, by selecting the lower M with $p_W(M)>\alpha$ (Figure 3b)
- The empirical probability distribution of M_c^B values, computed in previous step, represents the probability that S is complete above ascending M; therefore the $(1-\alpha)$*100% percentile of this distribution gives the completeness magnitude of S, Mc, with a confidence level equal to α (Figure 3c)

The ND-test is applied to D_c and D_{inc} data sets, with results shown in Figure 4, for $\alpha = 0.05$ and $N_B = 1000$. The ND test significantly improves the GRL estimation: there is a full agreement between the values $\hat{b}$and the predicted random variability, for both D_c and D_{inc} datasets (Figures 4a and 4b). Moreover, the joint underestimation of Mc and b for D_{inc} data sets disappears (Figure 4d).

The ND test was applied and is shown in the Supporting Information, for data sets simulated assuming $b = 0.5$ (with $\mu = 1.3$ and $\sigma = 0.6$) and $b = 2.0$ (with $\mu = 0.1$ and $\sigma = 0.25$), without reaching different results (Figures S4 and S5).

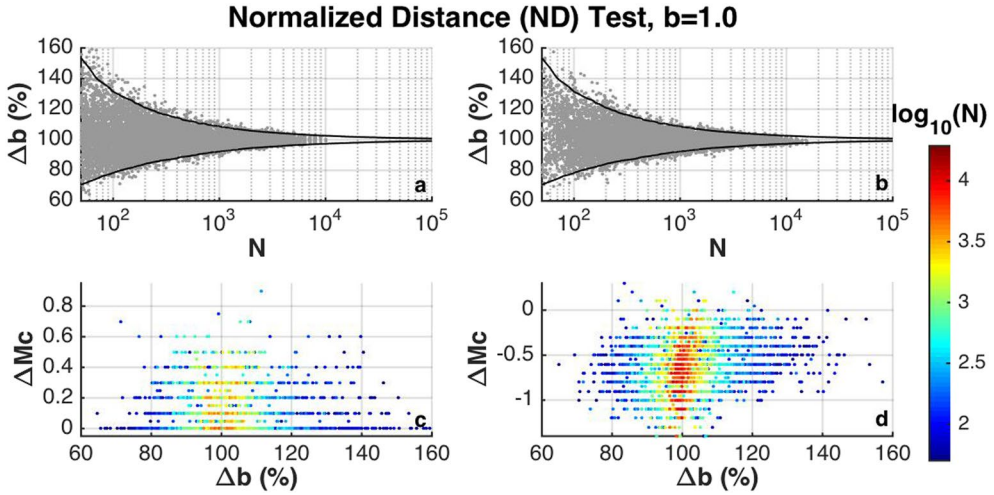

Figure 4. Results of the ND test for $b = 1.0$ and $\alpha = 0.05$, on D_c and D_{inc} databases. (a) Plot of $\Delta b = \hat{b}/b*100$ versus N for D_c samples. Solid black lines mark the 99% confidence bounds of aleatory uncertainty of Δb. (b) The same as (a) but for D_{inc} databases. (c) Plot of ΔMc versus Δb, for D_c samples. (d) The same as (c) but for D_{inc} databases.

5. Searching for *b*-Value Variations: The Central Italy Sequence (2016)

Over the last 2 decades, detailed studies have correlated *b*-value heterogeneity to depth (Gerstenberger et al., 2001), fault asperities (Wiemer & Wyss, 1997), focal mechanisms (Schorlemmer et al., 2005), high fluid flow and pore pressures in geothermal and volcanic regions (Lombardi et al., 2006; Wiemer et al., 1998). Moreover, the decrease of *b*-value before the occurrence of main-shocks or during aftershock activity is a common recognized feature of seismicity (De Gori et al., 2012; Gulia et al., 2016; Gulia & Wiemer, 2019; Nanjo et al., 2012; Schurr et al., 2014), even if it has been questioned (Hainzl, 2016; Kamer & Hiemer, 2015; Mignan, 2011). To delve deeper into this issue, an analysis of the most recent Italian sequence is presented in the following.

On October 30, 2016, the strongest Italian earthquake (M_w6.5) of last 35 years hit the town of Norcia, preceded by two months of strong seismicity, starting on August 24, 2016, with the Amatrice earthquake (M_w6.0) (Improta et al., 2019). To compute possible changes in *b*, before and after the main shocks, the shallow events (above 30 km of depth) were selected from the Italian Seismic Bulletin (BSI, Bollettino Sismico Italiano; http://terremoti.ingv.it/en/bsi) of the Istituto Nazionale di Geofisica e Vulcanologia (INGV), occurring since 2006 and with $M_L \geq 1.5$, and in a 15 and 12 km radius areas, centered on the Norcia and Amatrice shocks, respectively.

The first step of this analysis consisted in estimating a reference *b* for the seismicity that occurred from January 1, 2006 until the last event preceding the Amatrice earthquake (971 and 1425 events in the Amatrice and Norcia regions, respectively), with each of the methods discussed in the previous sections. These last identify a *Mc* varying from 1.5 to 2.4 and a reference *b* from 1.2 to 1.5. Each reference *b*-value ($\hat{b}_B$) is compared with a time series of *b*-values computed by the same approach. This is done by estimating *Mc* and *b* for windows of 500 events (T_i), moved forward by one event through the catalog. What is shown below does not critically depend on the choice of window size from 250 to 500. Figure 5 shows the percent variation of *b*-values time series ($\hat{b}_{T_i}$), with respect to the reference *b*-value $\hat{b}_B$ in the Amatrice and Norcia regions, respectively. The *Mc* time series are shown in the Supporting Information (Figure S6). All methods, except MBS and ND, identify a period of persistent low *b*-value, between the occurrence of Amatrice and Norcia main events, whereas *b*-values estimated by the MBS and ND tests remain mostly unchanged (Figure 5). In parallel, *Mc* is generally below 2.0 for the GFT, NLI, KS-Clauset&al, and KS-Corral&al methods, whereas the MBS and ND methods provide larger values (see Figure S6), confirming that MBS and ND give, on average, larger *Mc*/*b* values with respect to GF and NLI methods.

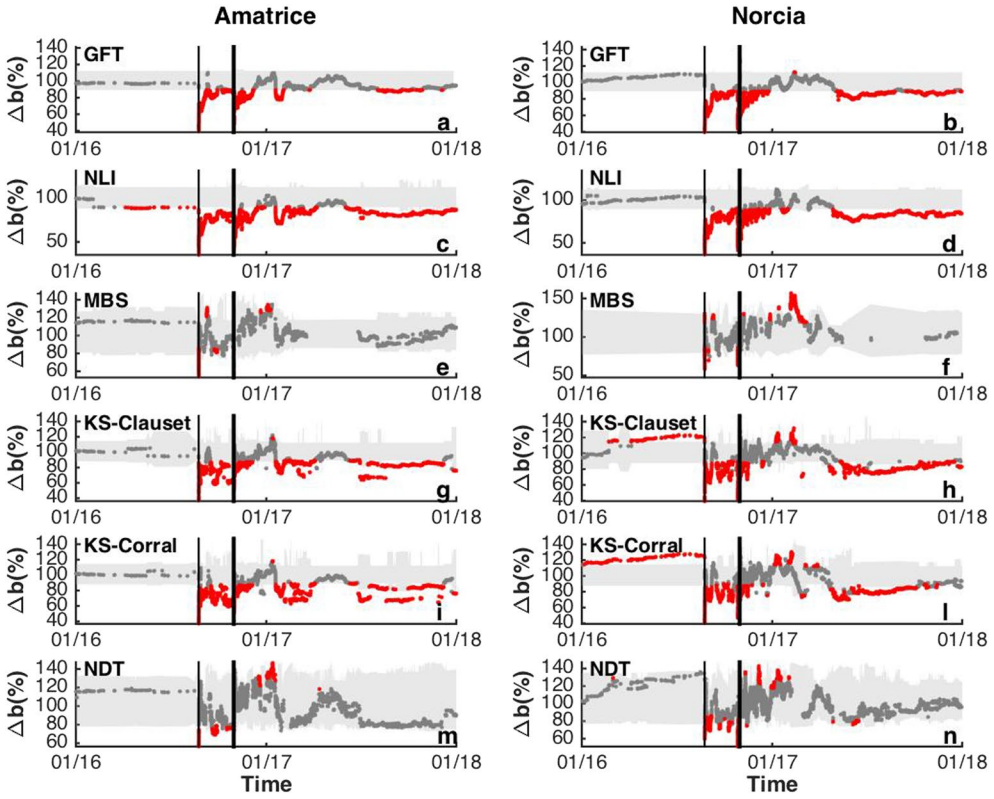

Figure 5. Time analysis of *b* values, for the source region of the Amatrice and Norcia mainshocks, by using all methods discussed in the text. The red (gray) points indicate the *b* values significantly (not significantly) different from the reference (background) *b*-value. The vertical solid gray lines represent the times of the Amatrice and Norcia earthquakes. The gray shaded areas show the 99% confidence bound of the aleatory *b*-values uncertainty.

The results, overall, show that the decrease of *b*-values between the Amatrice and Norcia mainshocks depends on the method used and, specifically, on *Mc* estimation.

In contrast with simulations, the KS-Clauset&al/KS-Corral&al methods are more in agreement with the GF and NLI methods than the ND test. This shows that the ND test is not a simple refinement of the KS-Clauset&al and KS-Corral&al methods but may provide a significant improvement to GRL estimation.

6. Discussion

The present study has been undertaken with the main objectives (1) to discuss, from a statistical point of view, an exhaustive list of previously published methods, that estimate the GRL, and (2) to propose a new method, fully placed in a statistical testing framework.

The first important result of present study is that most of algorithms, commonly used to infer the GRL, are misleading, since they are not formulated in terms of a rigorous statistical test, which gives quantitative statements both on the reliability of a reference hypothesis and on the probabilities of wrong judgments (Casella & Berger, 2001; Gibbons & Chakraborty, 2003). The GF (Wiemer & Wyss, 2000) and NLI (Tormann et al., 2014) methods provide a strong joint underestimation of *Mc* and *b*, no matter how large a sample is (Figures 1a–1d), whereas the MBS (Woessner & Wiemer, 2005) method may strongly overestimate *Mc* (Figures 1e and 1f). These inefficiencies of catalog and tested-based models examined here may be ascribed to several reasons.

- First, the GF, NLI, and MBS methods do not clearly define the reference hypothesis to the test, which should be a precise statement about data, without doubts on its definition. The GF and NLI methods check the empirical GRL (Equation 1) but adopt the *ML*exp method to estimate *b*, implying the

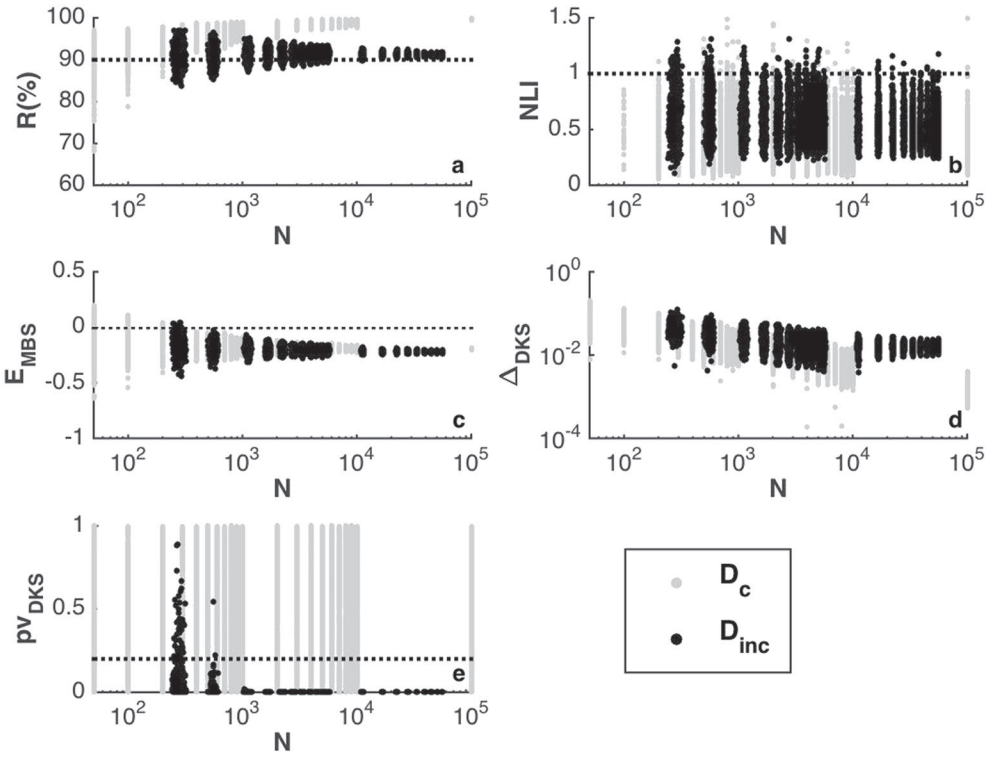

Figure 6. Plot of the key parameters, used by previously published methods to select *Mc*, versus the sample size *N*, for simulated complete D_c (magnitudes above $M_T = 0$) and incomplete D_{inc} (magnitudes above $M_T = \mu = 0.4$) databases. (a) Plot of *R* statistic (Equation A1) of GF test; (b) Plot of NLI statistic of NLI test (see Appendix A); (c) plot of E_{MBS} variable (see Appendix A) for the MBS tests; (d) plot of the Δ_{DKS} variable for the KS-Clauset&al test (see Appendix B and the main text for details); (e) plot of probabilities p_{DKS} for the KS-Corral&al test (see Appendix B). MBS, *Mc* by *b*-value stability; NLI, Non Linear Index test.

equivalence between the empirical GRL and H_0, in facts. The MBS method checks a general "stability" of *b*-values, without any further specification of a reference hypothesis to test

- Second, any statistical test requires the definition of the acceptance/rejection regions. These are the complementary intervals in which the test statistic, that is a representative summary of data, is consistent/inconsistent with the null hypothesis. The acceptance regions of catalog-based tests are: (a) [90% 100%], for the statistic *R* of GF test; (b) [0 1], for the NLI statistic of NLI test, and (c) (0 ∞) for the E_{MBS} statistic of MBS test (see Appendix A). However, no quantitative explanation for these choices is given and values of *R,* NLI, and E_{MBS} are not indicative of the reliability of GRL/H_0. Similarly, the choice of minimizing the KS distance, chosen by the KS-Clauset&al method, as well as the significance level equaling 0.2, chosen by the KS-Corral&al method, are somewhat arbitrary. Under the hypothesis H_0, none of these definitions are justified, as shown in Figure 6. The statistic *R* of GF method is partially lower than 90%, for complete D_c datasets, above $M_T = 0$ and $N < 200$, and systematically above 90%, for large incomplete D_{inc} catalogs, above $M_T = \mu = 0.4$ (Figure 6a; see also Woessner & Wiemer, 2005). Similarly, and even worse, NLI values may be well lower than one for incomplete large D_{inc} samples (Figure 6b), and E_{MBS} is systematically negative for D_c data sets, with $N > 500$ (Figure 6c). The variable $\Delta_{DKS} = D_{KS}(M_T) - \text{argmin}_{M_i \geq M_T} D_{KS}(M_i)$ is positive for both D_c and D_{inc} samples (Figure 6d), showing that the criterion used by Clauset et al. (2009) to select *Mc* was unfounded. Finally, the p-values p_{DSK} of observing $D_{KS}(M_T)$ span the whole probability range from 0 to 1 (Figure 6e), also for D_c samples, showing that the threshold of 0.2, fixed by Corral and colleagues (Corral & González, 2019), is unjustified
- Finally, none among GF, NLI, MBS, KS-Clauset&al, or KS-Corral&al tests allowed choosing and varying a significant level (i.e., the probability of make errors in judgments), as should be done in any statistical test (Casella & Berger, 2001; Gibbons & Chakraborty, 2003)

The importance of placing the GRL estimation question into a statistical testing framework is proven by the fact that the KS-Clauset&al and KS-Corral&al methods (Clauset et al., 2009; Corral et al., 2011) significantly improve the *Mc* and *b* estimation on simulations. The new method proposed here, the ND test, follows what done by Clauset et al. (2009) and Corral et al. (2011), but overcomes their limitations, placing the problem of *Mc* evaluation into a full statistical testing framework, without arbitrariness. Bootstrapping is not used here as a way of characterizing *Mc* uncertainty (Corral & González, 2019; Woessner & Wiemer, 2005), but as a measure of reliability in assessing *Mc*: in this view, the chosen value of *Mc* is requested to reach the $(1-\alpha)\%$ confidence level of the empirical distribution of bootstrapped *Mc*, where α is the pre-fixed significance level of the test. This formulation of the problem seems to be successful on both simulated and real data. Among the methods compared here, only the ND test provides results for simulations fully in agreement with the aleatory uncertainties.

The example of Central Italy sequence shows that the ND test may provide results which are significantly different from both catalog and test-based methods, proving that the significance of *b*-value variations depends on the suitability of adopted method. This result opens a new perspective on the debate about the significance of *b*-value variations and, most of all, on the potential of *b*-value variations as an earthquake precursor (Brodsky, 2019; Dascher-Cousineau et al., 2019; Gulia & Wiemer, 2019; Helmstetter et al, 2003; Hiemer & Kamer, 2015; Mignan, 2014; Wiemer & Wyss, 1997).

7. Conclusions

The following main conclusions can be drawn from the present study:

- Statistical testing is the most widespread, natural and rigorous strategy for dealing with issue of *b*-value and *Mc* estimations and the assessment of their variations
- The published methods discussed here were not fully placed in a statistical test context and, therefore, their technical and predictive skills remain controversial
- The ND test was fully placed into a statistical testing framework and improve the performance of previously published methods on both simulated and real data

Appendix A: The Catalog-Based Methods

The Goodness of Fit (GF) test (Wiemer & Wyss, 2000) is evaluated by a parameter *R*, which is the absolute difference of observed (O_i) and expected (E_i) numbers of events above each magnitude bin

$$R\left(Mc\right) = 100 - \frac{\sum_{M_i \geq Mc} \left|O_i - E_i\right|}{\sum_{M_i \geq Mc} O_i} \cdot 100. \tag{A1}$$

The expected rates E_i are given by the empirical Gutenberg-Richter Law (Equation 1) and the *ML*exp *b*-value estimator $\tilde{b}$ (Equation 3). The completeness magnitude *Mc* is estimated as the first magnitude cutoff above which the observed data have a log-linear behavior, that is, where *R* reaches a prefixed threshold of 90% or 95%.

The Non Linear Index (NLI) test (Tormann et al., 2014) judges the linearity of a sample based on *ML*exp *b*-value estimates, $\tilde{b}$. It starts at a magnitude defined by the Maximum Curvature (MAXC) method (Wiemer & Wyss, 2000) and increases up to the highest magnitude above which at least 50 events are still observed. If five or more *b*-value estimates can be calculated by this definition, NLI is computed as the ratio of the standard deviation of these *b*-value estimates divided by the largest individual uncertainty in the single *b*-value estimates ($\sigma_{\tilde{b}}$, Equation 4; Shi & Bolt, 1982). *Mc* is the lower magnitude cutoff for which NLI ≤ 1.

The Mc by *b*-value stability (MBS) method (Cao & Gao, 2002; Woessner & Wiemer, 2005) is based on the assumption that *b*-value estimates ascend for cutoff magnitudes smaller than *Mc* and remain constant for larger magnitude thresholds. In the original version of the MBS method (Cao & Gao, 2002), *Mc* is defined as the magnitude for which the change in *ML*exp *b*-value estimates $\tilde{b}$ (Equation 3), between two successive magnitude bins, is smaller than 0.03. Later, Woessner and Wiemer (2005) found this criterion to be unstable and proposed a new criterion, based on b-value uncertainty $\sigma_{\tilde{b}}$ (Equation 4; Shi & Bolt; 1982). For a

magnitude binning $\delta M = 0.1$ and a window length $\Delta M = 0.5$, Mc is, then, defined as the first magnitude at which $E_{\mathrm{MBS}} = \sigma_{\tilde{b}} - |b_{\mathrm{ave}} - \tilde{b}(Mc)| \geq 0$, where $b_{\mathrm{ave}} = \frac{\delta M}{\Delta M} \sum_{M=Mc}^{Mc+\Delta M} \tilde{b}(M)$.

Appendix B: The Test-Based Methods

The Clauset et al. (2009) method consists of estimating the exponent of a power law as a function of cutoff, and then, in choosing the couple of parameters that minimize the KS distance, D_{min}. Finally, a p-value, that is the probability that power-law data have a minimum KS distance larger than D_{min}, is computed, to test the inferred distribution on data. This probability is computed by applying the same fitting procedure, as the one applied on empirical data, to Monte Carlo simulations (see Clauset et al., 2009 for details).

In the particular case of the GRL inference and a geometric distribution (Equation 5) for magnitudes above Mc, the KS distance is given by

$$D_{\mathrm{KS}}(Mc) = \max_{M_j \geq Mc} \left[\left| F_{\mathrm{geo}}(M_j \mid \hat{b}) - F_E(M_j) \right| \right] \tag{B1}$$

where F_{geo} and F_E are the geometric and empirical cumulative distribution functions for magnitudes.

The Clauset et al. (2009) method estimates Mc (and therefore b) as the magnitude value that minimizes D_{KS}:

$$Mc = \underset{i}{\operatorname{argmin}}\, D_{\mathrm{KS}}(M_i) = \underset{i}{\operatorname{argmin}} \max_{j \geq i} \left[\left| F_{\mathrm{geo}}(M_j \mid \hat{b}) - \tilde{F}(M_j) \right| \right] \tag{B2}$$

A drawback of the Clauset et al. (2009) method is that the authors do not provide any explanation for why this should work, as recognized by Corral and colleagues (Corral et al, 2011; Corral & González, 2019; Deluca & Corral, 2013). Moreover, this method does not take into account the sample size, on which the KS distance depends, to compute the p-value of the test (Deluca & Corral, 2013; Gibbons & Chakraborty, 2003). To overcome these limitations, Corral and his colleagues developed an alternative method to estimate power law functions, still based on the KS distance (Corral et al, 2011; Corral & González, 2019; Deluca & Corral, 2013). Mc is estimated as the lower value for which the probability p_{DSK} to observe the computed KS distance is almost equal to 0.2 (Corral & González, 2019). The theoretical distribution of the KS distance is obtained by applying all steps (fit of b and calculation of KS distance with the new value of b) on simulated samples, with the same size as the empirical sample.

Data Availability Statement

Italian earthquake data were made available by the Bollettino Sismico Italiano (INGV) at http://terremoti.ingv.it/en/bsi

Acknowledgments
The author is grateful to the associate editor and reviewers for useful comments, suggestions, and references that significantly improved the manuscript.

References

Aki, K. (1965). Maximum likelihood estimate of b in the formula log N = a - bM and its confidence limits. *Bulletin of the Earthquake Research Institute of the University of Tokyo*, *43*, 237–239.
Amorèse, D. (2007). Applying a change-point detection method on frequency-magnitude distributions. *Bulletin of the Seismological Society of America*, *97*(5), 1742–1749.
Bender, B. (1983). Maximum likelihood estimation of b values for magnitude grouped data. *Bulletin of the Seismological Society of America*, *73*(3), 831–851.
Brodsky, E. E. (2019). Determining whether the worst earthquake has passed. *Nature*, *574*, 185–186. https://doi.org/10.1038/d41586-019-02972-z
Cao, A. M., & Gao, S. S. (2002). Temporal variations of seismic b-values beneath northeastern japan island arc. *Geophysical Research Letters*, *29*(9), 48-1–48-3. https://doi.org/10.1029/2001GL013775
Casella, G., & Berger, R. L. (2001). *Statistical inference* (p. 688). Duxbery Press.
Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661–703.
Conover, W. J. (Ed.) (1999). *Practical nonparametric statistic* (3rd ed., p. 608). New York, NY: John Wiley & Sons.
Corral, A., Font, F., & Camacho, J. (2011). Non-characteristic half-lives in radioactive decay. *Physical Review E*, *83*, 66103.
Corral, A., & González, A. (2019). Power law distributions in geoscience revisited. *Earth and Space Science*, *6*, 673–697.

Dascher-Cousineau, K., Lay, T., & Brodsky, E. E. (2019). Two foreshock sequences Post Gulia and Wiemer (2019). *Seismological Research Letters*, *91*(5), 2843–2850. https://doi.org/10.1785/0220200082

De Gori, P., Lucente, F. P., Lombardi, A. M., Chiarabba, C., & Montuori, C. (2012). Heterogeneities along the 2009 L'Aquila normal fault inferred by the b-value distribution. *Geophysical Research Letters*, *39*. https://doi.org/10.1029/2012GL052822

Deluca, A., & Corral, A. (2013). Fitting and goodness-of-fit test of non-truncated and truncated power law distributions. *Acta Geophysica*, *61*, 1351–1394.

Gerstenberger, M., Wieme, r S., & Giardini, D. (2001). A systematic test of the hypothesis that the b value varies with depth in California. *Geophysical Research Letters*, *28*, 57–60. https://doi.org/10.1029/2000GL012026

Gibbons, J. D., & Chakraborty, S. (2003). Nonparametric statistical inference (4th ed., Vol. 36). New York, NY: Marcel Dekker Inc.

Gomberg, J. (1991). Seismicity and detection/location threshold in the southern great basin seismic network. *Journal of Geophysical Research*, *96*, 16401–16414.

Gulia, L., Tormann, T., Wiemer, S., Herrmann, M., & Seif, S. (2016). Short-term probabilistic earthquake risk assessment considering time-dependent b values. *Geophysical Research Letters*, *43*, 1100–1108. https://doi.org/10.1002/2015GL066686

Gulia, L., & Wiemer, S. (2019). Real-time discrimination of earthquake foreshocks and aftershocks. *Nature*, *574*(7777), 193–199.

Gutenberg, B., & Richter, C. F. (1942). Earthquake magnitude, intensity, energy, and acceleration. *Bulletin of the Seismological Society of America*, *32*(3), 163–191.

Hainzl, S. (2016). Rate-dependent incompleteness of earthquake catalogs. *Seismological Research Letters*, *87*(2A), 337–344. https://doi.org/10.1785/0220150211

Helmstetter, A., Sornette, D., & Grasso, J. R. (2003). Mainshocks are aftershocks of conditional foreshocks: How do foreshock statistical properties emerge from aftershock laws. *Journal of Geophysical Research*, *108*(B1). https://doi.org/10.1029/2002JB001991

Hiemer, S., & Kamer, Y. (2015). Improved seismicity forecast with spatially varying magnitude distribution. *Seismological Research Letters*, *87*(2A), 327–336.

Kagan, Y. Y. (2002). Seismic moment distribution revisited: I. Statistical results. *Geophysical Journal International*, *148*(3), 520–541.

Kamer, Y., & Hiemer, S. (2015). Datadriven spatial b value estimation with applications to California seismicity: To b or not to be. *Journal of Geophysical Research: Solid Earth*, *120*. https://doi.org/10.1002/2014JB011510

Improta, L., Latorre, D., Margheriti, L., Nardi, A., Marchetti, A., Lombardi, A. M., et al. (2019). Multi-segment rupture of the 2016 Amatrice-Visso-Norcia seismic sequence (central Italy) constrained by the first high-quality catalog of early aftershocks. *Scientific Reports*, *9*, 6921.

Ishimoto, M., & Iida, K. (1939). Observations of earthquakes registered with the microseismograph constructed recently. *Bulletin of the Earthquake Research Institute, University of Tokyo*, *17*, 443–478.

Lilliefors, H. (1969). On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown. *Journal of the American Statistical Association*, *64*, 387–389.

Lombardi, A. M., Marzocchi, W., & Selva, J. (2006). Exploring the evolution of a volcanic seismic swarm: The case of the 2000 Izu Islands swarm. *Geophysical Research Letters*, *33*, 7310. https://doi.org/10.1029/2005GL025157

Marzocchi, W., & Sandri, L. (2003). A review and new insights on the estimation of the b-value and its uncertainty. *Annals of Geophysics*, *46*, 1271–1282.

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*(3), 1845–1856.

Mignan, A. (2011). Retrospective on the Accelerating Seismic Release (ASR) hypothesis: Controversy and new horizons. *Tectonophysics*, *505*. https://doi.org/10.1016/j.tecto.2011.03.010

Mignan, A. (2012). Functional shape of the earthquake frequency-magnitude distribution and completeness magnitude. *Journal of Geophysical Research*, *117*, B08302. https://doi.org/10.1029/2012JB009347

Mignan, A. (2014). The debate on the prognostic value of earthquake foreshocks: A meta-analysis. *Scientific Reports*, *4*, 4099.

Mignan, A. (2019). Generalized earthquake frequency magnitude distribution described by asymmetric Laplace mixture modeling. *Geophysical Journal International*, *219*. https://doi.org/10.1093/gji/ggz373

Mignan, A., Werner, M. J., Wiemer, S., Chen, C. C., & Wu, Y. M. (2011). Bayesian estimation of the spatially varying completeness magnitude of earthquake catalogs. *Bulletin of the Seismological Society of America*, *101*. https://doi.org/10.1785/0120100223

Mignan, A., & Woessner, J. (2012). *Estimating the magnitude of completeness for earthquake catalogs, community online resource for statistical seismicity analysis*. https://doi.org/10.5078/corssa-00180805 Retrieved from http://www.corssa.org

Nanjo, K. Z., Hirata, N., Obara, K., & Kasahara, K. (2012). Decade-scale decrease in b value prior to the M9-class 2011 Tohoku and 2004 Sumatra quakes. *Geophysical Research Letters*, *39*, L20304. https://doi.org/10.1029/2012GL052997

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

Ringdal, F. (1975). On the estimation of seismic detection thresholds. *Bulletin of the Seismological Society of America*, *65*, 1631–1642.

Rundle, J. B., Klein, W., Turcotte, D. L., & Malamud, B. D. (2000). Precursory seismic activation and criticalpoint phenomena. *Pure and Applied Geophysics*, *157*, 2165–2182.

Rydelek, P. A., & Sacks, I. S. (1989). Testing the completeness of earthquake catalogs and the hypothesis of self-similarity. *Nature*, *337*, 251–253.

Schorlemmer, D., Wiemer, S., & Wyss, M. (2005). Variation in earthquake-size distribution across different stress regimes. *Nature*, *437*, 539–542.

Schorlemmer, D., & Woessner, J. (2008). Probability of detecting an earthquake. *Bulletin of the Seismological Society of America*, *98*. https://doi.org/10.1785/0120070105

Schurr, B., Gunter, A., Hainz, S., Bedford, J., Hoechner, A., Palo, M., et al. (2014). Gradual unlocking of plate boundary controlled initiation of the 2014 Iquique earthquake. *Nature*, *512*, 299–302. https://doi.org/10.1038/nature13681

Shi, Y., & Bolt, B. (1982). The standard error of the magnitude-frequency b value. *Bulletin of the Seismological Society of America*, *72*(5), 1677–1687.

Tinti, S., & Mulargia, F. (1987). Confidence intervals of b-values for grouped magnitudes. *Bulletin of the Seismological Society of America*, *77*, 2125–2134.

Tormann, T., Wiemer, S., & Mignan, A. (2014). Systematic survey of high-resolution b value imaging along Californian faults: Inference on asperities. *Journal of Geophysical Research: Solid Earth*, *119*, 2029–2054. https://doi.org/10.1002/2013JB010867

Utsu, T. (1965). A method for determining the value of b in formula log N = a - bM showing the magnitude-frequency relation for earthquakes. *Geophysical Bulletin of Hokkaido University*, *13*, 99–103.

Wiemer, S., McNutt, S., & Wyss, M. (1998). Temporal and three-dimensional spatial analyses of the frequency-magnitude distribution near Long Valley Caldera, California. *Geophysical Journal International*, *134*, 409–421. https://doi.org/10.1046/j.1365-246x.1998.00561.x

Wiemer, S., & Wyss, M. (1997). Mapping the frequency-magnitude distribution in asperities: An improved technique to calculate recurrence times? *Journal of Geophysical Research*, *102*, 15115–15128.

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*(4), 859–869.

Woessner, J., & Wiemer, S. (2005). Assessing the quality of earthquake catalogues: Estimating the magnitude of completeness and its uncertainty. *Bulletin of the Seismological Society of America*, *95*, 684–698. https://doi.org/10.1785/012040007

Zhou, Y., Zhou, Y., & Zhuang, J. (2018). A test on methods for MC estimation based on earthquake catalog. *Earth and Planetary Physics*, *2*, 150–162. https://doi.org/10.26464/epp2018015