

An experiment on earthquake size distribution estimations reveals unexpected large epistemic uncertainty across methods

Anna Maria Lombardi,¹ Licia Faenza,² Francesco Pio Lucente,¹ Matteo Taroni,¹ Laura Gulia,³ Anna Tramelli,⁴ Emanuele Biondini⁵,⁶ Rodolfo Console^{1,6}, Vincenzo Convertito⁴, Cataldo Godano,⁷ Pasquale De Gori,¹ Giuseppe Falcone,¹ Alexander Garcia², Paolo Gasperini^{5,1}, Barbara Lolli², Caterina Montuori,¹ Maura Murru,¹ Laura Sandri,² Iliaria Spassiani¹ and Gianfranco Vannucci²

¹Istituto Nazionale di Geofisica e Vulcanologia, 00146, Rome, Italy. E-mail: annamaria.lombardi@ingv.it

²Istituto Nazionale di Geofisica e Vulcanologia, 40127, Bologna, Italy

³Swiss Seismological Service, ETH, 8006, Zurich, Switzerland

⁴Istituto Nazionale di Geofisica e Vulcanologia, 80124, Napoli, Italy

⁵Università di Bologna 'Alma Mater Studiorum', 40126, Bologna, Italy

⁶Center of Integrated Geomorphology for the Mediterranean Area, 85100, Potenza, Italy

⁷Università della Campania 'Luigi Vanvitelli', 81100, Caserta, Italy

Accepted 2025 May 22. Received 2025 May 20; in original form 2024 December 13

SUMMARY

The earthquake size distribution is well described by the Gutenberg Richter Law, controlled by the b -value parameter. In recent decades, a great variety of methods for estimating the b -value have been proposed by the scientific community, despite the simplicity of this relationship. All these methods underlie the different views of individual modellers and, therefore, often generate inconsistent results. In this study, we perform a seismological experiment in which we compare different, commonly adopted, methodologies, to estimate the completeness magnitude and the b -value, for seismicity in Central Italy. The intermethod differences are on average equal to 0.4 and 0.3, for M_c and b , respectively, but reach much larger values, especially during more intense seismic activity. This shows that epistemic uncertainty in the b -value plays a more crucial role than intramethod uncertainties, opening new perspectives in the interpretation of discrepant, single studies.

Key words: Statistical methods; Earthquake hazards; Statistical seismology.

1. INTRODUCTION

The b -value of the Gutenberg–Richter (GR) relationship (Ishimoto & Iida 1939; Gutenberg & Richter 1944) is a measure of the relative frequency of large versus small earthquakes. It has been proposed that it changes with differential stress (Mogi 1962; Scholz 1968), focal depth (Gerstenberger *et al.* 2001) and tectonic framework (Schorlemmer *et al.* 2005; Petruccioli *et al.* 2019) and monitoring its evolution over time has received increased amounts of attention for the purpose of earthquake forecasting (Gulia & Wiemer 2019; Gulia *et al.* 2024).

However, genuine and artificial b -value changes are often difficult to distinguish, due to the many factors affecting calculations, such as magnitude binning and measurement errors, sample size, width of the range covered by magnitude data, mixing of magnitude types, changes in seismic network detection, under or overestimation of completeness magnitude (M_c) and the presence of non-tectonic events (Kagan 1999; Wiemer & Baer 2000; Wiemer & Wyss 2000;

Marzocchi & Sandri 2003; Amorè *et al.* 2010; Gulia & Gasperini 2021; Geffers *et al.* 2022). Each of these points may interfere with b estimation, producing significant bias and misunderstood results, which have been widely discussed and quantified by the scientific community.

In particular, M_c is a crucial parameter for the estimation of the b -value: a too low estimated value of M_c gives a biased estimate of b , since we are attempting to fit a power-law model to non-power-law data. On the other hand, a too high value of M_c throws away legitimate magnitude data, affecting accuracy and precision in estimating b : if $b = 1$, an increase of 0.5 for M_c means a reduction of a factor of approximately 3 in sample size, the consequent increase of a factor close to 1.7 in b -value error (Aki 1965; Shi & Bolt 1982; Amorè *et al.* 2010) and more biased estimation of b , especially if the dynamic range (i.e. the range covered by magnitude data, after the M_c evaluation) is less than 2.0 (Geffers *et al.* 2022, 2023), as is often the case when the catalogue is divided into spatiotemporal cells.

Despite the considerable interest and intense debate on how to correctly estimate the parameters of the GR relationship (including Mc), a great variety of methods still remain in current use, even after years of attempted improvement, revealing that the question is far from being answered (Aki 1965; Utsu 1965; Bender 1983; Tinti & Mulargia 1987; Ogata & Katsura 1993; Wiemer & Wyss 2000; Cao & Gao 2002; Woessner & Wiemer 2005; Amorèse 2007; Leptokaropoulos *et al.* 2013; Tormann *et al.* 2014; Han *et al.* 2015; Corral & González 2019; Hermann & Marzocchi 2021; van der Elst 2021; Tinti & Gasperini 2024 and many others). The coexistence of such many methods generates the dissemination of ambiguous, if not conflicting, results on these topics by the scientific community. An example of this is the many valuable papers addressing the b -value computations and variations in the Central Italy region, especially with reference to the 2009 L'Aquila and the 2016 Central Italy sequences (e.g. Papadopoulos *et al.* 2010; De Gori *et al.* 2012; Sugan *et al.* 2014; Gulia *et al.* 2016; Gulia & Wiemer 2019; Hermann & Marzocchi 2021; van der Elst 2021; Lacidogna *et al.* 2023; Lombardi 2023, 2024 and many others). The conclusions of these papers go from the denial of any b -value variation, both out and during seismic sequences, to the statement that b changes may be used as precursors of near future large events. Certainly, several factors may be responsible for these different findings, such as the choice of data and of methods used for processing (Rhoades 1996; Marzocchi & Sandri 2003; Mignan & Woessner 2012; Kamer & Hiemer 2015; Leptokaropoulos *et al.* 2018; Marzocchi *et al.* 2020; Geffers *et al.* 2022, 2023) or criteria applied to detect possible b -value changes (Shi & Bolt 1982; Smyth & Mori 2009; Amorèse *et al.* 2010; Hermann *et al.* 2022), since statistical tests are rarely implemented to assess the significance of the observed differences. Less attention has been given to the impact that the choice of parameter (both Mc and b) estimation method has on the results (Woessner & Wiemer 2005; Zhou *et al.* 2018; Pavlenko & Zavyalov 2022), and this paper intends to provide a contribution in this respect.

Comparing and testing the published b -value estimation methodologies is difficult because different studies use different data, regions and timescales and, above all, because they are based on different hypotheses, going from the empirical log-linearity of the cumulative number of events to well-defined probability distributions for magnitudes, such as the exponential or the geometric ones. Therefore, each of these methods not only allows us to quantify the aleatoric uncertainty in the earthquake magnitude distribution, due to the inherent stochasticity of underlying physical processes, but also represents the perception and the knowledge of the problem by individual modellers. Therefore, the treatment of and the sensitivity analysis to this epistemic uncertainty play crucial roles in understanding the results of individual analyses. The best way to judge how this subjective knowledge affects results is to compare different estimation methods under a controlled environment, where pre-defined conditions on data are given to all modellers.

This study presents a collaborative experiment in which 19 different algorithms for estimating Mc and b parameters are compared. We quantify how much the estimates vary across different procedures and compare these differences with the uncertainties in individual estimates, provided by each algorithm, to quantify how much the choice of an estimation method (for both Mc and b) may affect the b -value computation and the interpretation of possible b -value variations. Specifically, in this paper we (1) quantify the sensitivity of Mc and b estimates (marked by $\overline{Mc}$ and $\overline{b}$, respectively, in the following) to the adopted estimation method; (2) investigate how much $\overline{Mc}$ affects the b -value estimation and (3) compare the b -value sensitivity with the uncertainty in $\overline{b}$, provided by each method.

This study aimed to compare the different methodologies but did not employ any ranking strategy. Moreover, our work does not intend to include all possible published methodologies: the participants were left free to adopt the algorithms with which they are more familiar. Finally, it does not face the question of possible spatiotemporal variation in b and of related physical interpretation.

The paper is structured as follows: (1) introduce the data and the details of the experiment; (2) provide a short description of the methodologies used to determine the Mc and b estimates; (3) compare and discuss the results, based on the aims previously described.

2. DATA AND DEFINITION OF THE SENSITIVITY EXPERIMENT

In this study, we analyse the data collected in the Italian Seismic Bulletin (BSI, Bollettino Sismico Italiano; <http://terremoti.ingv.it/en/iside>; <http://terremoti.ingv.it/en/bsi>) of the Istituto Nazionale di Geofisica e Vulcanologia (INGV), which refer to earthquakes occurring inside the Central Italy (CI) region (Fig. 1). This zone was struck by two important seismic sequences: the 2009 L'Aquila (main event 2009 April 6, ML5.9, M_w 6.1) and the 2016–2017 Central Italy (main event 2016 October 30, ML6.1, M_w 6.5) sequences (Fig. 1).

We identify a collection [42.6°–43.18°N, 12°28'–13°56'E] and a target [42°12'–43°12'N, 12°36'–13°48'E] (marked by dotted lines in Fig. 1) regions for our analysis. The algorithms select the earthquakes from the larger collection region as input data for their estimations within the testing region to minimize the boundary effect. We remove from the BSI data both the artificial events and the natural events for which only the M_d value is available (607 events above ML1.0 by the M_d /ML relation of Gasperini *et al.* 2013), in order to select data with homogeneous magnitudes. We select 150 503 events in the collection region, with depths less than 40 km and local magnitudes (ML) above 1.0, from 2005 April 16 (when the completely reorganized National Seismic Network of the INGV came into operation; Amato & Mele 2008) to 2021 December 31. We divide the selected catalogue into overlapping (in time and space) subsamples, for which any modeller must provide Mc and b . The subsampling method consists of defining temporally intersecting samples of $N_{tw} = 5000$ events, moved forward by $N_s = 500$ events through the catalogue. Calculations are made by dividing the CI target region by a $0.05^\circ \times 0.05^\circ$ grid and computing the b -value for the events occurred in an area of $R = 10$ km of radius around each node of the grid. The outermost cells have the central nodes on the border of the target region. In this way, the Mc and b -value maps are not equally distributed in time, since they follow the temporal evolution of seismicity. The sampling method identified 292 temporal windows and 525 nodes, for a total of 153 300 possible joint Mc and b estimations. In the [Supplementary Information](#) we provide the list of the 525 grid nodes and the list of the occurrence times of the first and the last event for the 292 temporal windows.

3. OVERVIEW OF METHODOLOGIES

In this study, we compare 19 different algorithms providing Mc and b evaluations. They use five different methods to estimate Mc and four methods to estimate b (or slight modifications of these methods). All these methods have been previously published; therefore, we provide a brief description of each of them in Appendix A and an overview in Table 1. In summary, the 19 algorithms estimate Mc by using (a) the Maximum Curvature method (Wiemer & Wyss 2000), that does not require any hypothesis on the probability distribution

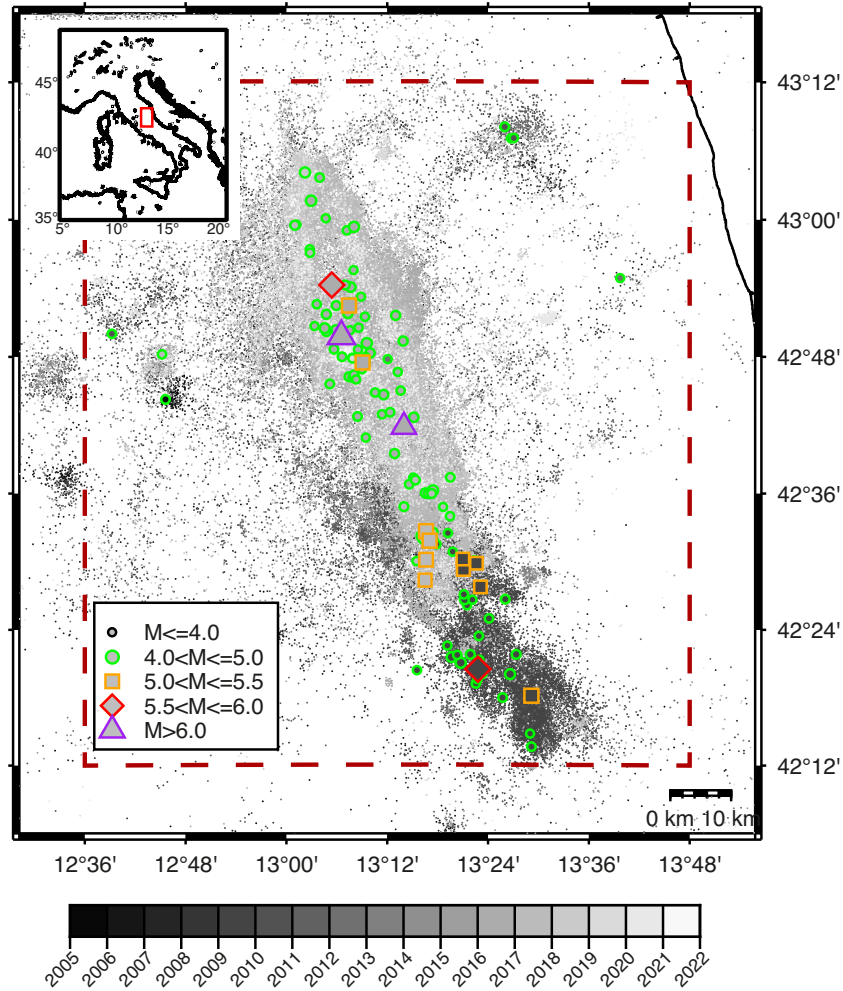


Figure 1. Map of the events that occurred in the Central Italy collection region, from 2005 April 16 to 2021 December 31, above 40 km in depth and with magnitude $M_L \geq 1.0$. The dashed lines delimit the target area adopted in this study. The symbol size is scaled with magnitude. The grey colour is scaled with the occurrence time of the events.

of magnitudes, or (b) by applying a goodness-of-fit test on magnitude data for an exponential (Goodness of Fit, Wiemer & Wyss 2000; Lilliefors test, Hermann & Marzocchi 2021; Variability Coefficient, Godano *et al.* 2023) or a geometric distribution (Normalized Distance Test, Lombardi 2021). Moreover, most of algorithms adopt the maximum likelihood method to estimate b -values, under the hypothesis of an exponential (Utsu 1965) or a geometric (Tinti & Mulargia 1987) distribution for magnitudes, or a double exponential distribution (i.e. Laplace) for differences in magnitude between successive earthquakes (van der Elst 2021). Only one algorithm adopted a robust regression analysis (Han *et al.* 2015), to check the GR relationship (meant as linearity of the cumulative number of events on a logarithmic scale), without a specific assumption for the magnitude distribution (see Table 1 and Appendix A for details).

To identify such many methods, we build two types of acronyms. The complete acronyms use M (for model) + incremental number (from 1 to 19) + abbreviation for M_c estimation approach (e.g. MaxC or GoF95 per cent) + abbreviation for b estimation approach (e.g. gMLE or B+) and are mostly used in figures and tables, where possible, to make them clearer. The short acronyms (M1, M2, M3, etc.) are used in the text, to preserve the reading flow, and when we cannot use the complete acronyms (see Table 1, for the association between the two types of acronyms).

All the 19 algorithms differ from each other, since any two of them applied different procedures for estimating M_c or b (or both), adapted different minimum sample criteria for obtaining a reliable estimate (see Table 1, last column), used different details of the adopted methodology (such as the significance level of statistical tests) or, finally, applied a different code. All working groups, which are responsible for one or more algorithms, worked autonomously and were free to apply their own practices and to select the samples satisfying the constraints required by their method; therefore, the number of joint M_c and b estimations changes among the methods. Moreover, for each time window, the modellers may choose to estimate a spatially invariant (i.e. on data of the whole region) or a spatially varying (i.e. for each node) M_c . Therefore, the methods used to evaluate M_c may be divided into two classes. The first one, G1, is composed of six algorithms, estimating a different M_c for each cell of the grid and each time frame (M5, M6, M7, M8, M10, M19), whereas the nine models estimating a spatial invariant M_c (M9, M11, M12, M13, M14, M15, M16, M17, M18) belong to the second group, G2. These last models estimate a different M_c for each temporal window, but valid for the whole region. The algorithms M1, M2, M3 and M4, which are based on the b -positive method (van der Elst 2021), do not evaluate any M_c , since M2 and M3 use all the data while M1 and M4 discard the data below the

Table 1. Overview of 19 Mc and b estimation algorithms

	Authors	Mc Class	Acronym	Estimation method Mc	Estimation method b	Reliability criteria
M1	WG1 Gulia L., Gasperini P., Lolli B., Vannucci G., Biondini E.		M1_NoMc_B+	no Mc	B positive (van der Elst 2021)	$N_{th} \geq 100$
M2	WG1 Gulia L., Gasperini P., Lolli B., Vannucci G., Biondini E.		M2_NoMc_B+	no Mc	B positive (van der Elst 2021)	$N \geq 100$
M3	WG1 Gulia L., Gasperini P., Lolli B., Vannucci G., Biondini E.		M3_NoMc_B+	no Mc	B positive (van der Elst 2021)	$N_{th} > 100$
M4	WG1 Gulia L., Gasperini P., Lolli B., Vannucci G., Biondini E.		M4_NoMc_B+	no Mc	B positive (van der Elst 2021)	$N_{th} \geq 100$
M5	WG1 Gulia L., Gasperini P., Lolli B., Vannucci G., Biondini E.	G1	M5_MaxC0.2.ceMLE	MaxC+0.2 (Wiener and Wyss 2000; Woessner and Wiener 2005)	corrected exponential MLE (Aki 1965; Ulsu 1965)	$N_{Mc} \geq 50$
M6	WG1 Gulia L., Gasperini P., Lolli B., Vannucci G., Biondini E.	G1	M6_rMaxC_RLR	'reinforced' MaxC (Wiener and Wyss 2000)	robust linear regression (Han <i>et al.</i> 2015)	$N_{Mc} \geq 75$
M7	WG2 Murru M., Falcone G., Montuori C., Console R.	G1	M7_BGoF90%.BeeMLE	Bootstrap Goodness of fit(10 per cent significant level) (Wiener and Wyss 2000, 2002; Woessner and Wiener; 2005)	Bootstrap corrected exponential MLE (Aki 1965; Ulsu 1965; Woessner and Wiener; 2005)	$N_{Mc} \geq 50$
M8	WG3 De Gori P., Lucente F.P.	G1	M8_GoF95%.ceMLE	Goodness of fit(5 per cent significant level) (Wiener and Wyss 2000, 2002)	corrected exponential MLE (Aki 1965; Ulsu 1965)	$N \geq 250$
M9	WG4 Taroni M., Spassiani I.	G2	M9_LT_ueMLE	Lilliefors test(10 per cent significant level) (Hermann & Marzocchi 2021)	unbiased corrected exponential MLE (Marzocchi <i>et al.</i> 2020)	$N_{Mc} \geq 50$; $Mc \geq 1.5$
M10	WG5 Lombardi A.M.	G1	M10_NDT_gMLE	Normalized distance test(1 per cent significant level) (Lombardi 2021, 2023)	geometric MLE (Tinti and Mulargia 1987)	$N_{Mc} \geq 50\Delta M \geq 2.0$
M11	WG6 Garcia A., Faenza L., Sandri L.	G2	M11_GoF95%.ceMLE	Goodness of fit(5 per cent significant level) (Wiener and Wyss 2000, 2002)	corrected exponential MLE (Aki 1965; Ulsu 1965)	$N_{Mc} \geq 50$
M12	WG6 Garcia A., Faenza L., Sandri L.	G2	M12_GoF95%.gMLE	Goodness of fit(5 per cent significant level) (Wiener and Wyss 2000, 2002)	geometric MLE (Tinti and Mulargia 1987)	$N_{Mc} \geq 50$
M13	WG6 Garcia A., Faenza L., Sandri L.	G2	M13_GoF90%.ceMLE	Goodness of fit(10 per cent significant level) (Wiener and Wyss 2000, 2002)	corrected exponential MLE (Aki 1965; Ulsu 1965)	$N_{Mc} \geq 50$
M14	WG6 Garcia A., Faenza L., Sandri L.	G2	M14_GoF90%.gMLE	Goodness of fit(10 per cent significant level) (Wiener and Wyss 2000, 2002)	geometric MLE (Tinti and Mulargia 1987)	$N_{Mc} \geq 50$
M15	WG6 Garcia A., Faenza L., Sandri L.	G2	M15_MaxC.ceMLE	MaxC (Wiener and Wyss 2000)	corrected exponential MLE (Aki 1965; Ulsu 1965)	$N_{Mc} \geq 50$
M16	WG6 Garcia A., Faenza L., Sandri L.	G2	M16_MaxC_gMLE	MaxC (Wiener and Wyss 2000)	geometric MLE (Tinti and Mulargia 1987)	$N_{Mc} \geq 50$
M17	WG6 Garcia A., Faenza L., Sandri L.	G2	M17_MGoF_ceMLE	Modified Goodness of fit (Leptokaropoulos <i>et al.</i> 2013)	corrected exponential MLE (Aki 1965; Ulsu 1965)	$N_{Mc} \geq 50$
M18	WG6 Garcia A., Faenza L., Sandri L.	G2	M18_MGoF_gMLE	Modified Goodness of fit (Leptokaropoulos <i>et al.</i> 2013)	geometric MLE (Tinti and Mulargia 1987)	$N_{Mc} \geq 50$
M19	WG7 Tranelli A., Convertito V., Godano C.	G1	M19_CV_ceMLE	Variability Coefficient (CV) (Godano <i>et al.</i> 2023)	correctedexponential MLE (Aki 1965; Ulsu 1965)	$N \geq 100$; $N_{Mc} \geq 50\Delta M \geq 1.5$

Table 2a. Proportion of bins with $\Delta McC = 0$, $\Delta McC \leq 0.1$ and $\Delta McC \leq 0.2$, in order, for all couples of models in group G1. The pairs of models for which $\Delta McC > 0.2$ for more than 50 per cent of bins are marked in boldface.

	M5	M6	M7	M8	M10	M19
M5		2 per cent 3 per cent 90 per cent	11 per cent 25 per cent 70 per cent	23 per cent 34 per cent 83 per cent	17 per cent 26 per cent 61 per cent	4 per cent 21 per cent 59 per cent
M6		—	35 per cent 53 per cent 93 per cent	27 per cent 37 per cent 69 per cent	7 per cent 11 per cent 33 per cent	31 per cent 45 per cent 81 per cent
M7		—	—	4 per cent 24 per cent 79 per cent	5 per cent 9 per cent 32 per cent	32 per cent 49 per cent 83 per cent
M8		—	—	—	11 per cent 17 per cent 48 per cent	13 per cent 31 per cent 61 per cent
M10		—	—	—	—	3 per cent 11 per cent 35 per cent

magnitude of maximum curvature (Wiemer & Wyss 2000), which, however, has not to be intended as the true completeness magnitude Mc . The algorithms M1, M2, M3 and M4 do not evaluate any Mc , since they are based on the b -positive method (van der Elst 2021), which does not require a preliminary estimate of Mc .

The only restriction given to workers was not to remove or modify the original data. In this way, the b -values estimated by two algorithms that provide the same $\overline{Mc}$ are computed exactly on the same data and their possible differences are ascribable only to differences in applied methodologies.

4. DESCRIPTIVE ANALYSIS OF THE RESULTS

Some general statistics on the results provided by all methods is summarized in Fig. 2. The first information regards the number of bins (on a total of 153 300) for which each algorithm provided joint $\overline{Mc}$ and $\bar{b}$ values; this differs among algorithms since these adopted different reliability criteria to select the samples (Table 1). The largest (41 024) and the smallest (13 311) numbers of joints $\overline{Mc}$ and $\bar{b}$ are given by M14 and M10, respectively (this last considers only subsamples with a magnitude range $\Delta M \geq 2.0$; Fig. 2a). The second smallest number of estimates (13 582) is provided by M8, since it analyses only bins with at least 250 events.

The second information given in Fig. 2(b) is the number of events above $\overline{Mc}$, $N[\overline{Mc}]$, on which the b -value is computed, that is crucial for the uncertainty about $\bar{b}$ values. $N[\overline{Mc}]$ has a (right) long-tailed distribution, with a median ranging from 123 (M9) to 367 (M8) and reaching a few thousand events. The smallest variability of $N[\overline{Mc}]$ are observed for M9 and M10.

Fig. 2(c) shows the variability of Mc evaluations: M19 (for G1), M13 and M14 (for G2) provide, on average, lower values of $\overline{Mc}$, with a median of 1.1 (Fig. 2c). No significant variation in the $\overline{Mc}$ distribution was observed, if we selected only bins with $N \geq 500$ and $\Delta M \geq 2.0$ (Fig. 2c, dashed lines), for which $\bar{b}$ values are more robust (Geffers *et al.* 2022, 2023). M9 and M10 are exceptions due to the higher resolution of the statistical tests, on which they are both based, for more robust samples. More symmetric distributions are obtained for $\bar{b}$ (Fig. 2d), with median values ranging from 1.1 (M13, M14 and M19) to 1.35 (M1, M2, M3, M4). M10 and M19 provide the smallest 90 per cent confidence intervals for $\bar{b}$, due to lower limits (2.0 and 1.5, respectively) imposed on the magnitude range, ΔM (Table 1). Finally, the distribution of b -values provided by the b -positive strategy is shifted towards larger values

than those provided by the other algorithms (Fig. 2d). All methods provide less variable $\bar{b}$ for samples with $N \geq 500$ and $\Delta M \geq 2.0$, since the overestimation effect, recognized for samples with small sizes and small dynamic ranges (Geffers *et al.* 2022, 2023), is minimized.

In the Appendix, Figs A1 and A2 show the behaviour of $\bar{b}$ as a function of $N[\overline{Mc}]$ and ΔM , for all the algorithms. From these figures, we infer that the $\bar{b}$ exhibits a decreasing variability with increasing $N[\overline{Mc}]$, as expected, but around a median value, which differs among the algorithms. Moreover, the already recognized overestimation of $\bar{b}$ for samples with small ΔM (Geffers *et al.* 2022, 2023) is associated with all methods, including the b -positive approach (note that M10 and M19 exclude samples with $\Delta M < 2.0$ and $\Delta M < 1.5$, respectively).

5. COMPARISON OF MC ESTIMATIONS

First, we compare the Mc values provided by different methods. Specifically, we separate the algorithms in the classes G1 and G2, since they use different sets of data, and we investigate both the overall differences in $\overline{Mc}$ for all algorithms (ΔMcO) and the differences for all couples of algorithms (ΔMcC). The algorithms M1, M2, M3 and M4 are excluded from this analysis, since they do not provide any estimation of Mc .

There are 6005 spatiotemporal bins for which all 19 algorithms provide a joint Mc and b estimation. When we separately consider the algorithms in G1 and G2, the numbers of bins in common are 6682 and 20 877, respectively. The maximum difference, ΔMcO , for a bin j in common with the Nm (6 and 9 for G1 and G2, respectively) algorithms, is defined by

$$\Delta McO(j) = \max [Mc_{ij}, i = 1, \dots, Nm] - \min [Mc_{ij}, i = 1, \dots, Nm], \quad (1)$$

where Mc_{ij} is the value of Mc for the j -th bin, estimated by the i -th model. ΔMcO represents how much the estimation of Mc may vary among the different methods. We prefer to compute the min-max range of $\overline{Mc}$ values, instead of the interquartile ranges, to better represent the real heterogeneity of methods. In this way, an algorithm that deviates from all the others is not considered as an exception, or as a ‘less correct’ approach. We found that ΔMcO ranged from 0.1 up to 2.1, for algorithms in G1, and from 0.2 up to 1.6 for algorithms in G2 (Fig. 3a). The median and the 90° percentile of ΔMcO were 0.5 and 0.9, respectively, for both G1 and G2. ΔMcO values greater than 1.0 are observed in the most intense periods of the 2009 and 2016 sequences, for both G1 and G2; the remaining values of ΔMcO do not show specific temporal features and have a time-independent median value of 0.4–0.5 (Fig. 3b). The spatial distribution of the median values of ΔMcO for the algorithms in G1 reveals larger values in the northern area, hit by the 2016 Norcia earthquake, than in the southern area, affected by the 2009 L’Aquila sequence (Figs 3b and c). ΔMcO depends on sample size, N , when it ranges from 500 to 1500, and on maximum magnitude, $M_{\max}$, when it is between 3.0 and 4.5 (Fig. 3d): the Spearman’s rank correlation test (Gibbons & Chakraborti 2020) provides a p -value < 0.01 for both correlations. In the Supplementary Information we show four cases with different ΔMcO values (Fig. S1).

In the second type of analysis, we compare all couples of algorithms, separately for G1 and G2. Specifically, we compute the relative difference by

$$\Delta McC(j) = Mc_{1j} - Mc_{2j}, \quad (2)$$

Table 2b. Proportion of bins with $\Delta McC = 0$, $\Delta McC \leq 0.1$ and $\Delta McC \leq 0.2$, in order, for all couples of models in group G2. The pairs of models for which $\Delta McC > 0.2$ for more than 50 per cent of bins are marked in boldface.

	M9	M11	M12	M13	M14	M15	M16	M17	M18
M9		2 per cent 10 per cent 42 per cent	2 per cent 10 per cent 42 per cent	0 per cent 0 per cent 5 per cent	0 per cent 0 per cent 5 per cent	0.5 per cent 2 per cent 14 per cent	0.5 per cent 2 per cent 14 per cent	36 per cent 37 per cent 73 per cent	36 per cent 37 per cent 73 per cent
M11		–	100 per cent 100 per cent 100 per cent	0.7 per cent 13 per cent 89 per cent	0.7 per cent 13 per cent 89 per cent	17 per cent 41 per cent 92 per cent	17 per cent 41 per cent 92 per cent	15 per cent 41 per cent 94 per cent	15 per cent 41 per cent 94 per cent
M12		–	–	0.7 per cent 13 per cent 89 per cent	0.7 per cent 13 per cent 89 per cent	17 per cent 41 per cent 92 per cent	17 per cent 41 per cent 92 per cent	15 per cent 41 per cent 94 per cent	15 per cent 41 per cent 94 per cent
M13		–	–	–	100 per cent 100 per cent 100 per cent	46 per cent 67 per cent 98 per cent	46 per cent 67 per cent 98 per cent	0 per cent 2 per cent 30 per cent	0 per cent 2 per cent 30 per cent
M14		–	–	–	–	46 per cent 67 per cent 98 per cent	46 per cent 67 per cent 98 per cent	0 per cent 2 per cent 0 per cent	0 per cent 2 per cent 30 per cent
M15		–	–	–	–	–	100 per cent 100 per cent 100 per cent	1 per cent 17 per cent 54 per cent	1 per cent 17 per cent 54 per cent
M16		–	–	–	–	–	–	1 per cent 17 per cent 54 per cent	1 per cent 17 per cent 54 per cent
M17		–	–	–	–	–	–	–	100 per cent 100 per cent 100 per cent

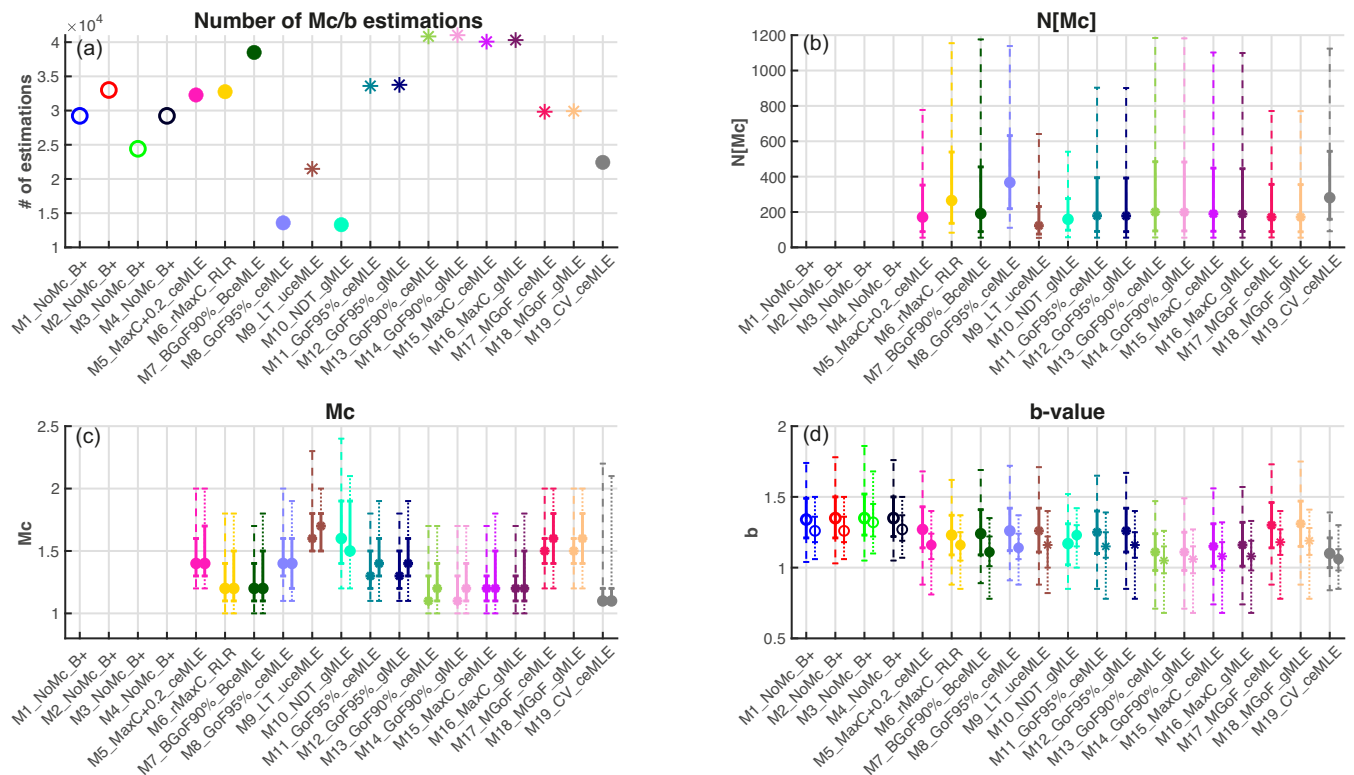


Figure 2. Descriptive analysis of the Mc and b estimations provided by the 19 algorithms. The filled circles and stars mark the methods in groups G1 and G2, respectively. The open circles refer to methods (M1–M4) which do not estimate any Mc . (a) Number of spatiotemporal bins (the total number of bins is 153 300) for which each method estimates both Mc and b . (b) Uncertainty of $N[Mc]$ for all methods (solid lines). The filled circles mark the median values. The solid and the dashed lines mark the interquartile range and the 90 per cent confidence interval, respectively. This information is not plotted for algorithms M1–M4, since they do not estimate any Mc . (c) Uncertainty of Mc estimations. The dotted and dashed lines mark the 90 per cent confidence interval for all bins and to those with $N > 500$ and $DM > 2.0$, respectively. The solid lines mark the interquartile ranges. This information is not plotted for algorithms M1–M4, since they do not estimate any Mc . (d) The same as (c) but for b -value estimations.

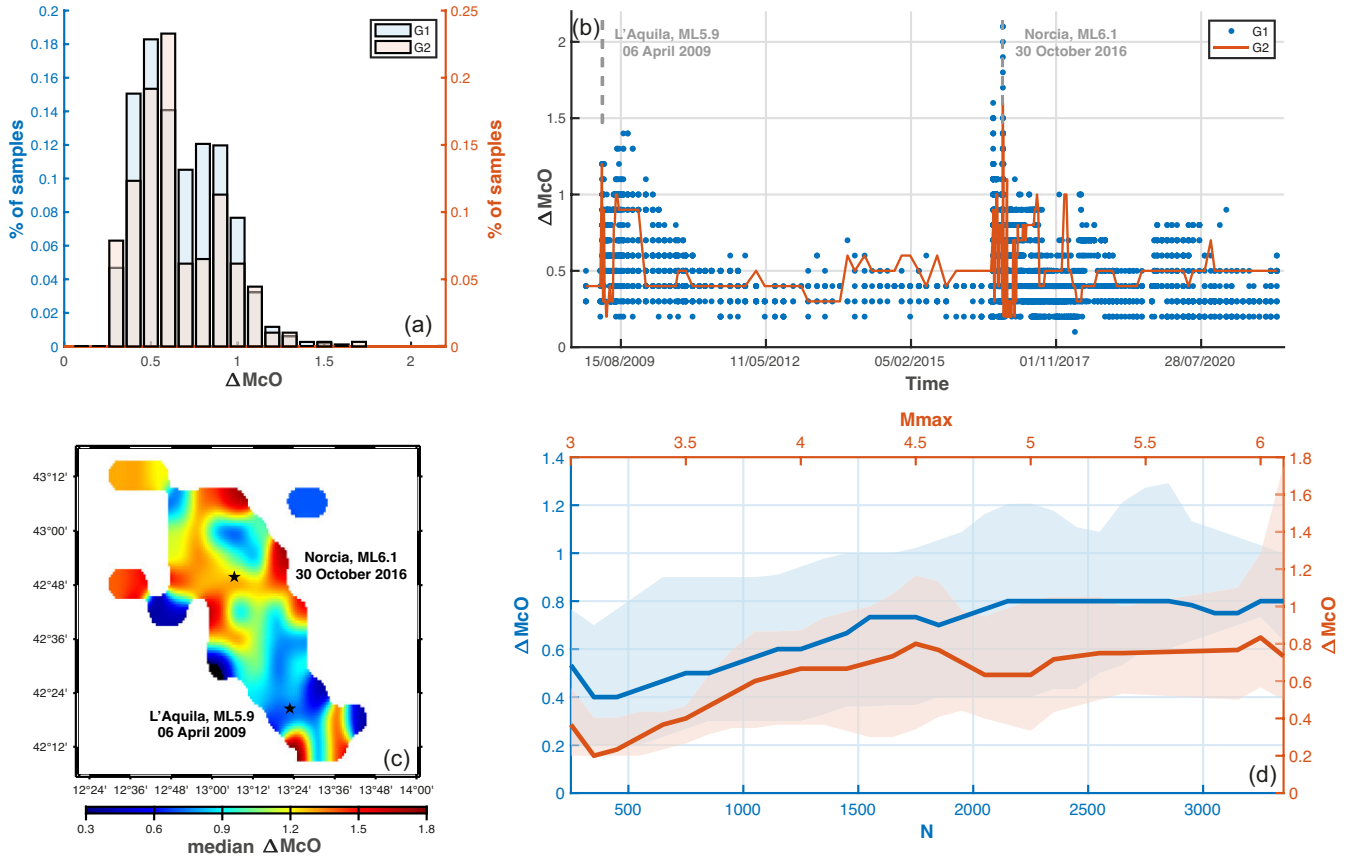


Figure 3. Overall differences in Mc values, ΔMcO , among all methods, but M1–M4, for the 6005 bins in common (i.e. for which all algorithms provide Mc and b estimations). (a) Histograms of ΔMcO values for algorithms in G1 and G2. (b) Plot of ΔMcO with time, for the algorithms in G1 (circles) and G2 (solid line). The dashed lines mark the occurrence times of the main shocks in the 2009 and 2016 sequences. (c) Map of median values of ΔMcO . The stars mark the locations of the main shocks in the 2009 and 2016 sequences. (d) Plot of ΔMcO values versus the sample size N (blue) and the maximum magnitude M_{max} (orange). The shadowed areas mark the 90 per cent confidence intervals. The solid lines mark the median values.

where Mc_{1j} and Mc_{2j} are the values of Mc provided by two algorithms, for the j -th bin. This variable quantifies how much the two compared approaches differ in time and space. Tables 2a,b list the percentage of bins in common, for which any couple of models provides equal Mc values ($\Delta McC = 0$) and differences equal to 0.1 and 0.2. All the equal values are obtained by only comparing identical algorithms, that is, in the following four cases: (1) M11 and M12; (2) M13 and M14; (3) M15 and M16 and (4) M17 and M18. The largest proportion for the remaining cases is equal to 45–46 per cent, which is reached by comparing the MaxC (Models M15 and M16) and the GoF90 per cent (M13 and M14) algorithms in G2. For all the other pairs of models, less than 36 per cent of the bins had equal Mc values. Several pairs of models in G2 always provide different estimations, such as the GoF90 per cent (M13 and M14) algorithm compared with the MGoF (M17 and M18) procedure or with LT (M9). Finally, equal Mc estimations are obtained for less than 1 per cent of cases, by comparing the following pairs of algorithms in G2: (a) MaxC (M15 and M16) and LT (M9); (b) GoF95 per cent (M11 and M12) and GoF90 per cent (M13 and M14). For 14 (22) pairs of algorithms (on a total of 51 pairs) more than 50 per cent (30 per cent) of bins have $\Delta McC > 0.2$. These cases occur when one model is M9 or M10 and when we compare M13/M14 with M18/M19.

Fig. 4 shows the confidence intervals of ΔMcC , for all possible pairs of algorithms. By comparing algorithms in G1, (Fig. 4a) we find that for the 90 per cent of the bins (a) rMaxC (M6) provides a

lower Mc , than do M5, M8 and M10; (b) M10 estimates a larger Mc than does any of the other algorithm, except for M5 and M19; (c) GoF90 per cent (M7) estimates smaller values than does GoF95 per cent (M8) and (d) the comparison of M19 with any other algorithm provides both negative and positive values of ΔMcC . The width of the 90 per cent confidence interval of ΔMcC is 0.5 at most, except if one of two compared models is M10 or M19. By comparing algorithms in G2 (Fig. 4b) we found that for the 90 per cent of the bins (a) GoF95 per cent (M11 and M12) provides a larger Mc than do GoF90 per cent (M13 and M14) and MaxC (M15 and M16); (b) MGoF (M17 and M18) estimates larger Mc than do MaxC, GoF90 per cent and GoF95 per cent and (c) M9 estimates larger Mc than do all the other algorithms. The width of the 90 per cent confidence interval of ΔMcC is 0.3, at most, except if one of two compared models is M9. By all this information, we can construct two diagrams, for G1 (Fig. 4c) and G2 (Fig. 4d), revealing the possible systematic sorting of Mc provided by the different methods. An increasing order may be assigned to Mc provided by (a) rMaxC and MaxC + 0.2 in G1, (b) GoF90 per cent and GoF95 per cent in G1 and (c) GoF90 per cent, GoF95 per cent and MGoF in G2.

Further details about the differences and similarities among the methods used are given in Movie S1 in the Supplementary Information, which shows a density plot for the two sets of Mc values, provided by all couples of methods in G1 and G2.

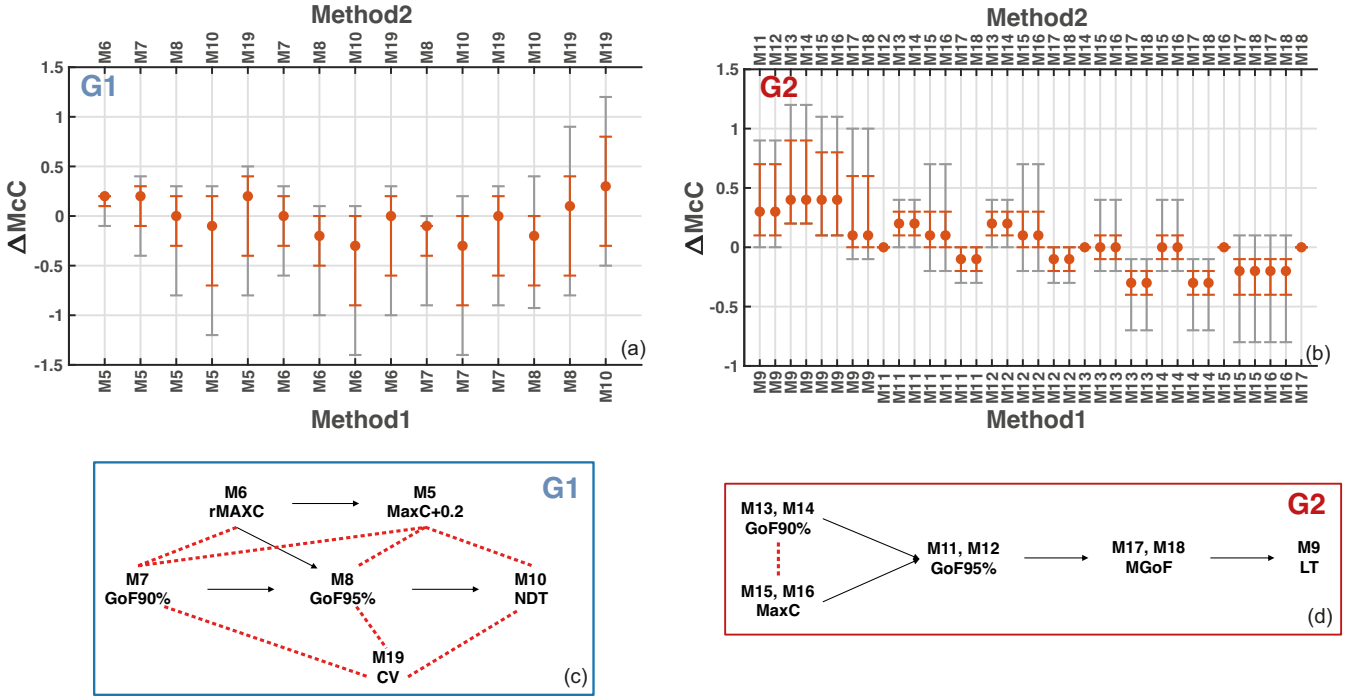


Figure 4. Relative differences in Mc values, ΔMcC , for all couples of algorithms in G1 and G2. (a) 99 per cent (grey lines) and 90 per cent (red lines) confidence intervals for all couples of algorithms in G1. The bottom and top x-axes show the algorithm1 and algorithm 2, respectively. (b) The same as (a) but for G2. (c) Possible sorting of algorithms in G1. The black arrows mark the couples of algorithms for which a sorting criterion may be specified (a method provides a lower Mc with respect to the other one, according to the arrow direction), based on the 90 per cent confidence interval of ΔMcC . The dotted red lines mark the couples of models for which ΔMcC values are both positive and negative. (d) The same as in (c) but for G2.

6. COMPARISON OF b ESTIMATIONS

We compare the b -value estimations provided by different methods, similarly to what done for Mc . The overall variability over the $Nm = 19$ estimated b -values, ΔbO , is defined by

$$\Delta bO(j) = \max[\bar{b}, i = 1, \dots, Nm] - \min[\bar{b}, i = 1, \dots, Nm] \quad (3)$$

where $\bar{b}_{ij}$ is the value of b for the j -th bin, estimated by the i -th model. This variable quantifies the variability of b -value estimation among the 19 algorithms. Similarly to the analysis of Mc variability, we prefer to compute the min-max range of b evaluations, to better represent the real heterogeneity of methods.

The values of ΔbO for the 6005 bins, which are common among the 19 algorithms, are shown in Fig. 5. These values ranged from 0.09 to 1.0, with a median of 0.3 and a 90th percentile equal to 0.5 (Fig. 5a). Again, the largest differences ($\Delta bO > 0.7$) are observed in the most intense periods of the 2009 and 2016 seismic sequences, whereas the remaining, however important, values of ΔbO do not show known temporal features (Fig. 5b). This overall large variability is not due to few ‘exceptional’ methods: by excluding the 20 per cent (the two largest and two smaller) of b -values, the overall variability is still large (Fig. 5b, green points). As shown for ΔMcO , the median of ΔbO is greater in the Norcia region, where the main shock of the 2016–2017 seismic sequence occurred, than in the L’Aquila area, but the spatial patterns of the two variables are different (the correlation coefficient between the medians of ΔMcO and ΔbO is equal to 0.03; Figs 3c and 5c). Moreover, the median ΔbO increases, from approximately 0.25 to 0.45, for $M_{\max} > 5.0$ (Fig. 5d), revealing greater disagreement among algorithms in the case of strong events, when the catalogues are highly heterogeneous. In the [Supplementary Information](#), we show two cases with different values of ΔbO (Fig. S2).

The difference in b -values provided by all the pairs of algorithms is quantified by the variable ΔbC . It does not allow us, in most cases, to establish an order between two or more algorithms, since ΔbC assumes both positive and negative values (Fig. 6). A few cases are exceptions: M13 and M14 mostly (i.e. for more than 90 per cent of common bins) estimate lower b -values than do the algorithms M1, M2, M3, M4 (all using the b -positive strategy), M8, M11, M12, M17 and M18: for these cases ΔbC is positive or negative for the 90 per cent of common bins (solid grey lines in Fig. 6). Moreover, M19 estimates lower b -values than do the b -positive algorithms.

We found different types of relationship between the b -values provided by any two of 19 algorithms (shown partially in Fig. 7 and in details in [Movie S2](#) in the [Supplementary Information](#)). If two algorithms provide similar results, the joint distribution of their estimated b -values is symmetric with respect to the bisector (pairs M1/M4, Fig. 7a), whereas, in other cases, ΔbC may have a bimodal distribution (Fig. 7b). As mentioned above, in some cases, one algorithm estimates lower b -values with respect to the other (Fig. 7c). Finally, some algorithms show a saturation effect when compared with some others (Fig. 7d). The type of relationship for all pairs of algorithms may be found in [Movie S2](#) of the [Supplementary Information](#).

A final analysis was performed to investigate the relationship between $\bar{b}$ and $\bar{Mc}$ differences. First, we analysed the variation of $\bar{b}$ among methods independently by $\bar{Mc}$, by investigating the bins for which two algorithms provide identical $\bar{Mc}$. We found that the b -values estimated by the exponential or geometric MLE approaches (M5, M8, M9, M10, M11–M18) are very similar (red solid lines in Fig. 6); therefore, we infer that, in these cases, the b -value differences are mostly due to differences in Mc (compare grey and red solid lines in Fig. 6). Larger ΔbC are observed if one

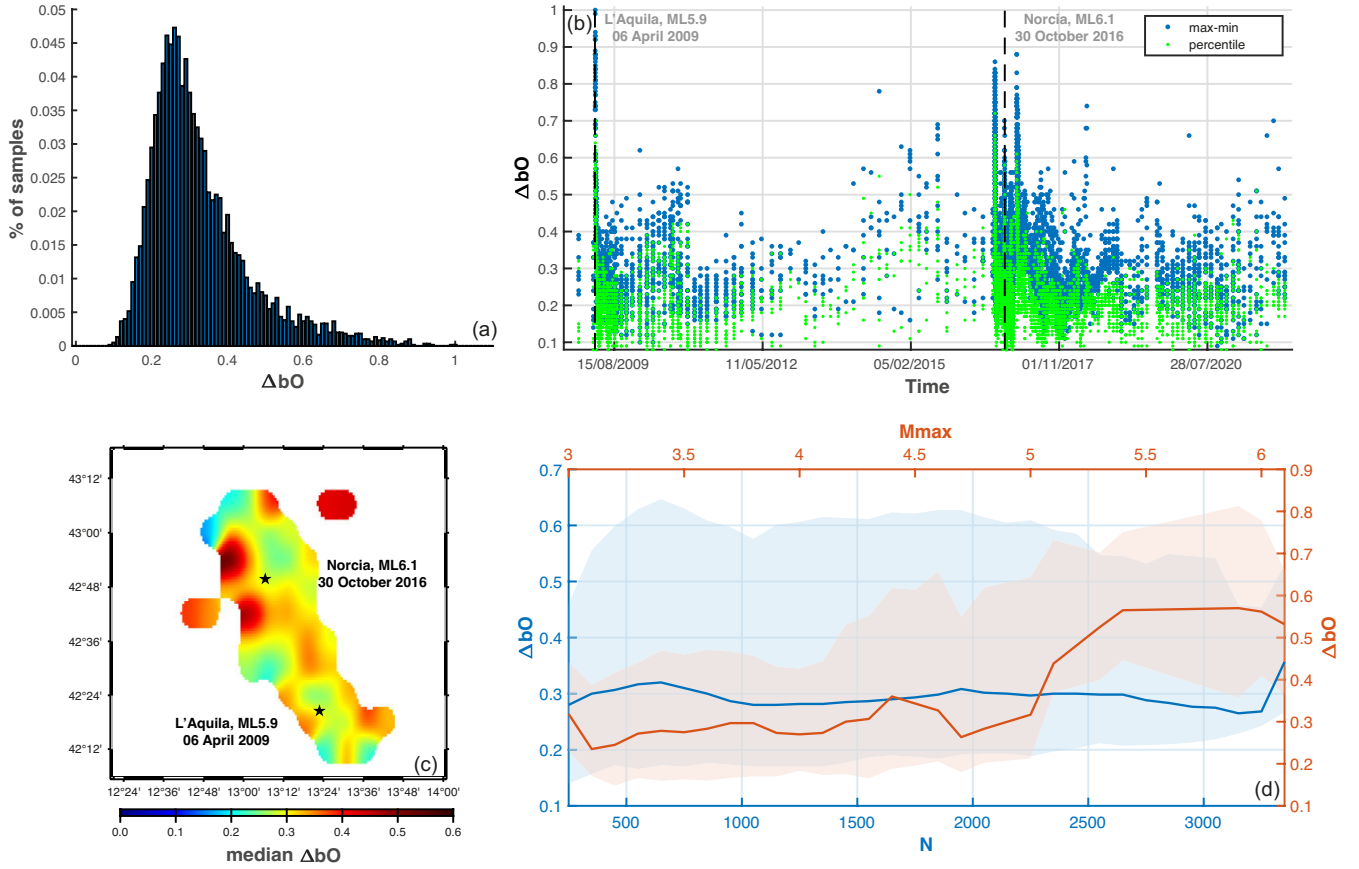


Figure 5. Overall differences in b values, ΔbO , among all algorithms, for the 6005 bins in common. (a) Histograms of ΔbO values. (b) Plot of ΔbO with time. The blue points mark the min-max range. The green points mark the ranges obtained by excluding the smaller two and larger two values, among the 19 provided by all methods. The grey dashed lines mark the occurrence times of the main shocks in the 2009 and 2016 sequences. (c) Map of median values of ΔbO . The black stars mark the locations of the main shocks in the 2009 and 2016 sequences. (d) Plot of ΔbO values versus the sample size N (blue) and the maximum magnitude $M_{\max}$ (orange). The shadowed areas mark the 90 per cent confidence intervals. The solid lines mark the median values.

of two models uses the b -positive approach (M1, M2, M3, M4) or is among M6, M7 and M19.

More generally, we investigate more in detail the relation between $\overline{Mc}$ and $\overline{b}$, by plotting ΔbO versus ΔMcO and ΔbC versus ΔMcC , for each pair of algorithms. The related results are outlined in Fig. 8 and shown in detail in [Movie S3](#), in the [Supplementary Information](#). Even if ΔMcO and ΔbO do not appear to be correlated (Fig. 8a), some peculiar relationships may be found by comparing couples of models. In some cases, one model mostly estimates lower Mc (or b) values with respect to the other model, but the differences in b (or Mc) values are both positive and negative (Figs 8b and c). Fig. 8(d) shows a case of linear relation between ΔMcC and ΔbC , for which a positive (negative) difference in Mc corresponds to a positive (negative) difference in b -values. A symmetric plot, centered on 0, with respect to both the ΔMcC and ΔbC axes, essentially refers to equivalent algorithms (Fig. 8e). The relation between ΔMcC and ΔbC for all couples of models may be observed in [Movie S3](#).

A comprehensive summary of the independent and joined relative differences among the values of Mc and b provided by the 19 methods is shown in Fig. 9. For each of the 6005 common bins, we compute the relative position of each of 19 Mc and b values with respect to their median value. Then we plot the median Mc and b relative positions on all bins, for each method. Joined relative Mc and b for the 15 methods M5–M19 show a monotone increasing relation: M13/M14 and M9/M10 provide the smallest and the largest

average values, respectively, for both Mc and b , with respect to the other methods. The b -positive methods M1–M4 provide the largest b -values on average.

7. DISCUSSION

Despite the simplicity of the empirical GR relationship, many practical problems have been raised, over time, in terms of parameter estimation, such as the effect of cumulative operation and of logarithmic transformation, the requirement that an acceptable data base must meet (homogeneous magnitudes, a long enough time span, a large range of magnitudes, a large enough sample size, no contamination by non-tectonic events, etc.) or the effect of measurement errors and decimal rounding of magnitudes (Marzocchi & Sandri 2003; Woessner & Wiemer 2005; Sandri & Marzocchi 2007; Gulia & Gasperini 2021; Geffers *et al.* 2022, 2023). The discussion of these and many other points has facilitated the development of new estimation algorithms, more than prevent the use of the already published algorithms. For this reason, there are many methods that may be chosen and implemented, without a clear prevalence of one over the others.

The main aim of this study was to quantify how much the choice of the estimation algorithm impacts the results and their interpretation. This matter is of utmost importance in the study of b -value variations and, more importantly, in the attempted use of the b -value

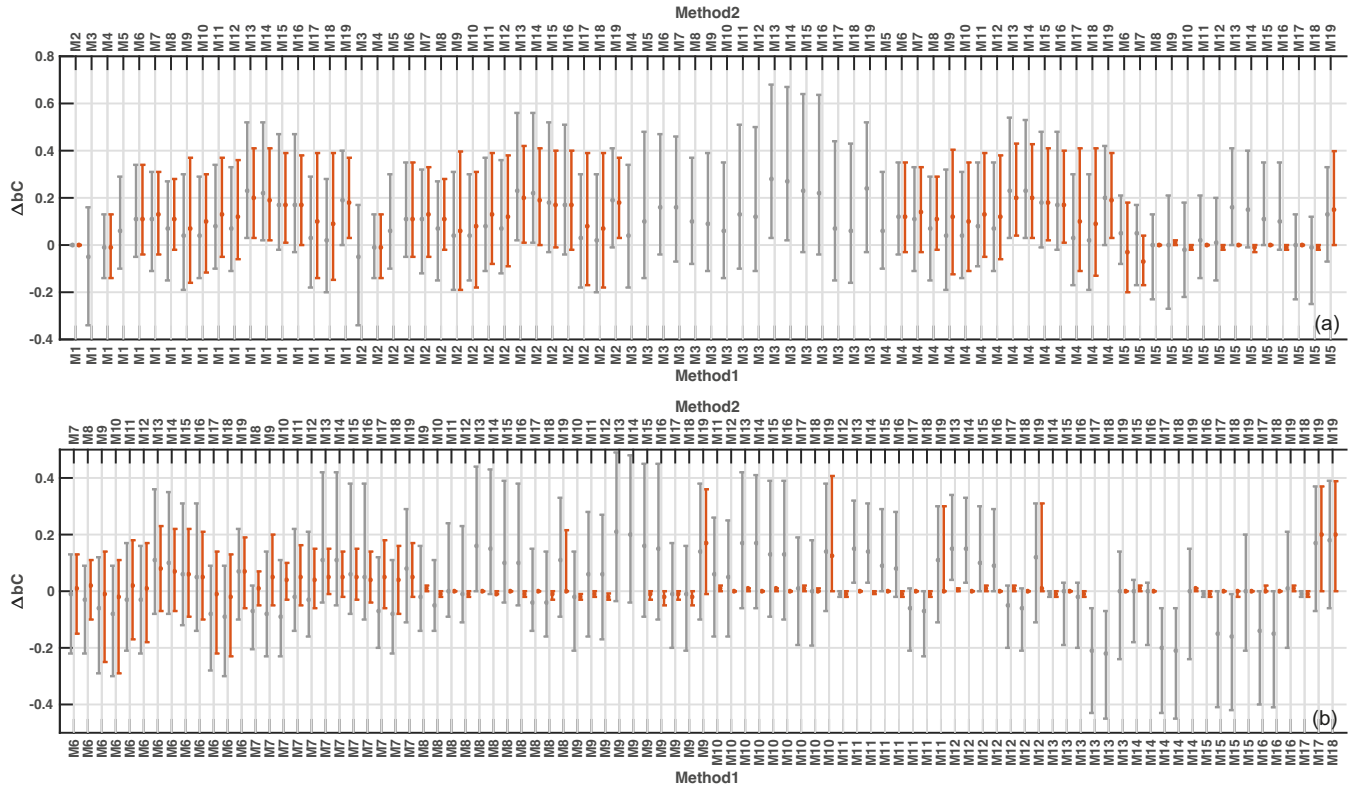


Figure 6. Relative differences in b values, ΔbC , for all couples of models. The grey lines mark the 90 per cent confidence intervals of ΔbC for all common bins. The orange lines mark the 90 per cent confidence intervals of ΔbC for bins for which the two algorithms provide equal M_c values. The bottom and top x-axes show the algorithm1 and algorithm2, respectively. All possible couples of algorithms are separated in panels (a) and (b).

as a precursor to large earthquakes. A method, also if applied in a self-consistent way through an analysis, does not guarantee proper results, since it might fail in some specific cases. Moreover, a large intermethod uncertainty about M_c and b estimations reveals individual differences in learning, knowledge and understanding that cannot be neglected.

The sensitivity study performed within this paper did not require the establishment of criteria for choosing the ‘best’ method or the development of strategies for ranking the algorithms: all methods considered in this study have the same reliability, since they were published and are used by the scientific community. Moreover, the choice of methods to include has not been motivated by the strength of their influence or by other intrinsic merits, but it is simply based on perspectives of participants.

The first result obtained in our analysis concerns the large sensitivity of M_c to the estimation method. We observed a median difference of 0.5 within G1 and G2 groups (ΔMcO ; Fig. 3) and up to 0.4 for pairs of algorithms, in both G1 and G2 (ΔMcC ; Figs 4a and b). The methods that more differ from the others are those based on rigorous goodness of fit statistical tests, M9 (LT, Lilliefors test; Hermann & Marzocchi 2021) and M10 (ND, Normal Distance test; Lombardi 2021): when they are compared to whichever method, ΔMcC is larger than 0.2 for more than 30 per cent of bins (Tables 2a, b). The overall results allowed us to establish that MaxC and GFT90 per cent provide the lower M_c values, whereas LT and ND provide the larger values (Figs 4c and d). Differences in M_c estimations are greater during the most intense phases of major seismic sequences, revealing difficulties to find the completeness when the data are strongly inhomogeneous due to rapid variation of detec-

tion. However, differences in M_c are far from negligible during the remaining times (median of 0.4–0.5; Fig. 3b), showing that the significant diversity among M_c values, provided by different methods, cannot be ascribed only to inhomogeneity caused by the occurrence of strong events, or, even less, to limitations deriving from small sample sizes (Fig. 3d). These differences are a systematic consequence of the requirements, limitations and peculiarities of each method (Movie S1 in the Supplementary Information): the choice of a method for estimating M_c and of underlying assumptions (such as the significance level for algorithms based on statistical tests) are crucial to determine the results.

The origin of such differences in M_c is to be sought in what we mean by completeness magnitude. Its definition as ‘the lowest magnitude at which 100 per cent of the events in a space–time volume are detected’ (Rydelek & Sacks 1989) opens the way to several considerations. First, statistics is not a method that provides certainties but rather addresses the degree of confidence in a belief. A statistical method may provide only a magnitude level, above which we are confident that the catalogue is complete with a certain probability. A higher confidence level indicates a greater degree of certainty (or a lower degree of uncertainty), achievable for a larger M_c . Therefore, the first difference between methods concerns the margin of errors that we are willing to accept when interpreting outcomes. The second point concerns the assumptions underlying our beliefs. Additionally, in this respect, the methods are quite varied, since they go from a check of log cumulative frequencies (MaxC, MaxC + 0.2, rMaxC) to procedures based on well-defined hypotheses on the magnitude distribution (LT, NDT), and through some ‘hybrid’ methods incorporating both (GFT). These different

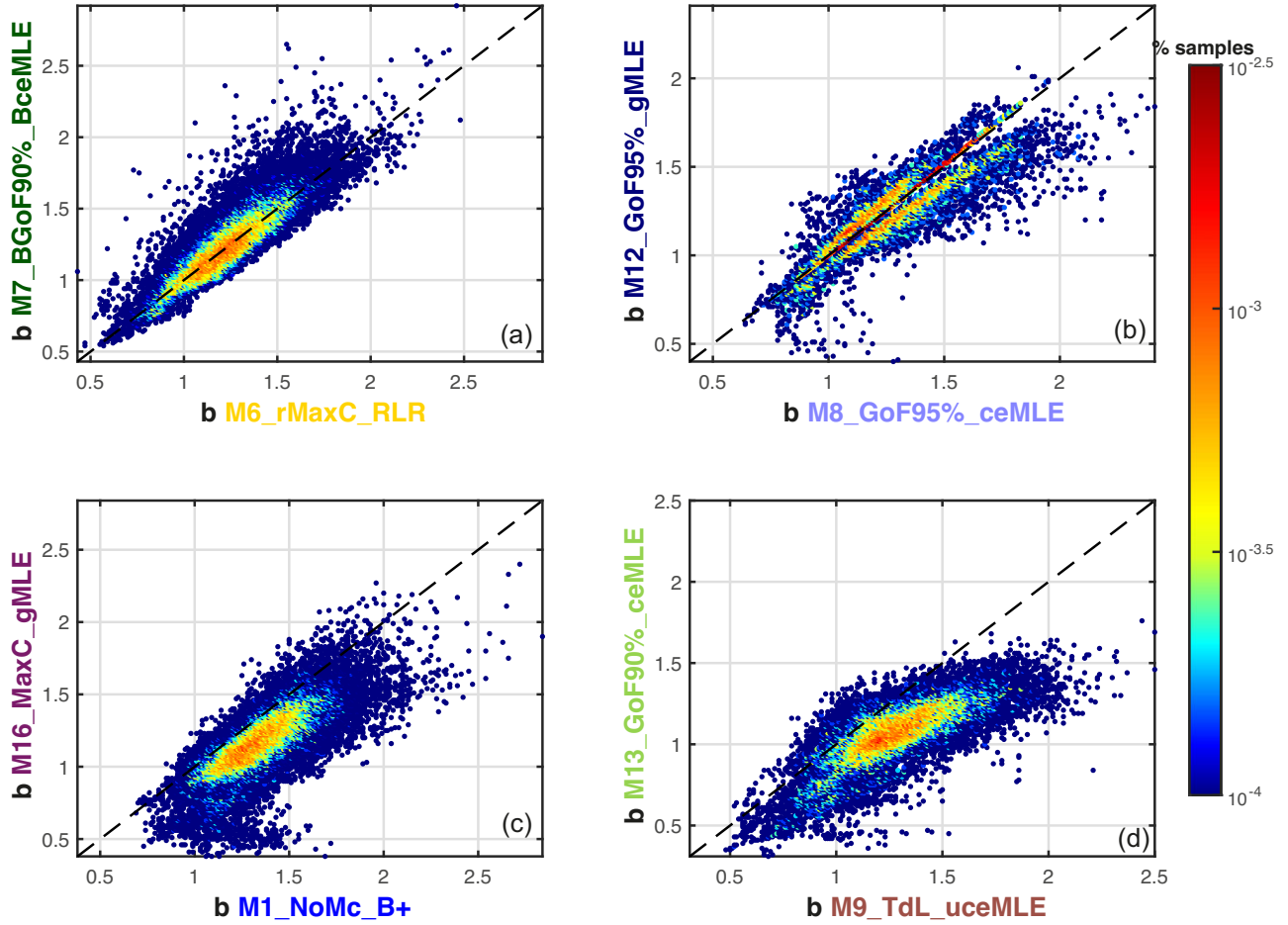


Figure 7. Four examples of the joint distributions of b -values estimated by two algorithms. The details of the comparison for all the couples of models are shown in [Movie S2](#), in the [Supplementary Information](#).

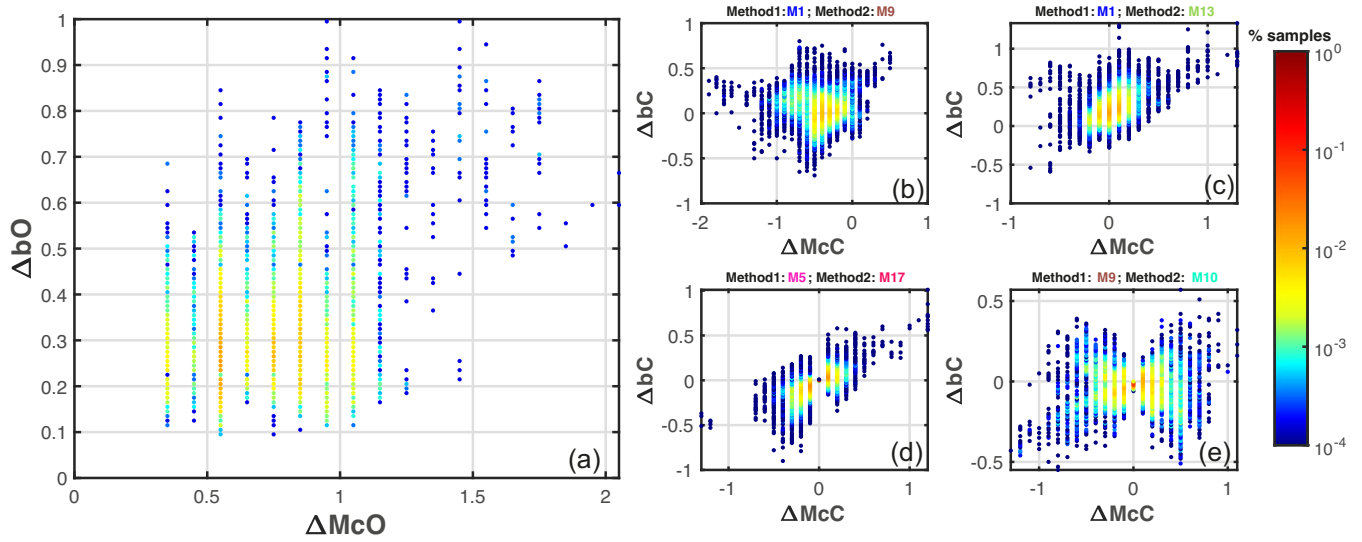


Figure 8. Relationships between the estimated values of M_c and b . (a) Plot of ΔbO versus ΔMcO for the 6005 bins in common to the 19 algorithms. Panels (b), (c), (d), (e) four examples of the joint distribution of ΔbC and ΔMcC . The details of the comparisons for all the couples of models are shown in [Movie S3](#), in the [Supplementary Information](#).

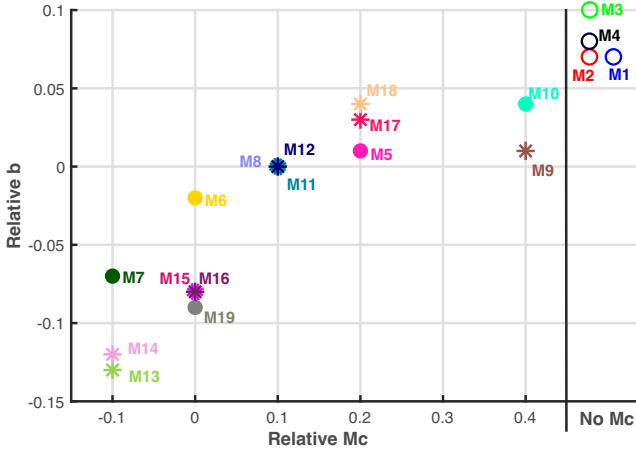


Figure 9. Relative average Mc and b for the 19 methods. The 0 indicates the median among the estimations provided by the 19 methods and, therefore, the negative and positive values indicates the smaller and the larger Mc and b , respectively, among the 19 methods. The details of computation are described in the text (Section 6). For methods M5–M19 the figure shows the average relative Mc versus the average relative b . The relative b values of the methods M1–M4, not providing Mc values, are plotted out the x-axis (No Mc area). The colour and symbols are the same as Fig. 2(a).

approaches inevitably produce different results, reflecting the individual perception of the problem. An example of this are the cases in Fig. S1a,b: the rigorous goodness-of-fit test, applied by algorithms M9 and M10, provides much larger Mc values with respect to the other methods, essentially checking the log-linearity of the cumulative distribution of magnitudes.

The second issue of our analysis concerns the b -value estimation, which is closely related to Mc evaluation. The overall difference, ΔbO , was greater than 0.3 for 50 per cent of the cases (Fig. 5a) and reached 1.0. Higher values (above 0.8) are obtained in conjunction with the occurrence of large events, but significant differences are observed also outside major sequences (Figs 5b and d). Substantial differences are also observed when comparing couples of algorithms (ΔbC , Fig. 6), but establishing an order among b -values provided by different methods is much more difficult with respect to Mc (Fig. 4).

The distributions of differences in Mc and b are not affected by the complexity of calculations during periods of intense seismic activity. In the Supplementary Information we report Fig. S1 showing the histograms like those of Figs 3(a) and 5(a) but obtained by excluding the subsamples of the first years of the 2009 L'Aquila and the 2016–2017 Central Italy sequences. No significant change is obtained for both distributions of ΔMcO and ΔbO . Moreover, the median value of ΔMcO is 0.4 and 0.5 for methods in G1 and G2, respectively. The median value of ΔbO is 0.28.

The quantification of differences among b -values inevitably also includes differences in Mc evaluation. If when we compare two algorithms, we isolate the contribution just due to b -value computation (by analysing bins with equal Mc evaluation), we find that differences are minimized for couples of algorithms with similar features, such as those based on geometric/exponential MLE (red lines Fig. 6; see also small panels in Movie S2). This finding provides some insight. The first is that the difference in $\bar{b}$ among algorithms based on MLE (M5 and M7–M18) is exclusively due to differences in Mc values (and, therefore, to differences in magnitude data), confirming that the evaluation of completeness is of utmost importance for b -value estimation (Woessner & Wiemer 2005; Amorèse *et al.* 2010).

The methodologies based on different criteria than MLE (M1–M4, M6) instead have intrinsic specificity, for which differences in outcomes cannot be entirely ascribed to Mc evaluations. The second remark concerns, more generally, the relation between Mc and b differences for each method, ranging from a linear behaviour, indicating lower b -values for lower Mc , to total independence (Fig. 8; Movie S3 in Supplementary Information). This finding shows the complexity of the relations between Mc and b and the need to interpret results on a case-by-case basis, also if, on average, the relative orders of Mc and b estimations given by each method are correlated (Fig. 9). This last result supports the existence of a strict, general, relationship between the b and the Mc values that cannot be neglected in a rigorous analysis of b -value changes. Both the underestimation and the overestimation of Mc can lead to inaccurate b -values due to the introduction of incomplete data and to the reduction of the sample size (Roberts 2015; Marzocchi *et al.* 2020; Geffers *et al.* 2022, 2023).

A comprehensive quantification of the uncertainty in b -value estimation and, specifically, of the variability of results among methods, is crucial for performing ‘trustworthy’ model-based inference, such as the detection of b -value variations. Here, epistemic uncertainty is classified as intramethod parameter uncertainty or intermethod parameter variability, which may be independently expressed as a coefficient of variation (standard deviation/mean). Therefore, we define the intramethod (COV^{intra}) and intermethod (COV^{inter}) uncertainty, for the 6005 bins in common among the 19 methods, as follows:

$$COV_{ij}^{\text{intra}} = \frac{\bar{\sigma}_{ij}}{\bar{b}_{ij}} \quad COV_j^{\text{inter}} = \frac{\tilde{\sigma}_j}{\tilde{b}_j},$$

where

- $\bar{b}_{ij}$ and $\bar{\sigma}_{ij}$ are the estimated value of b and the relative error, respectively, for the j -th bin and the i -th model
- $\tilde{b}_j$ and $\tilde{\sigma}_j$ are the mean and the standard deviation of the b -values estimated by the 19 algorithms for the j -th bin.

Fig. 10 shows the behaviour of $RCOV_{ij} = \frac{COV_j^{\text{inter}}}{COV_{ij}^{\text{intra}}}$ versus time for all algorithms. The intermethod variation is larger than the intramethod uncertainty, on the average, for most of the 19 algorithms, and for all methods during the most intense activity of major seismic sequences, reaching a maximum proportion ranging from 7 to approximately 30 times. This result is due to the impact of the estimation method of the Mc and b parameters on results, but also to the uncertainty propagation through the Mc and b parameters, which is not considered in the intramethod uncertainty, referring only to the b -value. This shows how the choice of the estimation method is decisive, often much more than uncertainty in the estimated b -value, and how much attention must be given to interpreting results, especially in the case of strong events. Each estimation method underlies a conceptual model; therefore, variations due to the use of different conceptual models should be considered as part of the uncertainty of the results and properly addressed (Neuman 2003).

8. CONCLUSIONS

The results of our seismological experiment serve as a basis for understanding the diversity of opinions and results about b -value changes. We find that the sensitivity of the estimated values of Mc and b to the 19 alternative model specifications is significant: the intermethod differences in Mc and b are on average equal to 0.4 and 0.3, respectively, but reach much larger values, especially

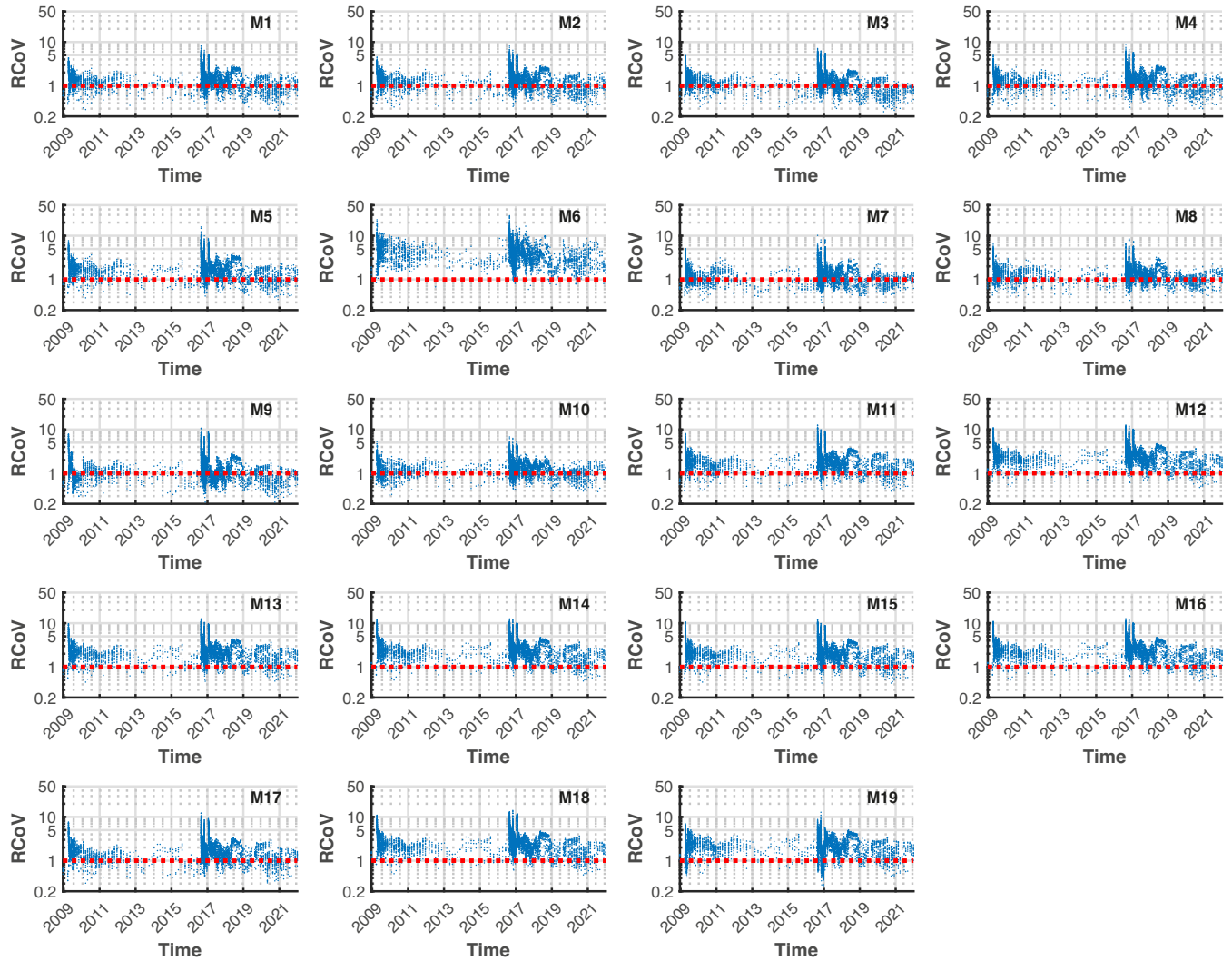


Figure 10. Relationships between the overall intra- and intermethod b -value covariances versus time. The dotted line separates the values of $RCoV$ for which the intermethod covariance is larger (above) or lower (below) than the intramethod covariance.

during the most intense activity periods. These results probably underestimate actual levels of intermethod uncertainty, since they do not include all possible methods available in literature. Even if more specific studies are required to additionally investigate the origin of the observed intermethod variations, this study aims to present a new contribution to the debate about b -value variations and, most importantly, on their potential as earthquake precursors. Ignoring the intermethod variability of estimated b -values may lead to suboptimal information, when we use the patterns of changes in b -values as precursors of subsequent large events or as stress metres. This has obvious consequences if this information is framed from a decision-making perspective. Well informed decision-making cannot be based on single model predictions and must acknowledge uncertainties in the concept of the b -values estimation method and consequent outcomes.

By this study, we are not overall rejecting the idea of the existence of b -value changes or the chance to use them as precursor of large events. The substantial differences across methods, found in our study, suggest that the changes in b -values could be due to methodological biases more than to physical causes and that they are probably overrated (Amorè *et al.* 2010). No method can probably

resolve all problems, but an objective judgement of methods and results may substantially reduce, if not eliminate, misunderstanding. Moreover, the large intermethod uncertainty shown in this study reveals individual differences in knowledge and in the meaning given to the Gutenberg–Richter Law that cannot longer be ignored.

We hope that this seismological experiment, together with similar comparative studies (e.g. Mai *et al.* 2016; Baltay *et al.* 2024), will pave the way for future similar enterprises, to illuminate the uncertainties on b -value and on different seismological parameters.

To conclude, the results of our experiment suggest that the efficient use of the b -value as an indicator of the stress level or of imminent large events requires an intermethod uncertainty reduction. In addition to individual studies, promoting collective efforts, as in this paper, will contribute significantly to the achievement of this goal.

ACKNOWLEDGMENTS

The authors are grateful to reviewers (J. Zhuang, D. Amorese and A. Christophersen) and the associate editor (M. Segou) for their constructive criticisms and valuable comments, which were of great

help in revising the manuscript and strongly improved the quality of the paper.

Author contributions: AML and LF outlined and led the research, performed the statistical analyses, created the figures and wrote the manuscript. FPL, MT, LG and AT contributed to writing the manuscript. All the authors contributed to the modelling of the data and code development, to the review of the original draft and to discussion of the results.

SUPPORTING INFORMATION

Supplementary data are available at [GJIRAS](https://doi.org/10.1016/j.gjiras.2024.100000) online.

File S1. List of the 525 grid nodes. The format is latitude, longitude.

File S2. List of the times of the first (Fev) and the last event (Lev) for the 292 temporal windows. The format is: Year Fev, Month Fev, Day Fev, Hour Fev, Minute Fev, Second Fev, Year Lev, Month Lev, Day Lev, Hour Lev, Minute Lev, Second Lev.

Movie S1. Density plots for the two sets of M_c values, provided by all couples of methods in G1 and G2. The colour of filled circles is scaled by proportion of samples. The dotted lines mark the bisector, indicating cases of equal M_c values.

Movie S2. Density plots for the two sets of b -values (two decimal digits are considered), provided by all couples of methods. The colour of filled circles is scaled by proportion of samples. The dotted lines mark the bisector, indicating cases of equal b -values. The top-left small panel show cases with equal M_c , only.

Movie S3. Density plots of ΔMcC versus ΔbC for all couples of methods. The colour of filled circles is scaled by proportion of samples. Two decimal digits are considered for b -values.

Figure S1. Histograms of ΔMcO (left panel) values, separated for algorithms in G1 and G2, and of ΔbO (right panel) values, obtained by excluding the subsamples of the first year of the 2009 and 2016–2017 sequences.

Figure S2. Magnitude distribution for four subsamples and relative M_c evaluations. The blue histogram shows the distribution of magnitude values, whereas the orange filled circles show the log cumulative number of events, versus magnitude (top x-axis). Dotted lines mark the values of M_c provided by different methods. The black points, in the bottom panel, show the magnitudes of the respective subsample versus time. (a) Magnitude sample with maximum value of ΔMcO for methods in G1. (b) The same of (a) but for methods in G2. Both examples in (a) and (b) belong to 2016 Central Italy sequence and includes three strong events: the main, Norcia (2016 October 30, $M_L6.1$, $M_w6.5$), the Visso (2016 October 26, $M_L5.8$, $M_w5.9$) and the Castelsantangelo sul Nera (2016 October 26, $M_L5.4$, $M_w5.4$) events. (c) Magnitude sample, including the main shock of 2009 L'Aquila sequence (2009 April 06, $M_L5.9$, $M_w6.1$), for which the M_c evaluations provided by methods in G1 are rather close. (d) Magnitude sample, including low-magnitude seismicity, for which the M_c evaluations provided by methods in G2 are quite different ($\Delta McO = 0.7$).

Figure S3. Magnitude distribution for two subsamples and relative M_c/b evaluations. The blue histogram shows the distribution of magnitude values, whereas the orange filled circles show the log cumulative number of events, versus magnitude (top x-axis). Solid and dotted lines mark the frequency magnitude fit provided by different methods. The black points, in the bottom panel, show the magnitudes of the respective subsample versus time. (a) Magnitude sample with maximum value of ΔbO , close to 1.0, including the main shock of 2009 L'Aquila sequence. The M_c differences are

significant (from 1.3 to 2.8), too. (b) Magnitude sample, including low-magnitude seismicity, for which b -value evaluations are close ($\Delta bO = 0.09$) and M_c , going from 1.0 to 1.7, are quite different.

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.

DATA AVAILABILITY

Earthquake data were collected in the Italian Seismological Instrumental and Parametric Data base (ISIDE; <http://terremoti.ingv.it/en/iside>), between 1985 January 1 and today. Starting in 2015, quarterly BSI releases were published on the web page <http://terremoti.ingv.it/en/bsi> in QuakeML format and with a DOI assigned to each file.

The Supporting Information includes [Figs S1–S2](#) and [Movies S1–S3](#).

REFERENCES

- Aki, K. 1965. Maximum likelihood estimate of b in the formula $\log(N) = a - bM$ and its confidence limits, *Bull. Earthq. Res. Inst. University of Tokyo*, **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.
- Amorè, D. 2007. Applying a change-point detection method on frequency–magnitude distributions, *Bull. seism. Soc. Am.*, **97**(5), 1742–1749
- Amorè, D., Grasso, J.R. & Rydelek, P.A. 2010. On varying b -values with depth: results from computer-intensive tests for Southern California, *Geophys. J. Int.*, **180**(1), 347–360
- Baltay, A., Abercrombie, R., Chu, S. & Taira, T. 2024. The SCEC/USGS Community Stress drop validation study using the 2019 Ridgecrest earthquake sequence, *Seismica*, **3**(1),.
- Bender, B. 1983. Maximum likelihood estimation of b values for magnitude grouped data, *Bull. seism. Soc. Am.*, **73**(3), 831–851
- Cao, A. & Gao, S.S. 2002. Temporal variation of seismic b -values beneath northeastern Japan island arc, *Geophys. Res. Lett.*, **29**(9), 48.1–48.3.
- Corral, A. & González, A. 2019. Power law distributions in geoscience revisited, *Earth Space Sci.*, **6**, 673–697
- 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, *Geophys. Res. Lett.*, **39**, L15304.
- Field, E.H. et al. 2023. The USGS 2023 continuous U.S. Time-independent earthquake rupture forecast, *Bull. seism. Soc. Am.*, **114**(1), 523–571.
- Gasparini, P., Lolli, B. & Vannucci, G. 2013. Empirical calibration of local magnitude data sets versus moment magnitude in Italy, *Bull. seism. Soc. Am.*, **103**, 2227–2246.
- 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., 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.
- Gerstenberger, M., Wiemer, S. & Giardini, D. 2001. A systematic test of the hypothesis that the b value varies with depth in California, *Geophys. Res. Lett.*, **28**, 57–60
- Gibbons, J.D. & Chakraborti, S. 2020. *Nonparametric Statistical Inference*, 6th ed., M. Dekker.
- Godano, C., Petrillo, G. & Lippiello, E. 2023. Evaluating the incompleteness magnitude using an unbiased estimate of the b value, *Geophys. J. Int.*, **236**(2), 994–1001.

- Godano, C., Tramelli, A., Petrillo, G. & Convertito, V. 2024. Testing the predictive power of b value for Italian seismicity, *Seismica*, **3**(1).
- Gulia, L. & Gasperini, P. 2021. Contamination of frequency–Magnitude slope (bb-Value) by quarry blasts: an example for Italy, *Seismol. Res. Lett.*, **92**(6), 3538–3551.
- Gulia, L. & Wiemer, S. 2019. Real-time discrimination of earthquake foreshocks and aftershocks, *Nature*, **574**(7777), 193–199.
- Gulia, L., Tormann, T., Wiemer, S., Herrmann, M. & Seif, S. 2016. Short-term probabilistic earthquake risk assessment considering time-dependent b values, *Geophys. Res. Lett.*, **43**, 1100–1108.
- Gulia, L., Wiemer, S., Biondini, E., Enescu, B. & Vannucci, G. 2024. Improving the Foreshock Traffic Light Systems for real-time discrimination between foreshocks and aftershocks, *Seismol. Res. Lett.*, **95**, 3579–3592.
- Gutenberg, B. & Richter, F. 1944. Frequency of earthquakes in California, *Bull. seism. Soc. Am.*, **34**, 185–188.
- Han, Q., Wang, L., Xu, J., Carpinteri, A. & Lacidogna, G. 2015. A robust method to estimate the b -value of the magnitude–frequency distribution of earthquake, *Chaos Solitons Fractals*, **81**, 103–110.
- Hermann, M. & Marzocchi, W. 2021. Inconsistencies and lurking pitfalls in the magnitude–frequency distribution of high-resolution earthquake catalogs, *Seismol. Res. Lett.*, **92**, 909–922.
- Hermann, M., Piegari, E. & Marzocchi, W. 2022. Revealing the spatiotemporal complexity of the magnitude distribution and b -value during an earthquake sequence, *Nat. Commun.*, **13**, 5087.
- Hoaglin, D.C. & Welsch, R.E. 1978. The hat matrix in regression and ANOVA, *Am. Stat.*, **32**(1), 17–22.
- Ishimoto, M. & Iida, K. 1939. Observations of earthquakes registered with the microseismograph constructed recently, *Bull. Earthq. Res. Inst., University of Tokyo*, **17**, 443–478.
- Kagan, Y. 1999. The universality of the frequency–moment relationship, *Pure appl. Geophys.*, **155**, 537–574.
- 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.: Solid Earth*, **120**, 5191–5214.
- Lacidogna, G., Borla, O. & De Marchi, V. 2023. Statistical seismic analysis by b -value and occurrence time of the latest earthquakes in Italy, *Remote Sens.*, **15**, 5236.
- Leptokaropoulos, K.M., Adamaki, A.K., Roberts, R.G., Gkaraouni, C.G. & Paradisopolou, P.M. 2018. Impact of magnitude uncertainties on seismic catalogue properties, *Geophys. J. Int.*, **213**(2), 940–951.
- Leptokaropoulos, K.M., Karakostas, V.G., Papadimitriou, E.E., Adamaki, A.K., Tan, O. & Inan, S. 2013. A homogeneous earthquake catalogue compilation for western Turkey and magnitude of completeness determination, *Bull. seism. Soc. Am.*, **103**(5), 2739–2751.
- Lilliefors, H.W. 1969. On the Kolmogorov–Smirnov test for the exponential distribution with mean unknown, *J. Am. Stat. Assoc.*, **64**(325), 387–389.
- Lippiello, E. & Petrillo, G. 2024. b -more-incomplete and b -more positive: insights on a robust estimator of magnitude distribution, *J. geophys. Res.: Solid Earth*, **129**, e2023JB027849.
- Lombardi, A.M. 2021. A normalized distance test for co-determining the completeness magnitude and b -value of earthquake catalogs, *J. geophys. Res.: Solid Earth*, **126**, e2020JB021242.
- 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**, 1545–1555.
- Lombardi, A.M. 2024. A statistical framework for detection of b -value anomalies in Italy, *Geophys. J. Int.*, **237**, 729–740.
- Mai, P.M. et al. 2016. The earthquake-source inversion validation (SIV) Project, *Seismol. Res. Lett.*, **87**(3), 690–708.
- 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.
- Mignan, A. & Woessner, J., 2012, Estimating the magnitude of completeness for earthquake catalogs, *Commun. Online Resource Stat. Seism. Anal.*, <https://doi.org/10.5078/corssa-00180805>.
- Mogi, K. 1962. Study of the elastic shocks caused by the fracture of heterogeneous materials and its relation to earthquake phenomena, *Bull. Earthq. Res. Inst.*, **40**, 125–173.
- Neuman, S. 2003. Maximum likelihood bayesian averaging of uncertain model predictions, *Stoch. Environ. Res. Risk Assess.*, **17**(5), 291–305.
- Ogata, Y. & Katsura, K. 1993. Analysis of temporal and spatial heterogeneity of magnitude frequency distribution inferred from earthquake catalogs, *Geophys. J. Int.*, **113**, 727–738.
- Papadopoulos, G.A., Charalampakis, M., Fokaefs, A. & Minadakis, G. 2010. Strong foreshock signal preceding the L'Aquila (Italy) earthquake (M_w 6.3) of 6 April 2009, *Nat. Hazards Earth Syst. Sci.*, **10**, 19–24.
- Pavlenko, V.A. & Zavyalov, A.D. 2022. Comparative analysis of the methods for estimating the magnitude of completeness of earthquake catalogs, *Phys. Solid Earth*, **58**, 89–105.
- Petrucelli, A., Gasperini, P., Tormann, T., Schorlemmer, D., Rinaldi, A.P., Vannucci, G. & Wiemer, S. 2019. Simultaneous dependence of the earthquake-size distribution on faulting style and depth, *Geophys. Res. Lett.*, **46**(20), 11 044–11 053.
- Rhoades, D.A. 1996. Estimation of the Gutenberg–Richter relation allowing for individual earthquake magnitude uncertainties, *Tectonophysics*, **258**, 71–83.
- Roberts, N.S. 2015. Are volcanic seismic b -values high, and if so when?, *J. Volc. Geotherm. Res.*, **308**, 127–141.
- Rydelek, P.A. & Sacks, I.S. 1989. Testing the completeness of earthquake catalogs and the hypothesis of self-similarity, *Nature*, **337**, 251–253.
- Sandri, L. & Marzocchi, W. 2007. A technical note on the bias in the estimation of the b -value and its uncertainty through the Least Squares technique, *Ann. Geophys.*, **50**, 329–339.
- Scholz, C.H. 1968. The frequency–Magnitude relation of microfracturing in rock and its relation to earthquakes, *Bull. seism. Soc. Am.*, **58**, 399–415.
- Schorlemmer, D., Wiemer, S. & Wyss, M. 2005. Variation 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. seism. Soc. Am.*, **72**, 1677–1687.
- Smyth, C. & Mori, J. 2009. Temporal variations in the Gutenberg–Richter distribution prior to the Kobe earthquake, *Annu Disaster*, **2009**(52B), 255–261.
- Sugan, M., Kato, A., Miyake, H., Nakagawa, S. & Vuan, A. 2014. The preparatory phase of the 2009 M_w 6.3 L'Aquila earthquake by improving the detection capability of low-magnitude foreshocks, *Geophys. Res. Lett.*, **41**, 6137–6144.
- Tinti, S. & Gasperini, P. 2024. The estimation of b -value of the frequency–magnitude distribution and of its 1σ intervals from binned magnitude data, *Geophys. J. Int.*, **238**, 433–458.
- Tinti, S. & Mulargia, F. 1987. Confidence intervals of b -values for grouped magnitudes, *Bull. seism. Soc. Am.*, **77**, 2125–2134.
- Tormann, T., Wiemer, S. & Mignan, A. 2014. Systematic survey of high-resolution b value imaging along Californian faults: inference on asperities, *J. geophys. Res.*, **119**, 2029–2054.
- Utsu, T. 1965. A method for determining the value of b in formula $\log N = a - bM$ showing the magnitude–frequency relation for earthquakes, *Geophys. Bull. Hokkaido University*, **13**, 99–103.
- van der Elst, N.J. 2021. B-positive: a robust estimator of aftershock magnitude distribution in transiently incomplete catalogs, *J. geophys. Res.: Solid Earth*, **126**, e2020JB021027.
- Wiemer, S. & Baer, M. 2000. Mapping and removing quarry blast events from seismicity catalogs, *Bull. seism. Soc. Am.*, **90**(2), 525–530.
- Wiemer, S. & Wyss, M. 2000. Minimum magnitude of completeness in earthquake catalogs: examples from Alaska, the western United States and Japan, *Bull. seism. Soc. Am.*, **90**(4), 859–869.
- Wiemer, S. & Wyss, M. 2002. Mapping spatial variability of the frequency magnitude distribution of earthquakes, *Adv. Geophys.*, **45**, 259–302.
- Woessner, J. & Wiemer, S. 2005. Assessing the quality of earthquake catalogs: estimating the magnitude of completeness and its uncertainty, *Bull. seism. Soc. Am.*, **95**, 684–698.
- Zhou, Y.J., Zhou, S.Y. & Zhuang, J.C. 2018. A test on methods for MC estimation based on earthquake catalog, *Earth planet. Phys.*, **2**, 150–162. <http://doi.org/10.26464/epp2018015>

APPENDIX A: DESCRIPTION OF THE 19 ESTIMATION METHODS

In the following, we describe the estimation methods that were used in the experiment. Figs A1 and A2, show the behaviour of $\bar{b}$ as a function of $N[M_c]$ and ΔM , for all methods.

A1. Algorithms 1–4 (G1): The b -positive method

(M1_NoMc_B+, M2_NoMc_B+, M3_NoMc_B+, M4_NoMc_B+)

The novel b -positive b -value estimator, developed by van der Elst (2021), does not require an accurate estimate of M_c and is less impacted by transient changes in catalog completeness. Therefore, it can be used to compute more robust b -values during periods of strongly varying incompleteness levels (e.g. Field *et al.* 2023; Gulia *et al.* 2024; Lippiello & Petrillo 2024; Tinti & Gasperini 2024).

This approach is based on the positive-only subset of the differences in magnitude between successive earthquakes. According to van der Elst (2021), the distribution of non-null magnitude differences is identical to that of earthquake magnitudes, with the same b -value, but positive magnitude differences are less affected by the small earthquake detection problem.

Although Tinti & Gasperini (2024) showed that there is no reason to prefer estimations based on positive-only magnitude differences, here we use the original formula (A3 in Appendix A of van der Elst 2021), based only on trimmed positive differences only:

$$\bar{b}_{B+} = \frac{1}{\frac{\delta M}{2} \ln(10)} \coth^{-1} \left(\Delta M - \Delta M'_c + \frac{\delta M}{2} \right) \times \left[\frac{1}{\frac{\delta M}{2}} \left(\overline{\Delta M} - \Delta M'_c + \frac{\delta M}{2} \right) \right], \quad (A1)$$

where $\coth^{-1}$ is the inverse of the hyperbolic cotangent function, $\overline{\Delta M}$ is the average of positive magnitude differences, $\Delta M'_c$ is the minimum allowed magnitude difference, and δM is the magnitude binning size (here equal to 0.1).

The b -positive estimator is not completely insensitive to incompleteness, and this latter may be treated in two different ways (Tinti & Gasperini 2024):

- the cutoff magnitude (M_{cut}) is estimated by the Maximum Curvature Method (Wiemer & Wyss 2000), which calculates the maximum value of the first derivative of the log-linear frequency–magnitude cumulative distribution, that is the magnitude bin with the highest occurrence frequency of events; then, the magnitudes below M_{cut} and the magnitude differences ΔM below δM (i.e. $\Delta M'_c = \delta M = 0.1$) are discarded;
- all the magnitudes are used, and the magnitude differences ΔM below 0.4–0.5 are discarded.

In view of the above, four different b -value estimates are provided, based on the b -positive approach:

- M1_NoMc_B+: the magnitudes below M_{cut} (estimated for each spatiotemporal bin) are discarded, and the b -positive method is applied to bins having at least 100 events (above M_{cut}), at least, by fixing $\Delta M'_c = 0.1$ and by considering only positive magnitude differences.
- M2_NoMc_B+: the bins having at least 100 events are selected, and the b -positive method is applied on magnitudes above M_{cut} (estimated for each spatiotemporal bin), by fixing $\Delta M'_c = 0.1$ and by considering only positive magnitude differences.

- M3_NoMc_B+: the bins having at least 50 events are selected, and the b -positive method is applied to all the data, by fixing $\Delta M'_c = 0.5$ and by considering the absolute magnitude differences.

- M4_NoMc_B+: the magnitudes below M_{cut} (estimated for each spatiotemporal bin) are discarded, and the b -positive method is applied to bins having at least 100 events (above M_{cut}) by fixing $\Delta M'_c = 0.1$ and by considering disjointed (not consecutive) absolute differences.

A2. Method 5 (G1): the corrected exponential maximum-likelihood method

(M5_MaxC + 0.2.ceMLE)

This method estimates M_c by adding a correction factor of 0.2 to the value provided by the Maximum Curvature Method (Wiemer & Wyss 2000), for which a systematic underestimation error has been recognized (Woessner & Wiemer 2005).

The b -value is estimated by the maximum-likelihood formula proposed by Aki (1965), deduced by assuming a continuous, exponential distribution for magnitudes, corrected by the factor proposed by Utsu (1965), to avoid bias from binned (discrete) magnitudes:

$$\bar{b}_{\text{ceMLE}} = \frac{1}{\ln(10) (\bar{M} - M_c + \frac{\delta M}{2})}, \quad (A2)$$

where $\bar{M}$ is the average magnitude (above M_c) and δM is the magnitude bin size (here, equal to 0.1).

A3. Algorithm 6 (G1): the robust linear regression

(M6_rMaxC_RLR)

This method estimates M_c by using the ‘reinforced’ Maximum Curvature Method (Wiemer & Wyss 2000), which consists of choosing the highest magnitude value, if multiple magnitude values have the same, larger, occurrence frequency and, therefore, all correspond to the point of maximum curvature.

The b -value is estimated by the robust linear regression (RLR; Han *et al.* 2015), providing the best-fit line for the logarithm of the cumulative frequency–magnitude distribution, for at least 75 events. Like in the standard linear regression, the objective of RLR is to minimize the sum of squared residuals, r_i ; however, by assigning lower, bisquare, weights, w_i , to data points with higher residuals, RLR mitigates the influence of outliers (Han *et al.* 2015). The RLR method consists of an iterative reweighted least-squares procedure, aimed at minimizing the quantity

$$S_{\text{RFM}} = \sum_{i=1}^N w_i r_i^2 = \sum_{i=1}^N w_i (y_i - \hat{y}_i)^2, \quad (A3)$$

where N is the number of observations y_i (cumulative number of events for each magnitude bin) and $\hat{y}_i$ denotes the values predicted by the model. Initially, the residuals r_i are computed relative to an initial ordinary least-squares estimation. Adjusted residuals and robust weights are then computed by means of leverages (Hoaglin & Welsch 1978), measuring the outliers with respect to the independent variable (here, the magnitude bins). The aim is to downweigh high-leverage data points, which have a large effect on the ordinary least-squares fit. Any details on the residual and weight computations are explained in Han *et al.* (2015; subsection 2.3).

The iterative procedure continues until convergence is achieved, that is further adjustments of model parameters do not significantly improve the quality of the fit. Upon achieving convergence, the b -value is defined as the slope of the best-fitting line.

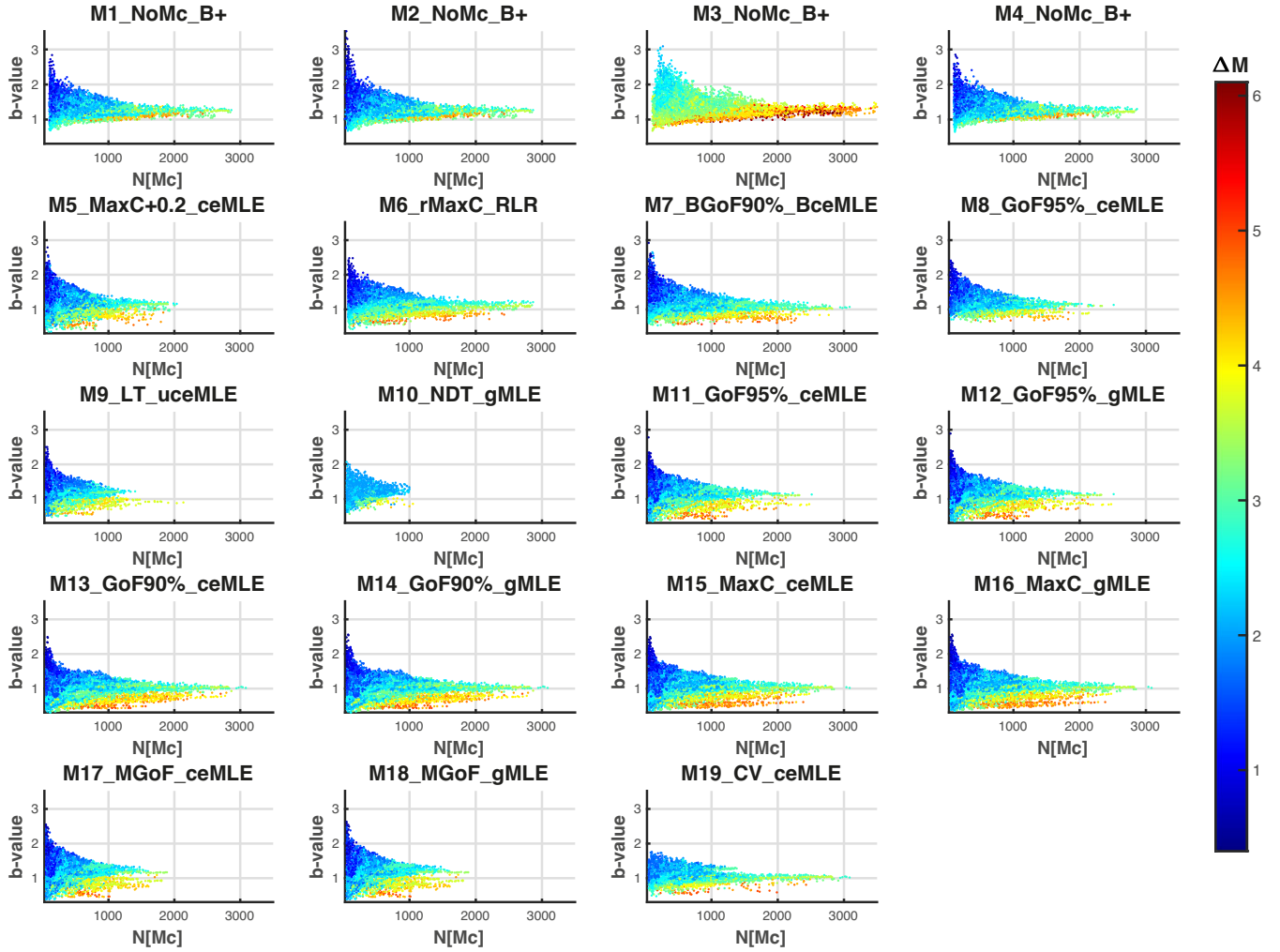


Figure A1. Behaviour of $\bar{b}$ as a function of $N[M_c]$ for all 19 algorithms. The colour of points is scaled with the magnitude range ΔM .

A4. Algorithms 7–8 (G1): the goodness-of-fit method

(M7_BGoF90 per cent_BceMLE, M8_GoF95 per cent_ceMLE)

Algorithms 7 and 8 both use the goodness-of-fit test (Wiemer & Wyss 2000) to estimate M_c . This method calculates M_c by comparing the observed cumulative frequency–magnitude distribution with the theoretical distribution. The latter is calculated by using the corrected maximum likelihood estimation of the b -value (eq. A2; Aki 1965; Utsu 1965).

The method consists of computing for each ascending cutoff magnitude, M_{cut} ,

- a and b parameters of the GR relation,
- the measure R

$$R(M_{\text{cut}}, a, b) = 100 - \frac{\sum_{M_{\text{cut}}}^{M_{\text{max}}} |O_i - E_i|}{\sum_{M_{\text{cut}}}^{M_{\text{max}}} O_i} 100, \quad (\text{A4})$$

where O_i and E_i are the observed and the predicted cumulative numbers of events in each magnitude bin, respectively. M_c is the smallest M_{cut} where R achieves a given confidence level, fixed to 90 per cent and 95 per cent, for Model 7 and Model 8, respectively.

To include the uncertainty in the M_c evaluation, Model 7 applies the whole procedure to 1000 bootstrapped resamples of the original

data set and estimates M_c and b as the means of the respective bootstrapped values (Woessner & Wiemer 2005).

A5. Algorithm 9 (G2): the Lilliefors test (LT) method

(M9_LT_uceMLE)

The Lilliefors test (LT) was first applied to earthquake catalogs by Herrmann & Marzocchi (2021). For each ascending cutoff magnitude, $M_{\text{cut}} \geq 1.5$, the exponentiality of magnitudes above M_{cut} is tested using the Lilliefors test (Lilliefors 1969). Uniform random noise $[-0.05, 0.05]$ was added to (binned) magnitudes to obtain continuous variables, as requested by the Lilliefors test (Herrmann & Marzocchi 2021). We performed 100 different runs of random noise, and M_c was determined to be the smallest magnitude for which the median of 100 p -values of the Lilliefors tests was greater than or equal to 0.1.

The b -value is computed, for subcatalogues containing at least 50 events above M_c , by the unbiased corrected exponential MLE (Marzocchi *et al.* 2020)

$$\bar{b}_{\text{uceMLE}} = \frac{N-1}{N} \cdot \bar{b}_{\text{ceMLE}}, \quad (\text{A5})$$

where N is the total number of events and $\bar{b}_{\text{ceMLE}}$ is given in equation (A2).

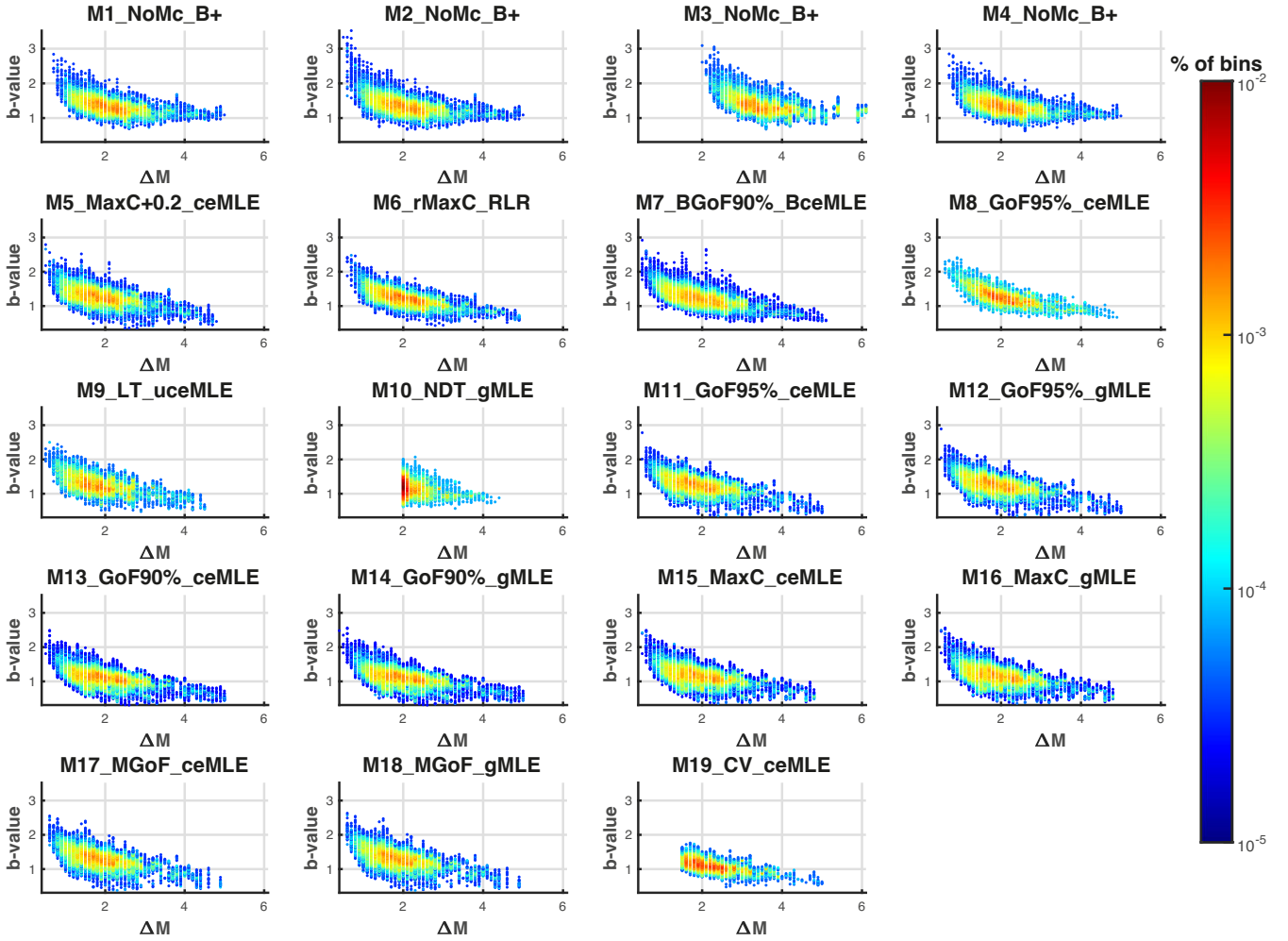


Figure A2. Density plot of $\bar{b}$ as a function of the magnitude range ΔM for all 19 algorithms.

A6. Algorithm 10 (G1): the normalized distance (ND) test method

(M10_NDT_gMLE)

The normalized distance (ND) test (Lombardi 2021, 2024) is structured as follows. Given a sample S of N magnitudes M (rounded to the first decimal place), NB bootstrap resampling catalogues are generated. For each of them, three quantities are computed, as a function of the ascending cutoff magnitude, M_{cut} :

(i) the geometric maximum likelihood b -value estimator $\bar{b}_{\text{ND}}(M_{\text{cut}})$

$$\bar{b}_{\text{ND}}(M_{\text{cut}}) = \frac{1}{\ln(10) \cdot \delta M} \ln(1 - \bar{p}), \quad (\text{A6})$$

where $\bar{p} = \frac{\delta M}{\bar{M} + \delta M}$, $\delta M = 0.1$ and $\bar{M}$ is the mean magnitude for $M \geq M_{\text{cut}}$ (Tinti & Mulargia 1987);

(i) the W-statistic $W(M_{\text{cut}}) = N_{\text{cut}} \cdot D[\bar{b}(M_{\text{cut}})]$, where $D[\bar{b}(M_{\text{cut}})]$ is the statistic of the Lilliefors test (Lilliefors 1969) and N_{cut} is the sample size above M_{cut} ;

(ii) the probability $P_W(M_{\text{cut}})$ of exceeding $W(M_{\text{cut}})$, under the hypothesis that magnitudes above M_{cut} follow a geometric distribution with parameter $\bar{b}_{\text{ND}}(M_{\text{cut}})$. This probability is independent of sample size and is empirically computed as a function of b , in Lombardi (2024).

Given a prefixed significance level α , here fixed to 0.01, the completeness magnitude for each resampling catalogue is computed by selecting the lowest magnitude BMc for which $P_W(\text{BMc}) > \alpha$. The empirical distribution of values BMc represents the uncertainty about the completeness magnitude M_c of the sample; therefore, the $(1 - \alpha) \times 100$ -th percentile of values BMc gives the value of M_c , and, subsequently, of b , with a significance level equal to α .

A7. Algorithms 11–18 (G2)

(M11_GoF95 per cent_ceMLE, M12_GoF95 per cent_gMLE, M13_GoF90 per cent_ceMLE, M14_GoF90 per cent_gMLE, M15_MaxC_ceMLE, M16_MaxC_gMLE, M17_MGoF_ceMLE, M18_MGoF_gMLE)

Algorithms from 11 to 18, are obtained by combining four methods to estimate M_c

- the GFT method at 90 per cent confidence level (Wiemer & Wyss 2000; subsection A4)
- the GFT method at 95 per cent confidence level (Wiemer & Wyss 2000; subsection A4)
- the MaxC method (Wiemer & Wyss 2000; subsection A2)
- the modified GFT method (Leptokaropoulos *et al.* 2013)

with two methods to estimate b

- the corrected exponential MLE (Aki. 1965; Utsu 1965; eq. A2)
- the geometric MLE (Tinti & Mulargia 1987; eq A6).

The modified GFT method substitutes, for each magnitude cutoff M_{cut} , the theoretical cumulative numbers of events with the empirical distributions obtained by simulated catalogues (Leptokaropoulos *et al.* 2013). The mean value of the difference between the real and the synthetic cumulative distributions is computed as a function of M_{cut} and Mc is selected as the magnitude threshold giving the lower mean difference.

A8. Algorithm 19 (G1): the variability coefficient method

(G1, M19_CV_ceMLE)

The variability coefficient method (Godano *et al.* 2023) is structured as follows.

For the ascending magnitude cutoff M_{cut} , the variability coefficient is defined (for bins with at least 100 events) by

$$CV(M_{\text{cut}}) = \frac{\sigma_{\tilde{M}}}{\mu_{\tilde{M}}}, \quad (\text{A7})$$

where $\tilde{M} = M - M_{\text{cut}}$ are rescaled magnitudes and $\mu_{\tilde{M}}$ and $\sigma_{\tilde{M}}$ are the mean and the standard deviation of $\tilde{M}$, respectively. If $\tilde{M}$ follows an exponential distribution, the theoretical value of CV is 1. Therefore, CV assumes a value $\simeq 1$ at $M_{\text{cut}} = Mc$. Considering the empirical fluctuation of the CV for exponential samples, a threshold of 0.93 is reliable for identifying exponential data (Godano *et al.* 2024); therefore, Mc is defined as the lower value of M_{cut} for which $CV(M_{\text{cut}}) \geq 0.93$.

The b -value is computed by the corrected exponential MLE (Aki 1965; Utsu 1965; eq. A2), for subcatalogues containing at least 50 events above Mc and covering a magnitude range above 1.5.