

A statistical framework for detection of b -value anomalies in Italy

Anna Maria Lombardi

Istituto Nazionale di Geofisica e Vulcanologia, Italy. E-mail: annamaria.lombardi@ingv.it

Accepted 2024 February 22. Received 2024 February 8; in original form 2023 October 20

SUMMARY

This study presents a new robust statistical framework, in which to measure relative differences, or deviations from a hypothetical reference value, of Gutenberg–Richter b -value. Moreover, it applies this method to recent seismicity in Italy, to find possible changes of earthquake magnitude distribution in time and space. The method uses bootstrap techniques, which have no prior assumptions about the distribution of data, keeping their basic features. Excluding Central Italy, no significative b -value variation is found, revealing that the frequency–magnitude distribution exponent is substantially stable or that data are not able to reveal hidden variations. Considering the small size of examined magnitude samples, we cannot definitively decide if the higher b -values in Central Italy, consistently founded by all applied tests, have a physical origin or result from a statistical bias. In any case, they indicate short-lived excursions which have a temporary nature and, therefore, cannot be associated solely to spatial variations in tectonic framework. Both the methodological issues and the results of the application to seismicity in Italy show that a correct assessing of b -value changes requests appropriate statistics, that accurately quantify the low accuracy and precision of b -value estimation for small magnitude samples.

Key words: Monte Carlo methods; Statistical methods; Computational seismology; Statistical seismology.

1 INTRODUCTION

The b -value, meant as a measure of the relative amount of large to small earthquakes, turns out to be, on average, equal to 1.0, at regional scale (Frohlich & Davis 1993; Kagan & Jackson 2000). Anyway, peculiar seismicity, such as volcanic or induced earthquakes (Wiemer *et al.* 1998; Henderson *et al.* 1999) or specific tectonic regimes (Schorlemmer *et al.* 2005), seems to show departures from the universal value, indicating that the spatial and temporal heterogeneity of b -values is an important clue for forecasting potentially large earthquakes and for seismic hazard assessment (Wyss *et al.* 2000; Wiemer & Wyss 1997; Schorlemmer & Wiemer 2005).

On the other hand, some research studies, more focused on statistical proprieties of earthquakes magnitude distribution, show that the apparent variability in b -values may be misunderstood. For example, Amorèse *et al.* (2010) show that the hypothesis of variation in b -values in southern California to the depth of the crust was rejected on the grounds of not being statistically significant. Kamer & Hiemer (2015) show that b -values in some tectonically different locations of California are distributed within a very limited range and that the larger variability, reported in previous studies, is mainly due to subjective arbitrariness. These and many other studies draw attention on non physical factors of b -value differences: network-related discrepancies in the magnitude computation or earthquake detection, bias introduced by binning magnitudes or departures from the Gutenberg–Richter Law, influence of specific

properties of magnitude samples, such as low size and dynamic range (Kamer & Hiemer 2015; Marzocchi *et al.* 2020, Herrmann & Marzocchi 2021; Geffers *et al.* 2022, 2023). All of them conclude that particular care is, then, required when we face with b -values estimation and we judge their variations. Specifically, particular attention must be paid in evaluating completeness magnitude, M_c , that may significantly affect b -value estimation (Roberts 2015; Marzocchi *et al.* 2020; Geffers *et al.* 2022). More generally, the analysis of spatial and temporal heterogeneity of b -values requires statistically robust methodologies and a rigorous testing of possible differences (Kagan 1999).

Based on above two viewpoints, for and against (or more cautious, at least, on) b -value variations, the calculation reliability for research on the spatiotemporal heterogeneity of b -values still needs to be solved, and the significance of b -value computations for hazard assessment and forecasting capabilities needs to be further investigated.

In this study, we deal with two points. First, we discuss how the low accuracy and precision of b -value estimate, for small sample sizes and magnitude ranges (Marzocchi *et al.* 2020; Geffers *et al.* 2022, 2023), affect the statistical measures of b -value variation and we develop a robust statistical framework in which to test b -value differences. Secondly, we apply this statistical approach to investigate the temporal and spatial heterogeneity of b -value, in space–time windows that are not associated with seismic sequences, in Italy. The influence of tectonic regimes on the earthquake size distribution in

Italy has been analysed by Gulia & Wiemer (2010) and Gulia *et al.* (2010). By building a seismotectonic zonation, consisting of large-scale tectonic zones, they show that reverse faulting is associated with lower b -values than those associated with strike-slip and normal faulting, although no statistical test was performed to assess the significance of the reported b -value differences. Later, Taroni *et al.* (2021) investigated the spatial variability of the magnitude frequency distribution in Italy, by a weighted maximum-likelihood approach, that has the advantage of gradually decreasing the importance of observations with distance, under the hypothesis of spatial continuous changes of b -value. The significance of detected spatial variability of b -values was, then, tested with respect to a spatially uniform b -value, by a Bayesian approach.

Following the above mentioned studies, we perform a spatio-temporal scanning of b -value in Italy, in light of recent recommendations to ensure statistical rigour (Marzocchi *et al.* 2020; Geffers *et al.* 2022 and references therein). Particular care is devoted to test the significance of possible different b -values and to avoid misunderstanding caused by poor statistical approaches. The final goal of this paper is to create a statistical framework in which it is possible to test the significance of possible b -value variations.

2 SEISMIC DATA

In this study, we analyse the data collected in the Italian Seismic Bulletin (BSI, Bollettino Sismico Italiano; <http://terremoti.ingv.it/en/inside>; <http://terremoti.ingv.it/en/bsi>) of the Istituto Nazionale di Geofisica e Vulcanologia (INGV). Locations and magnitudes of earthquakes are evaluated in real-time in the surveillance room of the INGV in Rome and then revised by the analysts of the BSI (all events with $M_L \geq 3.5$ are quickly revised, whereas the standard review is done for smaller events, within an agreed timeframe; Marchetti *et al.* 2016). The earthquake data come from the manual picking of seismic signals recorded by the National Seismic Network of the INGV, of which a significantly re-organized arrangement came into operation at 16 April 2005 (Amato & Mele 2008); therefore, only earthquakes after this time, and before 01 January 2022, are analysed in this study. Moreover, we select events shallower than 30 km in depth, that is about 95 per cent of all earthquakes, that occurred within the collection area, previously defined for the CSEP Italian earthquake forecast experiment (Schorlemmer *et al.* 2010a). To improve the quality of data, we apply the following selection constraints.

(i) Since this study is focused on eventual b -value changes not associated with large earthquakes, we ruled out the most intense phases of major seismic sequences, having a main shock with $M_L \geq 5.5$: the 2009 L'Aquila (main event 2009 April 6, M_L 5.9, M_w 6.1), the 2012 Emilia (main event 2012 May 20, M_L 5.9, M_w 5.8) and the 2016–2017 Central Italy (main event 2016 October 30, M_L 6.1, M_w 6.5) sequences (Table 1; Fig. 1). However, these sequences were investigated by Lombardi (2023), who found no firm evidence for short-term b -value variations.

(ii) The seismicity that occurred in the Sicilian (Etna, Eolian Islands and Pantelleria) and Campanian (Vesuvius, Phlegraean Fields and Ischia Island) volcanic areas (Table 1; Fig. 1) is removed from the catalogue, since the data of these areas are inhomogeneous, being obtained by both national and regional monitoring networks (<https://www.ingv.it/en/monitoring-and-infrastructure/monitoring-networks/ingv-and-its-networks>).

(iii) Starting in 2015, the BSI only revises events with magnitude above M_L 1.5 (Marchetti *et al.* 2016), corresponding to the detection

magnitude of a significant part of the national network (Schorlemmer *et al.* 2010b). This threshold is raised in exceptional cases, as the first 2 yr of 2016–2017 Central Italy sequence, which is, however, excluded from the present analysis. Anyway, we decided to consider events with $M_L \geq 1.5$ and revised by BSI analysts.

By applying these constraints, we select about 152 000 events (Fig. 1).

3 SPATIO-TEMPORAL b -VALUE SERIES

To search for any spatio-temporal variation of b -value in Italy, we perform a spatio-temporal scan of b and M_c parameters, using the Normalized Distance (ND) test (Lombardi 2021, 2023), at $\alpha = 0.01$ significance level. This method adopts the geometric distribution for magnitudes, which account for rounded measurements at first decimal place (Bender 1983; Marzocchi & Sandri 2003; Tinti & Mulargia 1987; Lombardi 2021). Even if rounded magnitudes do not significantly affect the b -value estimation (Marzocchi & Sandri 2003), at least for grouping interval $\delta M = 0.1$, we prefer adopting a discrete distribution, to make best use of statistical tests (Spinelli 2001; Haschenburger & Spinelli 2005), which are the focus of this work.

The cumulative density function for magnitudes is

$$F(M|b) = 1 - (1 - p)^{i+1} \quad i = 0, 1, \dots \quad (1)$$

where $M = M_c + i \cdot \delta M$, $\delta M = 0.1$ and $p = 1 - e^{b \cdot \ln(10) \cdot \delta M}$. In the following, $f(M|b)$ will mark the related probability density function of $F(M|b)$. Without loss of generality, we may assume $M_c = 0$, by considering shifted magnitudes $M - M_c$. The Maximum Likelihood Estimate (MLE) of b , \hat{b} , and the associated asymptotic error, $\sigma_{\hat{b}}$, are, then, given by

$$\hat{b} = \frac{-\ln(1 - \hat{p})}{\ln(10) \cdot \delta M} \quad \text{and} \quad \sigma_{\hat{b}} = \frac{\hat{p}}{\ln(10) \cdot \delta M \cdot \sqrt{N \cdot (1 - \hat{p})}}, \quad (2)$$

where $\hat{p} = \frac{\delta M}{(M + \delta M)}$, N is the sample size and \bar{M} is the average magnitude in the data (Tinti & Mulargia 1987; Lombardi 2021).

The formula for $\sigma_{\hat{b}}$ refers to the asymptotic gaussian distribution of the MLE estimator, obtained by applying the Limit Central Theorem, that may be unreliable for small sample size. Moreover, the small accuracy of MLE estimator for small magnitude ranges (Geffers *et al.* 2022, 2023) calls into question the indiscriminate use of this formula.

Note that the whole statistical development, described below, depends on assumption represented by eq. (1). Further models might be certainly considered for magnitudes, such as the tapered GRL (Main & Burton 1984; Kagan 1991). Anyway, the discrimination between models requires a number of data, which is not compatible with constraints of present study (Marzocchi *et al.* 2020; Geffers *et al.* 2022).

Considering the moderate seismic rate of Italy, we have to apply an adequate subsampling scheme, to avoid too small samples. The ND test is applied on intersecting samples of NT events, temporally moved forward through the catalogue by 1 per cent of NT events. Note that the b -value maps are not equally distributed in time, since they follow the temporal evolution of seismicity. Each sample is spatially distributed in sub-samples, having locations inside overlapping circles, covering the whole Italian mainland (excluding the Sicilian and Campanian volcanic zones and the almost aseismic Sardinia island), with centers C_j , equally spaced at distance $D = 0.4^\circ$. The radius R of each cell may be fixed (Fixed Radius, FR, method) or defined by the NR nearest neighbor events to C_j (Variable Radius, VR, method). Finally, since the b -value estimator may

Table 1. Spatio-temporal details on seismicity ruled out from the analysis.

| | | Coordinates polygon vertices | Temporal period |
|-------------------------|---|------------------------------|---------------------------|
| seismic sequence | L'Aquila 2009 | 12.70E 42.00N | 06 Apr 2009 / 05 Apr 2010 |
| | | 12.70E 43.30N | |
| | | 13.80E 43.30N | |
| | | 13.80E 42.00N | |
| seismic sequence | Emilia 2012 | 10.60E 44.50N | 20 May 2012 / 19 May 2013 |
| | | 10.60E 45.30N | |
| | | 11.90E 45.30N | |
| | | 11.90E 44.50N | |
| seismic sequence | Central Italy 2016–2017 | 12.70E 42.00N | 24 Aug 2016 / 23 Aug 2018 |
| | | 12.70E 43.30N | |
| | | 13.80E 43.30N | |
| | | 13.80E 42.00N | |
| Sicilian volcanic area | Etna | 14.70E 37.50N | 06 Apr 2005 / 31 Dec 2021 |
| | | 14.70E 37.90N | |
| | | 15.30E 37.90N | |
| | | 15.30E 37.50N | |
| Sicilian volcanic area | Eolian Islands | 14.84E 38.32N | 16 Apr 2005 / 31 Dec 2021 |
| | | 14.84E 38.48N | |
| | | 14.28E 38.48N | |
| | | 14.28E 38.63N | |
| | | 15.00E 38.63N | |
| | | 15.00E 38.71N | |
| | | 15.11E 38.71N | |
| | | 15.11E 38.87N | |
| | | 15.32E 38.87N | |
| | | 15.32E 38.71N | |
| | | 15.18E 38.71N | |
| | | 15.18E 38.58N | |
| | | 15.07E 38.58N | |
| | | 15.07E 38.32N | |
| Sicilian volcanic area | Pantelleria | 11.87E 36.69N | 16 Apr 2005 / 31 Dec 2021 |
| | | 11.87E 36.89N | |
| | | 12.12E 36.89N | |
| | | 12.12E 36.69N | |
| Campanian volcanic area | Vesuvius Phlegraean Fields Ischia Island | 13.78E 40.63N | 16 Apr 2005 / 31 Dec 2021 |
| | | 13.78E 40.92N | |
| | | 14.53E 40.92N | |
| | | 14.53E 40.63N | |

be strongly biased for small sample sizes, N and magnitude ranges, DM , covered by data (Geffers *et al.* 2022; Lombardi 2023), we consider samples that (a) cover a magnitude range $DM \geq 2.0$, the minimum value suggested to correctly estimate and test a GR model with $b = 1.0$ (Geffers *et al.* 2022) and (b) have more than 50 events above M_c . Different analyses are performed to evaluate the impact of spatiotemporal subsampling on results. In the following, FRA1 ($NT = 1000$ and $R = 100$), FRA2 ($NT = 5000$ and $R = 50$) and VRA ($NT = 5000$ and $NR = 300$) mark two analyses with FR method and one analysis with VR method, respectively. For VRA we consider radii below 200 km.

By applying FR and VR strategies, we select samples covering highly variable time intervals, ranging from about 20 d to 6 months (with a median of 3 months) for FRA1, from 6 months to 2 yr (with a median of 1 yr) for FRA2 and from 1 month to 2 yr (with a median of 1 yr) for VRA. The radius of cells for VRA goes from 7 to 200 km, with a median of about 100 km, depending on seismic rate of each zone. The number of events above M_c goes from 50 to 583 (with a median of 97) for FRA1, from 50 to 884 (with a median of 96) for FRA2 and from 50 to 300 (with a median of 134) for VRA.

We find that the values of \hat{b} are, on average, larger in Central Italy (zone A; median \hat{b} equal to 1.2), than the rest of territory, whatever subsampling method would be used (Fig. 2). The spatial distributions of median and confidence limits at 99 per cent of \hat{b} show similar results for all sampling methods. Lower \hat{b} (median \hat{b} equal to 0.9) are close to national borders, but this is an artifact of different network detection inside and outside them. The Northern Italy region may be analysed only by VRA, due to lower seismic rate of this zone. The northeastern zone (zone B) stands out for the greater precision of b estimations, since both 99 per cent confidence bounds of \hat{b} are close to 0.8.

The lower completeness magnitude is estimated in Central Italy (zone A), where the 99 per cent percentiles of M_c is mostly below 2.0 (Fig. 3). This is due to good network detection of this zone (Schorlemmer *et al.* 2010b), but also to removal of two important sequences from database, during which M_c significantly increased. The largest incompleteness is observed in Northern Italy (zone B) and in Sicily (zone C), where M_c reaches 2.8 and 2.5, respectively, both for the occurrence of stronger earthquakes and for the lower network detection in part of these areas.

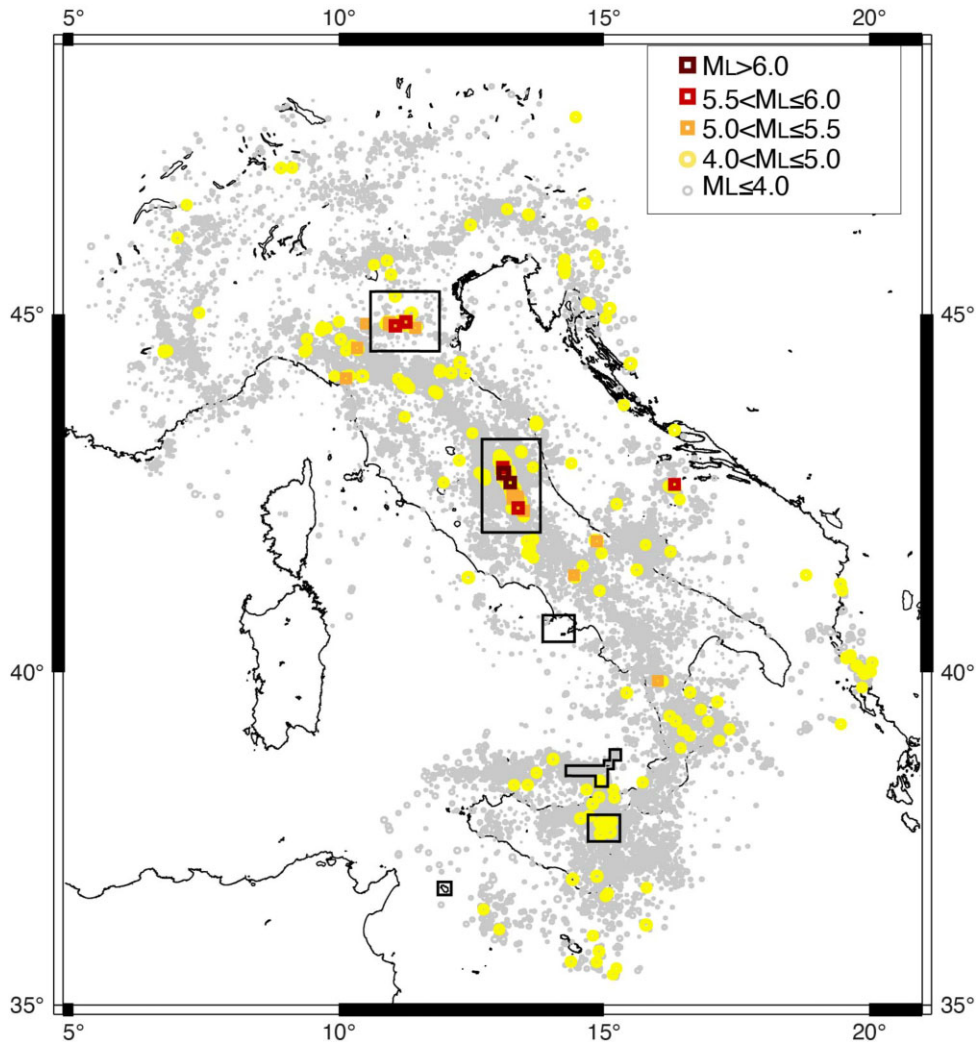


Figure 1. Map of the events that occurred in Italy, from 16 April 2005 to 31 December 2022, above 30 km of depth and with magnitude above M_L 1.5. Solid black lines mark areas in which seismicity is ruled out, as explained in main text and Table 1.

4 STATISTICAL TESTS

In this study, a special focus is laid on the significance of changes in b -value parameter. Specifically, a set of statistical tests is performed, at 0.01 significance level, devoted to judge (a) if estimated b -values are significantly different by a reference value b_0 and (b) if there are temporal and/or spatial significant variations in b -value time-series.

A proper approach to test b -value changes relies on a correct evaluation of uncertainty of b -values estimator. Traditionally, this is assumed asymptotically (for large N) normal, both for exponential (Aki 1965; Shi & Bolt 1982) and for geometric (Bender 1983; Tinti & Mulargia 1987; see also eq. 2) magnitudes. In any case, a detailed study of the spatiotemporal features of b -values usually deals with small data sets, for which the b -value estimator may be slightly accurate and precise (Geffers *et al.* 2022, 2023). In these cases, a mistaken assumption on \hat{b} uncertainty may affect the results of statistical tests for checking b -value changes (Schorlemmer *et al.* 2003; Amorè *et al.* 2010; Marzocchi *et al.* 2020); therefore, the use of bootstrap resampling to estimate more realistic errors in \hat{b} is recommended (Schorlemmer *et al.* 2003; Amorè *et al.* 2010).

We go into detail of this topic in the Supplementary Information, for both exponential and geometric magnitude distribution, by means of simulations. Specifically, we check the reliability of the

asymptotic predicted distribution on simulated data, for different values of real b -value b and sample size N . It shows that the distribution of b -value MLE significantly differs from a Gaussian one, for DM below a threshold, that depends on the (unknown) b and N (see Figs S1 and S2). Unfortunately, an analytic formula for the statistical distribution of the conditional (to N and DM) MLE of b cannot easily be computed and, therefore, we propose, in the following, a non-parametric approach.

In this study two types of tests are performed: (a) the t -test, assessing an hypothesis on population average, which is closely related to b -values estimator (eq. 1) and (b) the log-likelihood ratio test, comparing the goodness of fit of data to two specified distributions (Snedecor & Cochran 1989; Casella & Berger 2001). Since the aleatory uncertainty on \hat{b} depends on covered magnitude range DM , besides on sample size N (Geffers *et al.* 2022; Lombardi 2023), we construct statistical tests conditioned to both N and DM , to guarantee a rigorous judgement. This may be reached by generating synthetic data, that represent the sample distribution under the null hypothesis H_0 of a test, and, at the same time, account for values of N and DM of sample under testing. The easiest way to control both N and DM in simulations is to apply non-parametric bootstrap resampling algorithms (Efron & Tibshirani 1993). The use of

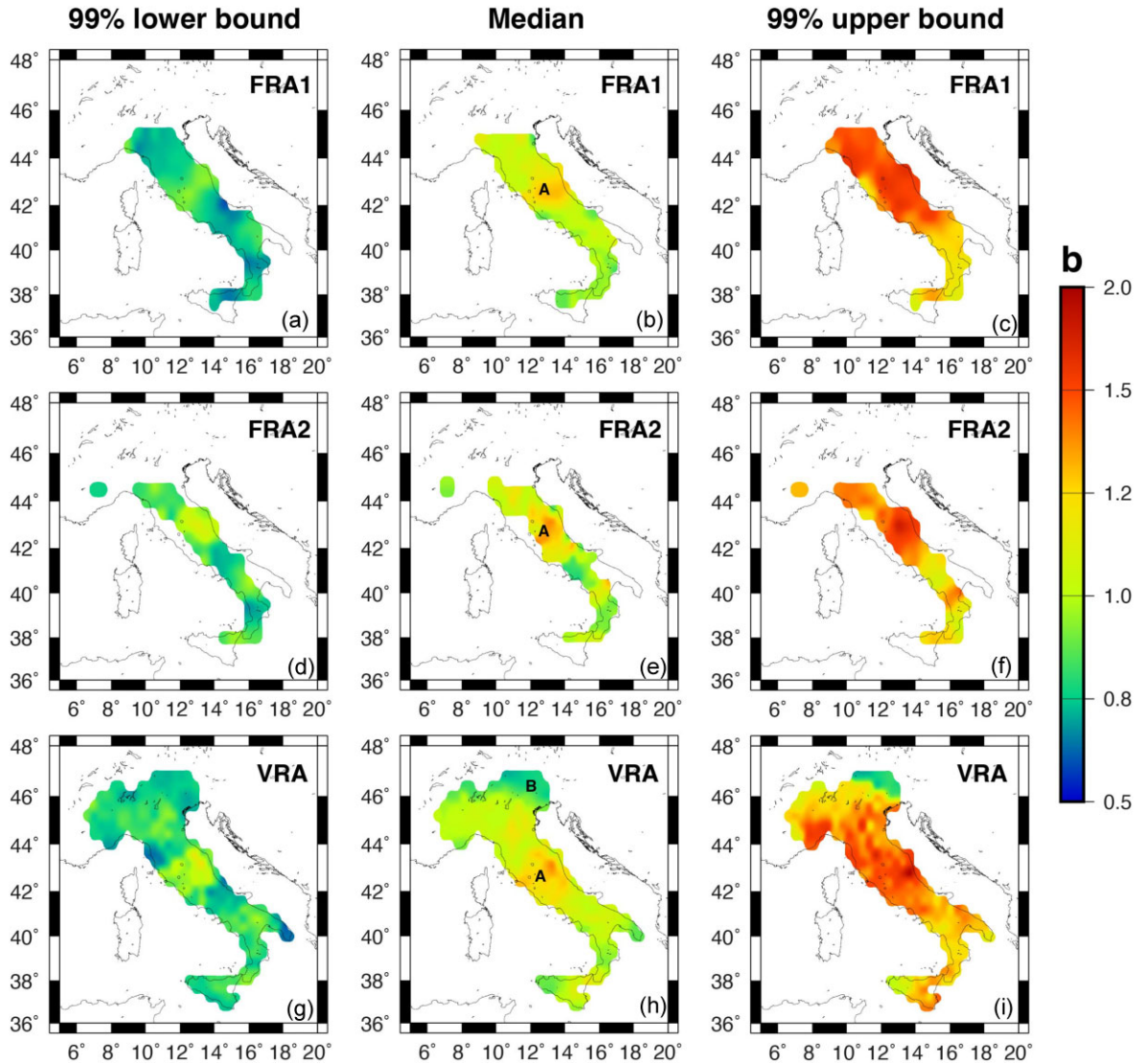


Figure 2. Maps of percentiles in time series of \hat{b} : the 99 per cent confidence bounds and the 50th percentile are plotted for FRA1 (panels a, b, c), FRA2 (panels d, e, f) and VRA (panels g, h, i). Zones marked by A and B refer to Central and Northeastern Italy regions, cited in the main text.

simulations has also advantage of ruling out arbitrary assumptions on distribution of test statistics under H_0 , valid only under specific conditions.

All applied statistical tests are described below, whereas in the Appendix we report results of simulation studies, performed to assess their robustness.

4.1 The maximum magnitude (M_{\max}) test

Any judgement of this study on b -values is done considering the values of N and DM . Therefore, we apply a preliminary test, devoted to judge if the maximum magnitude, M_{\max} , and, then, the magnitude range, DM , of a sample with N earthquakes is compatible with what expected for $b_0=1.0$ (null hypothesis H_0). The cumulative density function for M_{\max} depends on N , beyond b , and is given by

$$F_{M_{\max}}(M|b) = [F(M|b)]^N. \quad (3)$$

So, we can compare each value of M_{\max} with the 99 per cent confidence interval, given by b_0 , and reject b_0 for subsamples having a value of M_{\max} outside it.

We find that DM values are all inside the confidence intervals, given by eq. (3) with $b = b_0$; therefore they are all consistent with H_0 , at significance level $\alpha = 0.01$, for all three analyses FRA1, FRA2 and VRA (Fig. 4).

4.2 The one-sample bootstrap goodness of fit tests

The first set of tests is devoted to judge if values \hat{b} are significantly different by a reference value $b_0 = 1.0$ (null hypothesis H_0). To this end, we adopt two tests: the bootstrap t , BT, test, checking if mean magnitude is as expected under H_0 , and the bootstrap Log-Likelihood Ratio (BLLR) test, that judges reliability of H_0 , by mean of Log-Likelihood Ratio (Kalbfleisch 1985).

The BT test for a magnitude sample $S = \{M_1, \dots, M_N\}$, with mean $\mu(S)$ and standard deviation $\sigma(S)$, consists of following steps:

- (i) We compute the t -statistic

$$t[S] = \frac{\mu(S) - M^*}{\sigma(S)/\sqrt{N}}, \quad (4)$$

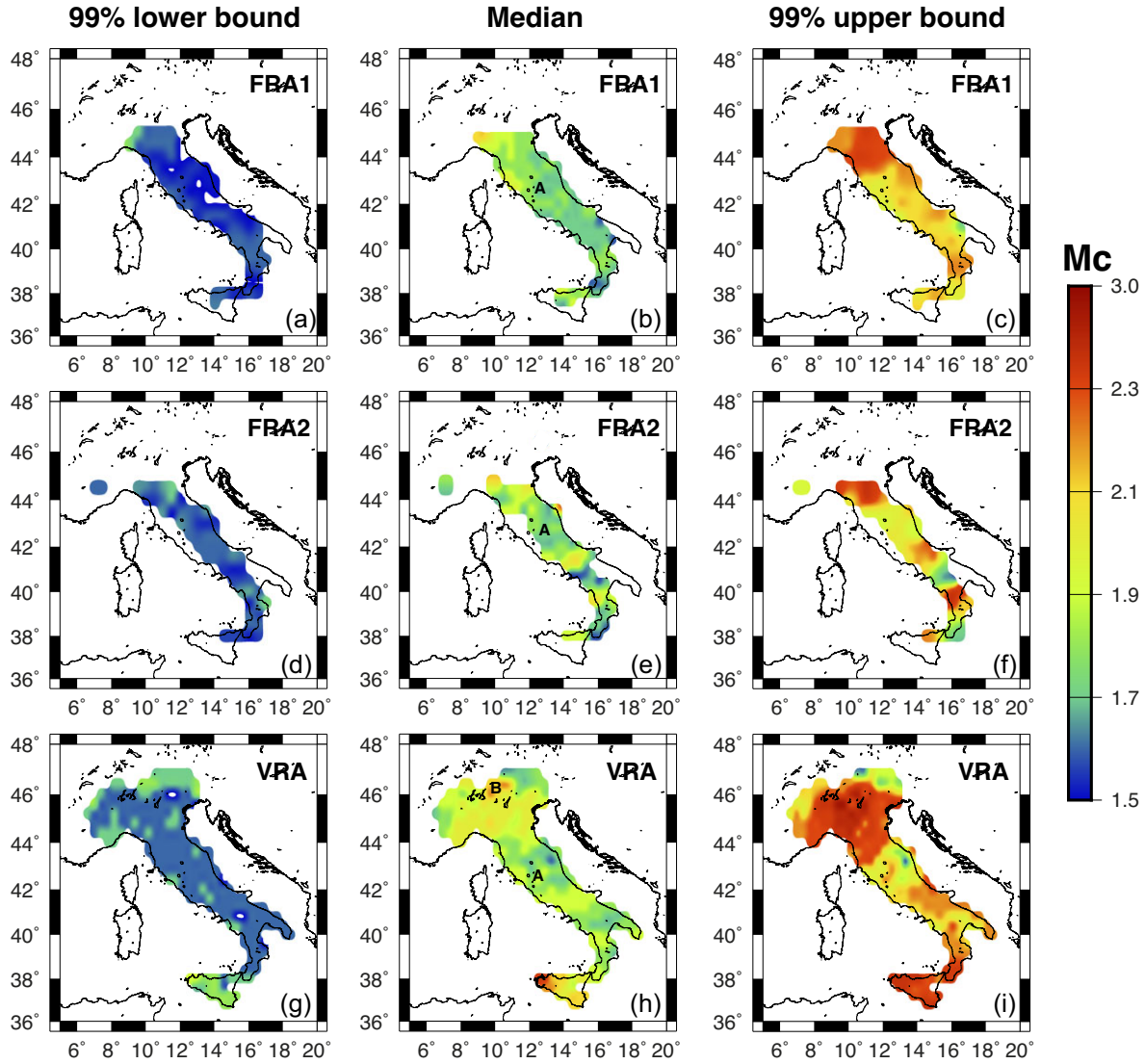


Figure 3. The same of Fig. 2, but for M_c . Zones marked by A, B and C refer to Central Italy, Northern Italy and Sicily regions, cited in the main text.

where $M^* = \frac{(1-p_0)}{p_0} \cdot \delta M$ is the mean expected for geometric variables with distribution $F(M|b_0)$.

- (ii) We produce N_B bootstrap samples $S_j, j = 1, \dots, N_B$, from S .
- (iii) We compute the quantity

$$t'[S_j] = \frac{\mu(S_j) - \mu(S)}{\sigma(S)/\sqrt{N}} \quad (5)$$

on all S_j ; their distribution represents the uncertainty of t -statistic under H_0 , for a sample with size N and range DM (Efron & Tibshirani 1993).

- (iv) We reject H_0 if $t[S]$ is out the $(1-\alpha)$ per cent empirical confidence interval, identified from values $t'[S_j]$.

The BLLR test compares the goodness of fit of geometric distributions with parameter b_0 and \hat{b} , whether to decide if the second is significantly better. For each magnitude sample $S = \{M_1, \dots, M_N\}$, the BLLR test consists of following steps (Davison & Hinkley 1997):

- (i) We estimate b -value on S , \hat{b} , and we compute the statistic of Log-Likelihood Ratio Test (Kalbfleisch 1985) for \hat{b} and b_0 ,

given by

$$LLR(S; \hat{b}, b_0) = -2 \ln \left[\frac{LL(S; \hat{b})}{LL(S; b_0)} \right]; \quad (6)$$

where

$$LL(S; b) = \prod_{i=1}^N f(M_i|b) \quad (7)$$

is the log-likelihood of the geometric distribution (eq. 1), with parameter b , for a sample S .

- (ii) We produce N_B bootstrap samples $S_j, j = 1, \dots, N_B$ from S , and we estimate b -values on all S_j (marked by \hat{b}_j).

(iii) We compute $LLR(S_j; \hat{b}_j, \hat{b})$ for all S_j , which represent the distribution of LLR under the null hypothesis (Davison & Hinkley 1997); therefore we reject H_0 if the proportion of $LLR(S_j; \hat{b}_j, \hat{b})$ exceeding $LLR(S; \hat{b}, b_0)$ is larger than α .

We apply both BT and BLLR tests, by producing $N_B = 10^5$ bootstrap resampling of all detected subsamples. The two tests identify anomalous b -values on the same subsamples. FRA1 and FRA2

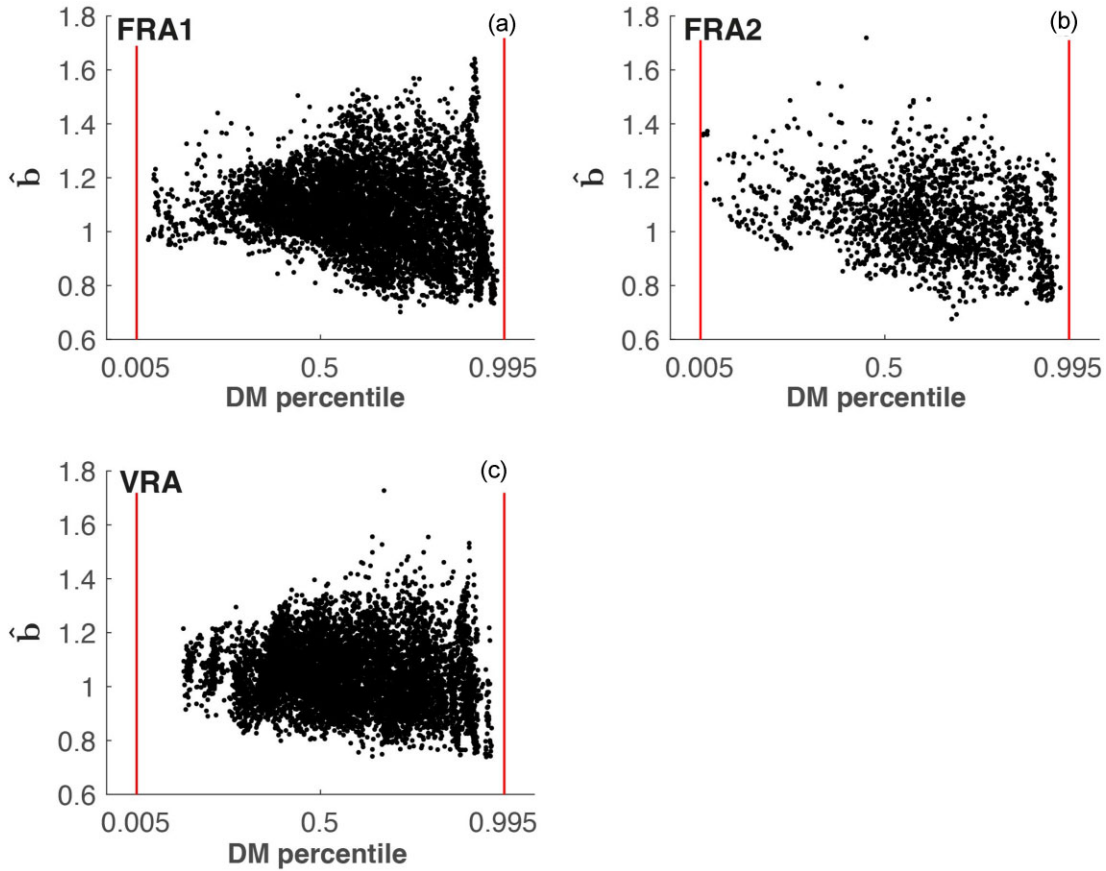


Figure 4. Plot of p -values of M_{\max} test for all subsamples and (a) FRA1, (b) FRA2 and (c) VRA. All of them are inside (0.05, 0.995) interval and, then, M_{\max} values are in agreement with the null hypothesis $b = b_0 = 1$.

detect anomalous high b -values, mostly in Central Italy and for subsamples with lower DM and M_c , whereas no significant anomaly is recognized by VRA (Fig. 5). Very scattered and occasional low b are detected in all three analyses, in border cells, due to inclusion of larger events out of national boundaries, where the detection is higher. No zone shows systematic anomalies, since each cell has 90 per cent, at least, of estimated b -values not significantly different from 1.0. The reason for which we do not find anomalous high b -values by VRA, differently from when we apply FRA1 and FRA2, is due to the different sampling methods. Considering the high seismic rate and the good detection of this zone, subsamples identified by VRA refer to smaller areas respect to FRA1 and FRA2 and cover smaller magnitude ranges DM . In more than 99 per cent of anomalous cases recognized by FRA2, VRA identifies samples with $DM < 2$ and/or $N < 50$, which, therefore, do not satisfy the requests to be selected and are kept out from the analysis.

4.3 The bootstrap two-samples tests

The second phase of testing analysis is devoted to check relative differences of b -values, in time and space. Specifically, they are devoted to judge if two independent (i.e. without events in common) magnitude samples have equal b -values (null hypothesis H_0). Also in this case, we use a non-parametric test, the two sample t -test (Snedecor & Cochran 1989), and a parametric one, the two-samples Log-Likelihood Ratio test (Kalbfleisch 1985). Moreover, we apply bootstrap resampling to take into account basic features of samples

under testing and to rule out arbitrary assumptions on distribution of test statistics under H_0 .

The two-samples bootstrap t -test, 2S-BT, is used to determine if two population means are equal (Snedecor & Cochran 1989). Given two independent samples $S_1 = \{M_1^1, \dots, M_{N_1}^1\}$ and $S_2 = \{M_1^2, \dots, M_{N_2}^2\}$, the test statistic, if equal variances are assumed, is given by

$$T_{2S}(S_1, S_2) = \frac{\mu(S_1) - \mu(S_2)}{s \cdot \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}, \quad (8)$$

where $s = \sqrt{\frac{(N_1 - 1) \cdot \sigma(S_1)^2 + (N_2 - 1) \cdot \sigma(S_2)^2}{N_1 + N_2 - 2}}$ is the pooled variance of S_1 and S_2 .

Also in this case, we apply a bootstrap version of this test, consisting of following steps:

- (i) For a couple of samples (S_1, S_2), we compute $T_{2S}(S_1, S_2)$.
- (ii) We combine S_1 and S_2 and produce N_B bootstrap resampling $S_j, j = 1, \dots, N_B$.
- (iii) For all j , we compute the quantities $T'_{2S}(S_j^1, S_j^2)$, where S_j^1 and S_j^2 are the samples given by the first N_1 and the last N_2 elements of S_j , respectively; their distribution represents the uncertainty of T_{2S} -statistic under H_0 .
- (iv) We reject H_0 , if $T_{2S}(S_1, S_2)$ is out the $(1 - \alpha)$ per cent empirical confidence interval, identified from values $T'_{2S}(S_j^1, S_j^2)$.

The 2S-BT test is similar to significance test proposed by Amorè *et al.* (2010), also based on bootstrap resampling, which,

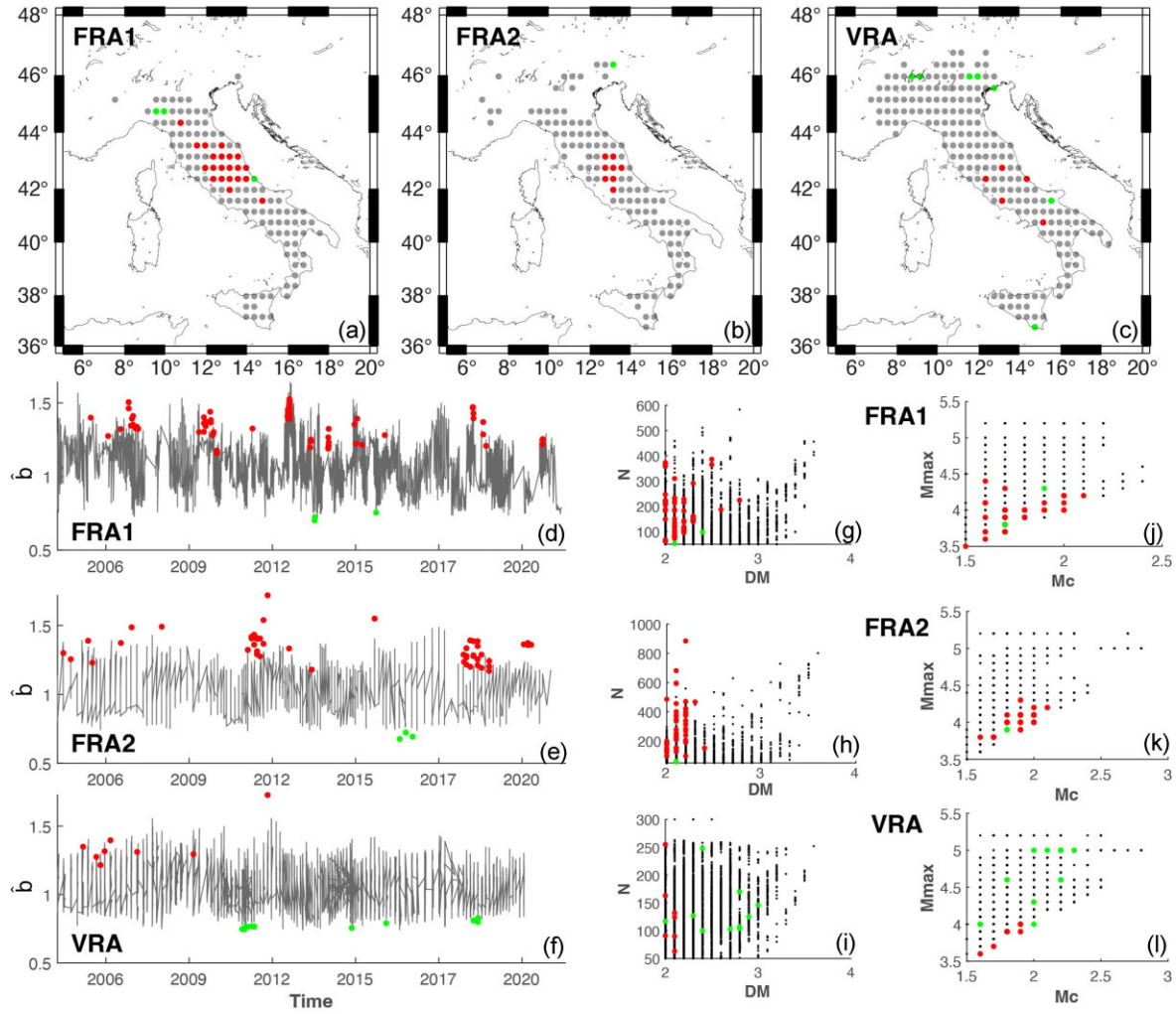


Figure 5. Results of bootstrap goodness of fit tests BT and BLLR, which are consistent in identifying b -values anomalies. Red and green circles mark the anomalous high and low b -values, respectively, for all panels, whereas grey circles mark b -values consistent with H_0 . (a) Map of cells for which values of \hat{b} are available by FRA1. (b) The same of (a) but for FRA2. (c) The same of (a) but for VRA. (d) Time versus \hat{b} plot for all subsampling identified by FRA1. (e) The same of (d) but for FRA2. (f) The same of (d) but for VRA. (g) Magnitude range DM versus sample size N of all subsamples for which b -value is estimated by FRA1. (h) The same of (g) but for FRA2. (i) The same of (g) but for VRA. (j) Maximum magnitude M_{\max} versus completeness magnitude M_c of all subsamples for which b -value is estimated by FRA1. (k) The same of (j) but for FRA2. (l) The same of (j) but for VRA.

however, does not include sample variances or uncertainty on b -value estimations.

The second test, the two-sample log-likelihood ratio test, 2S-LLRT, judges if S_1 and S_2 have the same distribution, by mean of Log-Likelihood Ratio test (Kalbfleisch 1985). In this case, the statistic test is

$$\overline{LLR}_{2S} = LLR_{2S}(S_1, S_2; \hat{b}_1, \hat{b}_2) = -2 \ln \left[\frac{LL(S_1; \hat{b}_1) + LL(S_2, \hat{b}_2)}{LL(S_{1,2}; \hat{b}_{1,2})} \right], \quad (9)$$

where $S_{1,2}$ is the sample obtained by combining S_1 and S_2 , and $\hat{b}_{1,2}$ is the estimated b -value for $S_{1,2}$. We apply this test to all couples of independent magnitude subsamples, in order to take into account both possible temporal and spatial b -values variations. The test for each couple of magnitude samples S_1 and S_2 consists of following steps:

(i) We estimate b -value on the magnitude samples S_1 and S_2 , \hat{b}_1 and \hat{b}_2 , and we compute \overline{LLR}_{2S} .

(ii) We produce N_B bootstrap resampling from $S_{1,2}$, we split each of them in two samples S_j^{1*} and S_j^{2*} , $j = 1, \dots, N_B$, with size N_1 and N_2 , respectively, and estimate b -values \hat{b}_j^{1*} and \hat{b}_j^{2*} .

(iii) We compute $LLR_{2S}^{j*} = LLR_{2S}(S_j^{1*}, S_j^{2*}; \hat{b}_j^{1*}, \hat{b}_j^{2*})$ and we reject the null hypothesis, $\hat{b}_1 = \hat{b}_2$, if the proportion of LLR_{2S}^{j*} exceeding \overline{LLR}_{2S} is larger than α .

P -values of BT2 and BLLR2 are consistently lower than $\alpha = 0.01$ when p -values of BT and BLLR tests are low for one of two samples in comparison (Fig. 6). In other words, significant relative variations of b -values are found when the null hypothesis $b_0 = 1$ of one-sample tests is rejected (or is close to being rejected) for one of two samples in comparison, showing that response of two samples tests is consistent with results provided by one-sample ones. The good degree of consistency among test judgements improves reliability of statistical inferences and helps to extract more efficiently information from data.

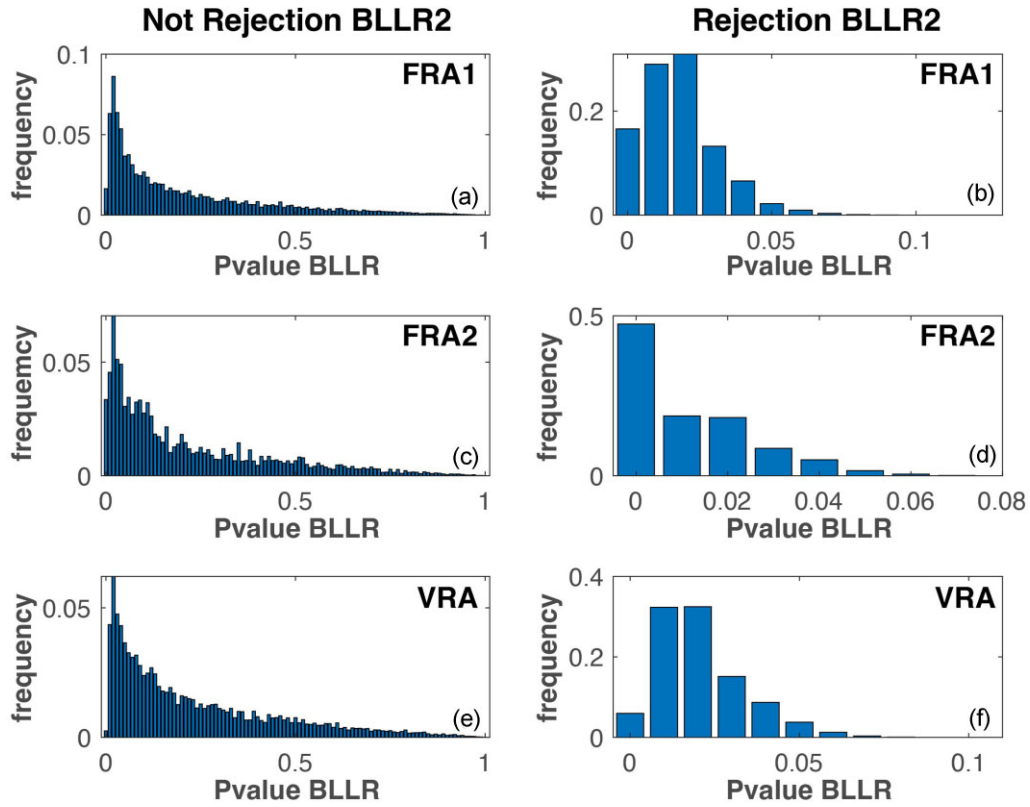


Figure 6. Comparison of p -values for BLLR and 2S-BLLR tests. (a) Histogram of minimum between two p -values of BLLR test for all couples of subsamples, defined by FRA1 and tested by 2S-BLLR test, for which the null hypothesis H_0 is not rejected. (b) The same of (a), but for cases of rejection of H_0 . (c) The same of (a), but for subsamples identified by FRA2. (d) The same of (b), but for subsamples identified by FRA2. (e) The same of (a), but for subsamples identified by VRA. (f) The same of (b), but for subsamples identified by VRA.

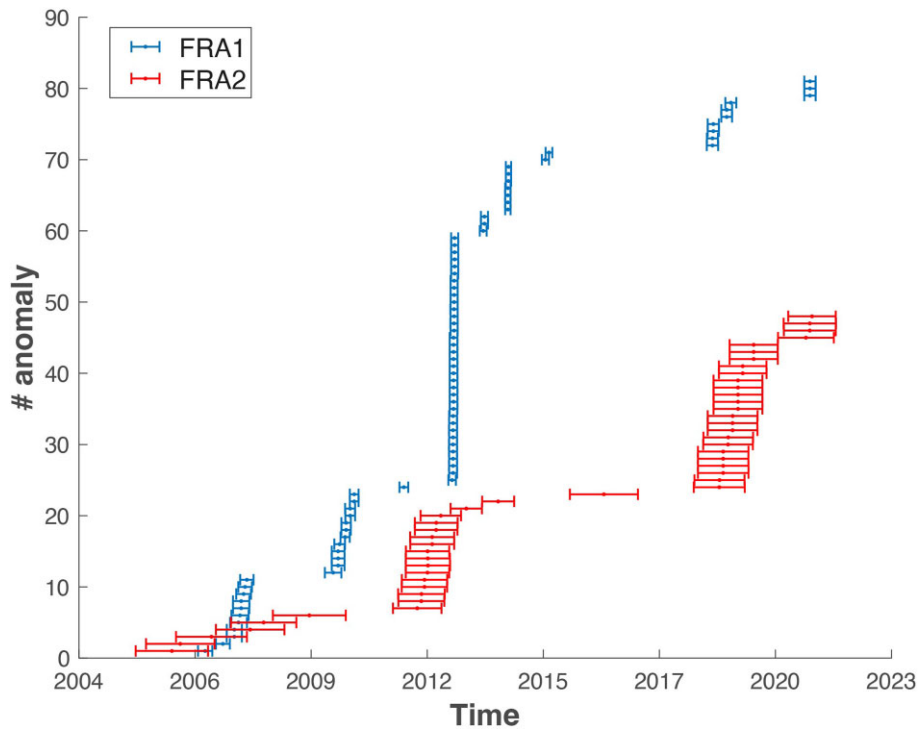


Figure 7. Comparison of cumulative numbers of time frames in which anomalous b -values are detected by FRA1 (blue) and FRA2 (red).

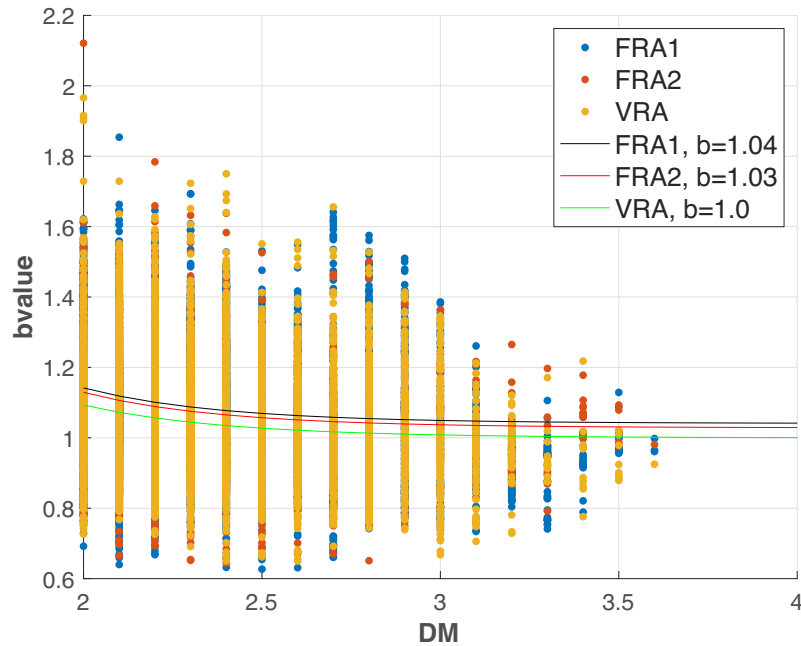


Figure 8. b -Values as a function of dynamic range, for the three analysis: FRA1, FRA2 and VRA. Solid lines represent the best fit of formula proposed by Geffers *et al.* (2023), to model the bias in estimating b at low magnitude ranges.

5 DISCUSSION AND CONCLUSIONS

This study is devoted to quantify the variability of b -value in Italy, in time and/or space, that is not associated to evolution of important seismic sequences. To achieve this goal some key aspects of statistic techniques and data selection have been addressed:

(i) We assume a discrete, geometric, probability distribution for magnitudes, to take into account the first decimal rounding, which has a negligible impact on b -value estimation, but may affect statistical test results (Spinelli 2001; Haschenburger & Spinelli 2005).

(ii) We exclude samples with too low size and/or magnitude range, to minimize bias in b -value estimation (Geffers *et al.* 2022).

(iii) We adopt the ND method (Lombardi 2021, 2023) to estimate b , since it includes uncertainty on M_c .

(iv) We apply different subsampling methods and parametric setting, to quantify their impact on results.

(v) We use non-parametric bootstrap algorithms (Efron & Tibshirani 1993; Davison & Hinkley 1997) in the testing phase, to control basic features of magnitude samples, with particular care for range.

The values of \hat{b} are on average larger in Central Italy, whatever subsampling method would be used (Fig. 2), confirming what Taroni *et al.* (2021) founded by applying a weighted maximum-likelihood estimation method on the whole catalogue (i.e. without temporal binning).

By testing the significance of estimated b -values, we find that no zone has a b -value systematically different from 1.0, since more than 90 per cent of b -value estimations in each cell is not recognized as anomalous. This result puts into question the influence of tectonic regimes and differential stress on earthquake size distribution (Schorlemmer *et al.* 2005; Gulia & Wiemer 2010; Taroni & Carafa 2023) and suggest that possible b -value variations may be mostly associated with temporary processes. Spotty anomalously

low b -values, that is significantly lower than $b_0 = 1$, are founded in boundary zones, but this is an artefact of variation of network detection across the borders. Consistently with Taroni *et al.* (2021), we conclude that no zone has a significantly low b -value. Some temporary high b -values, that is significantly larger than $b_0 = 1$, are consistently found in Central Italy by all tests by both FR analyses. Again, this result is consistent with Taroni *et al.* (2021), though they do not consider the nature, permanent or temporary, of these anomalies, since they analyse only spatial and not temporal changes of magnitude distribution.

The VR method does not identify zones and periods in which the magnitude distribution changes. As explained above, this is due to sampling method applied by VRA. Indeed, considering the high seismic rate of this zone, this method select the NR nearest neighbour events close to cell centre and covering a magnitude range lower than 2.0; therefore these samples are ruled out from testing. Also if the different subsampling methods, applied here, provide essentially consistent results (Figs 2 and 3), the comparison between FR and VR methods draw the attention on the sensitivity of the detection of b -values changes to the sampling method, confirming that data-driven strategies could significantly help to resolve the significance of any apparent b -value variation (Kamer & Hiemer 2015).

In view of these results, some questions arise. The first concerns why we find possibly significant anomalies only in Central Italy. Certainly, the low size of samples analysed in this work, which lead to relatively large aleatory uncertainty and low precision on the estimate, is a problem, whichever test we apply: just consider that Type II error, that is incorrectly failing to reject the null hypothesis, cannot be fully controlled, especially for small samples, and that, then, anomalies may be undetected. In other word, we must remember that the non rejection of an hypothesis does not mean that it is true, but only that data are not inconsistent with it. Therefore, anomalies out of Central Italy cannot be excluded as a possibility.

The same concerns the systematic nature of anomalies in Central Italy. We cannot exclude the possibility that *b* is always larger than 1, but there is indication against this thesis: (1) no anomaly is recognized for more than 90 per cent of temporal slots of each cell and (2) the seismic activity in this area may be strongly influenced by the presence of fluids (Miller *et al.* 2004; Chiadini *et al.* 2004) and of unrecognized anthropic activity (Cattaneo *et al.* 2014), that could justify temporary high *b*-values. It is hardy to analyse each single case, of course, and to diagnose the reasons for a different magnitude distribution, with respect to the rest of the national territory. The main question which should be answered is if detected anomalously high *b*-values are real or biased by statistical factors. Certainly, the overlapping timeframes of subsamples detected as anomalous by FRA1 and FRA2 (Fig. 7), which all have seismicity in common, and consistency of statistical test results strengthening the hypothesis of a real variation of magnitude distribution. On the other hand, some factors raise doubts on the nature of possible mechanisms. First, the anomalies are most recognized for magnitude subsamples covering a range close to 2.0 and for lower M_c (Fig. 5), for which there might be still some bias in *b*-value estimation (Marzocchi & Sandri 2003; Geffers *et al.* 2022). In this regard, Geffers *et al.* (2023) provides an analytic expression for the bias to high *b* at low *DM*. The best fit of \hat{b} data, as function of *DM*, converges to *b*-values close to 1.0 and indicates a bias up to $DM = 2.5$, for all three analysis (Fig. 8). This result identifies the statistical factors as the most likely cause of detected anomalous *b*-values. Secondly, the sizes of supposed anomalous samples are mostly well lower than 1000, threshold above which any consideration is statistically robust (Geffers *et al.* 2022). Finally, erroneously high *b*-values may be obtained if the GRL is applied on tapered small data samples (Marzocchi *et al.* 2020; Geffers *et al.* 2022). Therefore, even if no firm statement can be made about the geophysical relevance of *b*-values changes that we have detected, we interpret the anomalous high *b*-values to be statistically biased (contrary to Taroni *et al.* 2021).

In conclusion, this work supports a cautious approach towards assigning significance to apparent *b*-values variation, on medium spatio-temporal scales, since it indicates small and occasional changes, that are possibly artefacts of statistical bias. This calls into question hypotheses, such as correlation of magnitude distribution with crustal stress. The stringent statistical requirements, adopted in this study, highlight the need to produce improvements in detecting greater numbers of events, at national scale, that would strengthen their statistical treatment and help the understanding of underlying physical processes.

ACKNOWLEDGMENTS

The author is grateful to reviewers (J. Zhuang and one anonymous) and the associate editor (M. Segou) for their helpful reviews of the paper.

DATA AVAILABILITY

Earthquake data are collected in the Italian Seismological Instrumental and parametric Database (ISIDE, <http://terremoti.ingv.it/en/inside>), in the time frame between 1985 January 1 and today. Starting from 2015, quarterly releases of BSI are published on the web page <http://terremoti.ingv.it/en/bsi> in QuakeML format and with a DOI assigned.

SUPPORTING INFORMATION

Supplementary data are available at *GJI* online.

suppl.data

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the paper.

REFERENCES

- Aki, K., 1965. Maximum likelihood estimate of *b* in the formula $\log(N)=abM$ and its confidence limits, *Bull. Earthq. Res. Inst. Tokyo Univ.*, **43**, 237–239.
- Amato, A. & Mele, F.M., 2008. Performance of the INGV National Seismic Network from 1997 to 2007, *Ann. Geophys.*, **51**(2–3), 417–431.
- Amorèse, D., Grasso, J.R. & Rydelek, P., 2010. On varying *b*-values with depth, results from computer-intensive tests for Southern California, *Geophys. J. Int.*, **180**, 347–360.
- Bender, B., 1983. Maximum likelihood estimation of *b* values for magnitude grouped data, *Bull. seism. Soc. Am.*, **73**, 831–851.
- Casella, G. & Berger, R.L., 2001. *Statistical Inference*, 2nd edn, pp. 660, Duxbury, Pacific Grove.
- Cattaneo, M., Caffagni, E., Carannante, S., D'Alema, E., Frapiccini, M., Ladina, C., Marzorati, S. & Monachesi, G., 2014. A catalogue of non-tectonic earthquakes in central-eastern Italy, *Ann. Geophys.*, **57**, doi:10.4401/ag-6434.
- Chiadini, G., Cardellini, C., Amato, A., Boschi, E., Caliro, S., Frondini, F. & Ventura, G., 2004. Carbon dioxide Earth degassing and seismogenesis in central and southern Italy, *Geophys. Res. Lett.*, **31**, doi:10.1029/2004GL019480.
- Davison, A.C. & Hinkley, D.V., 1997. *Bootstrap Methods and Their Application*, Cambridge Univ. Press.
- Efron, B. & Tibshirani, T.J., 1993. *An Introduction to the Bootstrap*, pp. 436, Chapman and Hall.
- Frohlich, C. & Davis, S., 1993. Teleseismic *b*-values: or, much ado about 1.0, *J. geophys. Res.*, **98**(B1), 631–644.
- Geffers, G., Main, I.G. & Naylor, M., 2022. Biases in estimating *b*-values from small earthquake catalogues: how high are high *b*-values?, *Geophys. J. Int.*, **229**, 1840–1855.
- Geffers, G-M, Main, I.G. & Naylor, M., 2023. 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, *Geophys. J. Int.*, **232**(3), 2080–2086.
- Gulia, L. & Wiemer, S., 2010. The influence of tectonic regimes on the earthquake size distribution: a case study for Italy, *Geophys. Res. Lett.*, **37**, doi:10.1029/2010GL043066.
- Gulia, L., Wiemer, S. & Schorlemmer, D., 2010. Asperity based earthquake likelihood models for Italy, *Ann. Geophys.*, **53**(3), 63–75.
- Haschenburger, J.K. & Spinelli, J.J., 2005. Assessing the goodness-of-fit of statistical distributions when data are grouped, *Math. Geol.*, **37**, 261–276.
- Henderson, J.R., Barton, D.J. & Foulger, G.R., 1999. Fractal clustering of induced seismicity in The Geysers geothermal area, California, *Geophys. J. Int.*, **139**, 317–324.
- Herrmann, M. & Marzocchi, W., 2021. Inconsistencies and lurking pitfalls in the magnitude-frequency distribution of high-resolution earthquake catalogs, *Seismol. Res. Lett.*, **92**(2A), 909–922.
- Kagan, Y. Y., 1999. Universality of the seismic moment frequency relation, *Pure appl. Geophys.*, **155**, 537–573.
- Kagan, Y.Y. & Jackson, D.D., 2000. Probabilistic forecasting of earthquakes, *Geophys. J. Int.*, **143**, 438–453.
- Kagan, Y.Y., 1991. Seismic moment distribution, *Geophys. J. Int.*, **106**, 123–134.
- Kalbfleisch, J. G., 1985. *Probability and Statistical Inference (Volume 2: Statistical Inference)*, 2nd edn, pp. 1–360, Springer-Verlag.

- Kamer, Y. & Hiemer, S., 2015. Data-driven spatial b value estimation with applications to California seismicity. To b or not to b , *J. geophys. Res.*, **120**, 5191–5214.
- Lombardi, A.M., 2021. A normalized distance test for co-determining the completeness magnitude and b -value of earthquake catalogs, *J. geophys. Res.*, **126**, doi:10.1029/2020JB021242.
- Lombardi, A.M., 2023. Anomalies and transient variations of b -value in Italy during the major earthquake sequences: what truth is there to this?, *Geophys. J. Int.*, **232**(3), 1545–1555.
- Main, I. & Burton, P., 1984. Information theory and the earthquake frequency–magnitude distribution, *Bull. seism. Soc. Am.*, **74**(4), 1409–1426.
- Marchetti, A. et al., 2016. The Italian Seismic Bulletin: strategies, revised pickings and locations of the central Italy seismic sequence, *Ann. Geophys.*, **59**, doi:10.4401/ag-7169.
- Marzocchi, W. & Sandri, L., 2003. A review and new insights on the estimation of the b -value and its uncertainty, *Ann. Geophys.*, **46**, 1271–1282.
- Marzocchi, W., Spassiani, I., Stallone, A. & Taroni, M., 2020. How to be fooled searching for significant variations of the b -value, *Geophys. J. Int.*, **220**(3), 1845–1856.
- Miller, S.A., Colletini, C., Chiaraluce, L., Cocco, M., Barchi, M. & Kaus, B.J.P., 2004. Aftershocks driven by a high-pressure CO₂ source at depth, *Nature*, **427**, 724–727.
- Roberts, N.S., 2015. Are volcanic seismic b -values high, and if so when?, *J. Volc. Geotherm. Res.*, **308**, 127–141.
- Schorlemmer, D. & Wiemer, S., 2005. Microseismicity data forecast rupture area, *Nature*, **434**, doi:10.1038/4341086a.
- Schorlemmer, D., Christophersen, A., Rovida, A., Mele, F., Stucchi, M. & Marzocchi, W., 2010a. Setting up an earthquake forecast experiment in Italy, *Ann. Geophys.*, **53**(3), 1–9.
- Schorlemmer, D., Mele, F. & Marzocchi, W., 2010b. A completeness analysis of the National Seismic Network of Italy, *J. geophys. Res.*, **115**, doi:10.1029/2008JB006097.
- Schorlemmer, D., Neri, G., Wiemer, S. & Mostaccio, A., 2003. Stability and significance tests for b -value anomalies: example from the Tyrrhenian Sea, *Geophys. Res. Lett.*, **30**(16), doi:10.1029/2003GL017335.
- Schorlemmer, D., Wiemer, S. & Wyss, M., 2005. Variations in earthquake-size distribution across different stress regimes, *Nature*, **437**, 539–542.
- Shi, Y. & Bolt, B.A., 1982. The standard error of the magnitude–frequency b -value. *Bull. Seismol. Soc. Am.*, **72**, 1677–1687.
- Snedecor, G.W. & Cochran, W.G., 1989. *Statistical Methods*, 8th edn, Iowa State Univ. Press.
- Spinelli, J.J., 2001. Testing fit for the grouped exponential distribution, *Can. J. Stat.*, **29**(3), 451–458.
- Taroni, M. & Carafa, M.M.C., 2023. Earthquake size distributions are slightly different in compression vs extension, *Commun. Earth Environ.*, **4**, doi:10.1038/s43247-023-01059-y.
- Taroni, M., Zhuang, J. & Marzocchi, W., 2021. High-definition mapping of the Gutenberg–Richter b -Value and its relevance: a case study in Italy, *Seismol. Res. Lett.*, **XX**, 1–7.
- Tinti, S. & Mulargia, F., 1987. Confidence intervals of b -values for grouped magnitudes, *Bull. seism. Soc. Am.*, **77**, 2125–2134.
- Wiemer, S. & Wyss, M., 1997. Mapping the frequency–magnitude distribution in asperities, an improved technique to calculate recurrence times?, *J. geophys. Res.*, **102**, 15 115–15 128.
- Wiemer, S., McNutt, S.R. & Wyss, M., 1998. Temporal and three dimensional spatial analyses of the frequency–magnitude distribution near Long Valley Caldera, California, *Geophys. J. Int.*, **134**, 409–421.
- Wyss, M., Schorlemmer, D. & Wiemer, S., 2000. Mapping asperities by minima of local recurrence time, San Jacinto Elsinore fault zones, *J. geophys. Res.*, **105**, 7829–7844.

APPENDIX: NUMERICAL STUDY

We set up some simulation studies to assess the robustness of all bootstrap tests outlined in main text.

We apply one-sample tests, that is BT and the BLLR tests, on 10^5 synthetic samples, with size N equal to 50, 100, 300, 500, and 1000, generated from the geometric distribution (eq. 1 in main text), with $b_0 = 1$. The null hypothesis of all one-sample tests is $b = b_0$ and, therefore, the proportion of rejections of null hypothesis on simulated samples is the empirical Type I error rate and is expected to be close to significance level α . As is well known, the risk of committing a Type II error, that is incorrectly failing to reject the null hypothesis, cannot be fully controlled and can be minimized only by increasing the sample size.

Tables A1 lists the proportion of rejections of the null hypothesis, for BT, BLLR tests, for $\alpha = 0.01$ and 0.05. The number of bootstrap resampling is 10^5 .

The null hypothesis for the 2S-BT and 2S-BLLR tests is that two magnitude samples have the same b -value. The numerical study is performed by simulating couples of geometric samples, one of size N and one of size $2N$, with $b = 1$. Table A2 lists the proportion of rejections of the null hypothesis. Also in this case, the Type I error is well controlled.

Table A1. Proportion of simulations for which $b_0 = 1$ is rejected by one-sample tests at significance level $\alpha = 0.05$ and $\alpha = 0.01$. These values represent the Type I error of the tests.

| | N | | | | | | | | | |
|-----------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| | 50 | | 100 | | 300 | | 500 | | 1000 | |
| | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.05$ |
| BT test | 0.0096 | 0.048 | 0.011 | 0.051 | 0.0094 | 0.050 | 0.0089 | 0.049 | 0.0091 | 0.048 |
| BLLR test | 0.011 | 0.061 | 0.0082 | 0.049 | 0.0051 | 0.050 | 0.008 | 0.049 | 0.0073 | 0.046 |

Table A2. Proportion of simulations for which b_0 is rejected by 2S-BT and 2S-BLLR test. In each sample for all values of N and b , and significance level $\alpha = 0.05$ and $\alpha = 0.01$. These values represent the Type I error of the tests

| | N | | | | | | | | | |
|--------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| | 50 | | 100 | | 300 | | 500 | | 1000 | |
| | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.05$ |
| 2S-BT test | 0.011 | 0.048 | 0.0094 | 0.047 | 0.0059 | 0.051 | 0.0061 | 0.048 | 0.0092 | 0.046 |
| 2S-BLLR test | 0.0093 | 0.051 | 0.0089 | 0.049 | 0.0086 | 0.050 | 0.0084 | 0.049 | 0.0072 | 0.038 |