Bulletin of Earthquake Engineering (2021) 19:2325–2342
https://doi.org/10.1007/s10518-021-01064-6

ORIGINAL ARTICLE

Estimation of MCS intensity for Italy from high quality accelerometric data, using GMICEs and Gaussian Naïve Bayes Classifiers

Laura Cataldi[1] · Lara Tiberi[1] · Giovanni Costa[1]

Received: 17 July 2020 / Accepted: 12 February 2021 / Published online: 9 March 2021

Abstract

Macroseismic intensity provides a qualitative description of seismic damage. It can be associated with Ground Motion Parameters (GMPs), which are extracted in near real-time from instrumental recordings during an earthquake. Several formulations of this empirical association exist in literature for Italy, mainly focusing on the relationship between intensity expressed on the Mercalli-Cancani-Sieberg (MCS) scale and peak ground acceleration or velocity. They are usually in the form of Ground Motion to Intensity Conversion Equations (GMICEs), which treat intensity as a continuous quantity. We propose an alternative approach, the Gaussian Naïve Bayes (GNB) classifiers, which allows to correctly treat intensity according to its ordinal definition. As a comparison, we also implement a modified version of the standard GMICE approach. We expand the existing database of GMP/MCS-intensity points with new, high-quality accelerometric data recorded in Italy in the period from 2002 to 2016 and resample the database by treating the intermediate intensities with half integer values (which are not meaningful in the MCS description) as both belong to the above and below full integer classes with an assigned weight. As a result, we estimate a new set of regression relations and GNB probability distributions between integer MCS intensity classes and eight GMPs (peak acceleration, velocity, displacement, Arias and Housner intensities, spectral acceleration at 0.3, 1.0 and 3.0 s). Results based on PGA and PGV are the most stable on the whole intensity scale. GNB models score better than GMICEs in terms of performance on unseen data and classification scores.

Keywords Earthquake ground motions · Macroseismic Intensity · Probability distributions · Bayesian classifier

Lara Tiberi: Formerly at SeisRaM group, Department of Mathematics and Geosciences, University of Trieste, Trieste, Italy.

✉ Laura Cataldi
laura.cataldi@phd.units.it

[1] SeisRaM group, Department of Mathematics and Geosciences, University of Trieste, Via Weiss 4, 34128 Trieste, Italy

1 Introduction

After the occurrence of a large earthquake, the aim of the civil protection unit is to rapidly assess spatial distribution of damage levels with special attention to highest degrees. Macroseismic intensity can be used to provide a qualitative description of such consequences of a seismic event. Reliable and updated tools to immediately associate the ground motion information with expected macroseismic intensity can thus play an essential role in the timely implementation of civil defence emergency plans.

By definition, macroseismic intensity is an ordinal quantity, which means that it has natural, ordered categories and the distances between the categories are not known (Agresti 2013; Kuehn and Scherbaum 2010). Intensity classes are expressed as roman integer values and are always defined as a "collective" measurement, coming from the observation of many factors which are not linearly dependent on any single, directly measurable value. The first main implication of this fact is that the classes are not proportional to one another, meaning that there is no assurance that the effects observed for a degree II are two times those of a degree I (while for example we can exactly define the proportionality between the energy released by a $M_W = 4.0$ and a $M_W = 5.0$ earthquake). This also implies that intensity measures have a high error content, pushing strongly towards the impossibility of interpreting decimal values as an improvement in the actual intensity estimate.

Even so, Ground Motion to Intensity Conversion Equations (GMICEs) are the most common choice in defining instrumental intensity as a function of Ground Motion Parameters (GMPs). Many GMICEs have been estimated for Italy in the past (e.g., Margottini et al. 1992; Faccioli and Cauzzi 2006; Faenza and Michelini 2010; Caprio et al. 2015; Gomez-Capera et al. 2020). They are usually obtained in a simple linear regression form:

$$I = a + b \log X, \tag{1}$$

where I is the intensity and X is the GMP. In GMICE-related literature, it is common practice to use this functional form as it is, which implies treating intensity as a continuous quantity and obtaining decimal intensity values as a forecast.

In the light of the very definition of macroseismic intensity classes, the current formulation in (1) is not appropriate. With this reasoning in mind, we wanted to obtain GMICEs for Italy that are more compliant to the intensity scale by applying pre- and post-processing to the data, in order to use integer classes only. We also decided to propose a radically alternative description of intensity in terms of a direct probabilistic estimate through the Naïve Bayes classification. This approach has the benefit of treating intensity as a discrete variable throughout the whole definition procedure. It is also a machine learning oriented procedure that can be easily updated as more and new data become available.

In fact, the increase in the number of seismological stations and in the instrumentation quality and the occurrence of the latest destructive earthquakes have already provided new high quality data for a better definition of instrumental intensity for the Italian case. In particular, in this study, we re-elaborated the Faenza and Michelini (2010) dataset with the addition of 82 new data points related to 18 events which occurred in the time-span from 2002 to 2016 in Italy, using high quality accelerometric data. Such data points consist in GMP/MCS-intensity data couples obtained by coupling each expert-assessed intensity value with the nearest available waveform in a 3-km radius. The choice of using only high-quality macroseismic data mitigates the possible bias introduced with the use of the nearest value (cf. Lesueur et al. 2013).

We tested a set of eight ground motion parameters. For each one we present both our improved version of the GMICE, as a comparison, and the Naïve Bayes classification, as a suggested best practice.

2 Input data

The accelerometric input database consists of two parts, according to the availability and ownership of the data.

The first part of the dataset is composed of 72 analysed events from 1972 to 2004 (Table SI1 in the Online Resource), with the lowest local magnitude of 3.4, corresponding to 193 GMP-intensity pairs. These are part of the set of 87 events (266 GMP-intensity pairs) used by Faenza and Michelini (2010) to derive the GMICEs used to date for the Italian territory. Since part of these older recordings generally shows low quality, originally being analog traces with no information on the starting time, we conducted a thorough analysis in order to discard all the recordings with multiple events, or evident artificial signals arising from analog-to-digital conversion or with no clearly identifiable peaks in acceleration, velocity or displacement. Data for the remaining 72 events are taken from the ITACA 2.0 database (ITalian ACcelerometric Archive, version 2.0; Luzi et al. 2008; Pacor et al. 2011), and now all belong to the RAN network (Rete Accelerometrica Nazionale; Gorini et al. 2010; Costa et al. 2015), managed by the DPC (Italian National Civil Protection).

The second part of the database consists of a selection of 18 events (82 GMP-intensity couples) from 2002 to 2016, with 3.4 chosen as the lowest local magnitude, and which comes from the dataset used by Tiberi et al. (2018). These high-quality accelerometric data were collected by the CE3RN (Central Eastern European Earthquake and Research Network; Costa et al. 2010; Bragato et al. 2014) and RAN stations.

The associated macroseismic intensity data-points are taken from the 2015 version of the Italian Macroseismic Database (DBMI15; Locati et al. 2016), except for data related to the two 2016 events (Table SI1) that comes from QUEST reports taken for the Ibleo (Azzaro et al. 2016) and Amatrice (Galli et al. 2016; Tertulliani and Azzaro 2016) earthquakes. All intensity values are expressed in MCS scale. In order to exclude cumulative damage, we only took into consideration the main shocks.

The scope of the work is to improve intensity forecasts also for medium–high damage levels ($I \geq$ VI), where they become of interest for civil defence purposes. For this reason, and to guarantee the homogeneity of the database, we only considered macroseismic intensity measures issued from expert surveys. Furthermore, the inclusion of crowdsourced intensity data is not a trivial process and would go beyond the scope of this work.

The investigated set of ground motion parameters consists in peak ground acceleration (PGA), velocity (PGV), displacement (PGD), Arias intensity (I_A), Housner intensity (I_H), spectral acceleration at 0.3 s (PSA03), 1.0 s (PSA10) or 3.0 s (PSA30). The ground motion parameters for all events were calculated using a near-real-time procedure developed at the Department of Mathematics and Geosciences of the University of Trieste (Gallo et al. 2014), in order to process the signals in an as homogeneous as possible way. Its main features include a Butterworth filtering between 0.1 and 50 Hz, with range automatically selected based on the signal to noise ratio, and a trend removal used to compute PGV and PGD values. For older waveforms, when the pre-event noise trace was not available, we used fixed filter frequency values taken from the ITACA database.

GMPs were taken from the maximum between the two horizontal component values. The parameters and the observed intensity values were associated using the minimum distance criterion, with a maximum distance limit of 3 km. The complete database thus counts 90 events (Fig. 1) in the time-window from 1972 to 2016, corresponding to 275 associated GMP-intensity pairs (Table SI1 of the Online Resource). Intensity values range between II and X; the epicentral distances range between 1.6 and 150.7 km (Fig. 2) and are well distributed, especially for the central values of intensities (IV-VI).

It is common practice to treat intensity as a continuous value, for example by using half integer classes in assigning uncertain MCS intensity values, so this kind of data is widely present in dedicated macroseimic surveys and is also found in the Italian Macroseismic Database. Following Kuehn and Scherbaum (2010), in order to be consistent with the class definitions given by the MCS scale, we included this uncertainty information in the data by re-assigning the half integer values to the nearby integer classes, with the use of some weights, so that integer classes only will be used in the calculations. In particular, all data originally corresponding to half integer classes were assigned both to

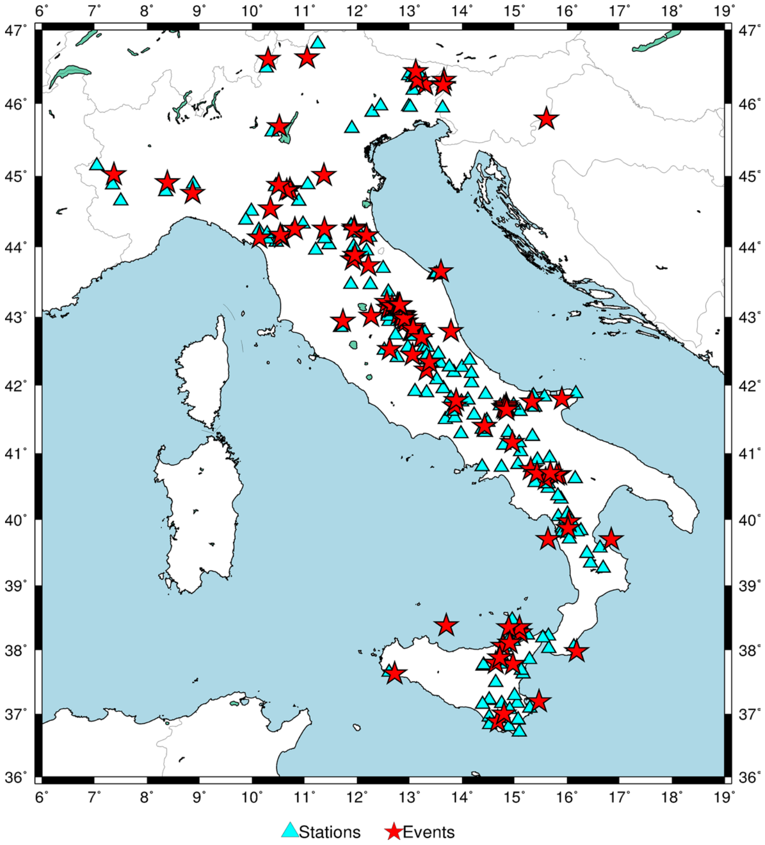


Fig. 1 Data set used for the definition of instrumental intensity: the red stars are the epicentral locations of the analysed events; the cyan triangles are the station sites for which the GMPs are estimated, and with an associated observed intensity value

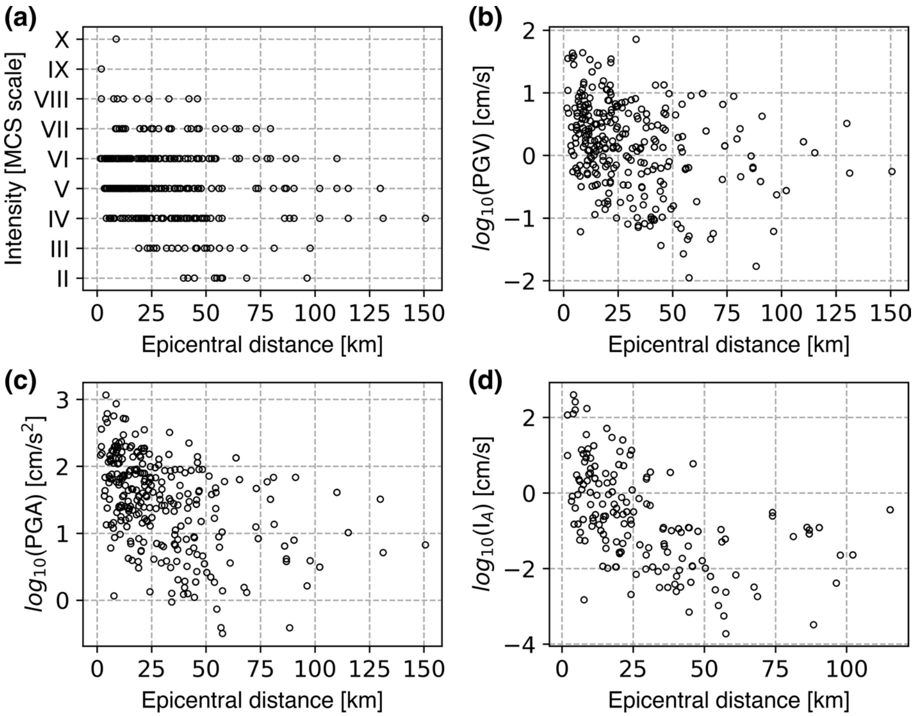
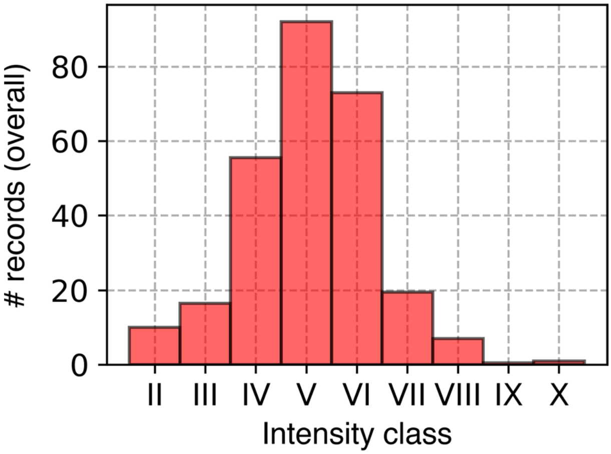


Fig. 2 Coverage of the intensity points in distance **a** and for parameters PGV **b**, PGA **c** and I_A **d**. The distance event-station is the epicentral distance between the epicentre and the station site in km. See Fig. SI1 in the Online Resource for the remaining parameters

the above and below integer classes, with a weight $w = 0.5$, whereas all data originally corresponding to integer classes were assigned a weight $w = 1$.

The resulting weighted data distribution in terms of intensity classes is shown in Fig. 3 (cf. Fig. SI2 of the Online Resource; for practical reasons, from here on only PGV, PGA

Fig. 3 Distribution of the GMP values binned into classes at integer intensity intervals



and I_A are explicitly shown in the figures, while we provide results for all studied parameters in the Online Resource). For the peak parameters (PGA, PGV, PGD) there are 376 points in the weighted database (of which, 174 with $w = 1$); for Arias intensity, PSA03 and PSA10, 220 points (100 with $w = 1$); for Housner intensity, 200 points (94 with $w = 1$), and for PSA30, 195 points (91 with $w = 1$). The difference in the number of available points for each parameter mainly comes from the fact that, in the case of partial or cut recordings, the extracted parameters were limited to peak amplitudes (PGA, PGV, PGD) and all integral quantities were discarded. The main reason behind this choice is the risk of underestimating the integral parameters due to missing part of the record. In the case of Housner intensity and PSAs, moreover, we also discarded the cases for which the used high-pass filter was so high that it would filter out the frequency values used in the parameter calculation.

3 GMICEs

Linear regressions are the most common tool to define instrumental intensity. Even so, they treat intensity as a continuous numerical value; therefore, predicted outcomes are not directly meaningful and either have to be rounded to the nearest integer value, or to be interpreted as reflecting an uncertainty between two intensity classes. For this reason, we calculated an updated version of GMICEs for Italy in the most intensity-compliant way, to confront the results with those obtained via the more rigorous GNB methodology. We decided to keep the log-linear functional form itself (Eq. 1) in performing the regression, but we applied some pre- and post-processing in order to take into consideration the caveats discussed so far. The first part of the pre-processing is described in detail in the Input Data section; as a form of post-processing, we rounded up the resulting forecast values to the nearest integer.

In order to consider both the dependent and the independent variables as affected by sampling variability, which is more correct given the nature of our data, we calculated the GMICEs by using the ODR methodology (Boggs et al. 1988; *odr* in scipy.org). ODR is a common technique for fitting data to models and we used it to extract the intercept and gradient parameters (a, b in Eq. 1). This algorithm minimizes the weighted orthogonal distances from the curve, taking into consideration both the vertical (σ_y) and horizontal (σ_x) uncertainties. We used ODR in its simplest form, assuming that the ratio of the standard deviation of the errors on dependent and independent data (σ_y/σ_x) is known and fixed. This also makes it possible to directly invert the relation, so the regression coefficients could be likewise used to express the GMPs as a function of intensity.

3.1 Application

Both intensity and strong motion data are characterised by an intrinsically high spatial variability. As for intensity, this fact has been addressed by defining macroseismic classes as a collective measurement, subtracting to the meaningfulness of "punctual" measures. On the other side, instrumental data are intrinsically punctual and strongly connected to specific, local geological conditions. The polar character of intensities also adds to the disequilibrium in the input dataset, mostly concentrated in the lower-central classes (IV-VI). A possible solution to address this variability is to perform a preliminary smoothing of data to filter out effects related to regional variability, random components, and geological conditions.

We decided to follow the approach proposed by Faenza and Michelini (2010) and to bin the GMP-intensity couples into integer intensity classes as a form of smoothing of the instrumental data. The regressions are thus performed on the mean of the logarithmic GMP values in each bin. This choice also depends on the distribution of the GMP standardised values with zero mean and a unit standard deviation, represented in Fig. 4 (cf. Fig. SI3 of the Online Resource). It is clearly visible that after the application of the logarithm in base 10 the data follow the normal Gaussian curves, allowing us to estimate the intrinsic variability of ground motion data for each class.

We calculated the mean values and the standard deviations of these distributions and used them in the GMICE inversions. In our case, in particular, we estimated the mean values μ_k^* as the weighted arithmetic mean of the logarithm in base 10 of the parameter ($\log X$), for each intensity class k between II and X:

$$\mu_k^* = \frac{\sum_{j=1}^{N_k} w_{jk} \log X_{jk}}{\sum_{j=1}^{N_k} w_{jk}}, \tag{2}$$

where w_{jk} is the weight assigned to the j-th point among the N_k data points with $I_k = k$. As for the associated errors, we must take into consideration that there is an evident lack of data for some classes with respect to others (e.g. classes IX and X versus class V; cf. Figure 3). For this reason, it was not possible to provide a robust estimate of the regular standard deviation for those classes, which would turn out very low or even null and would not reflect the actual distribution of the underlying data. The use of such standard deviation values would also excessively push the noise filtering resulting from the binning, leading to an artificial increase in the statistical parameters related to the goodness of the fit. Following the approach proposed by Kuehn and Scherbaum (2010), we thus estimated a standard deviation σ_{CSD} common to all intensity classes for each GMP, as the square root of the pooled variance:

$$\sigma_{CSD} = \sqrt{\frac{\sum_k \sum_{j=1}^{N_k} \left(\log X_{jk} - \mu_k^*\right)^2}{\left(\sum_{j=1}^{k} N_j\right) - 9}} = \sqrt{\frac{\sum_k \sum_{j=1}^{N_k} \left(\log X_{jk} - \mu_k^*\right)^2}{N_{tot} - 9}}, \tag{3}$$

where we divided by the total number of samples (N_{tot}) minus the number of different classes in which data were binned, nine in our case. Values of μ_k^* and σ_{CSD} for each parameter are reported in the Online Resource (cf. Table SI2 of the Online Resource). The

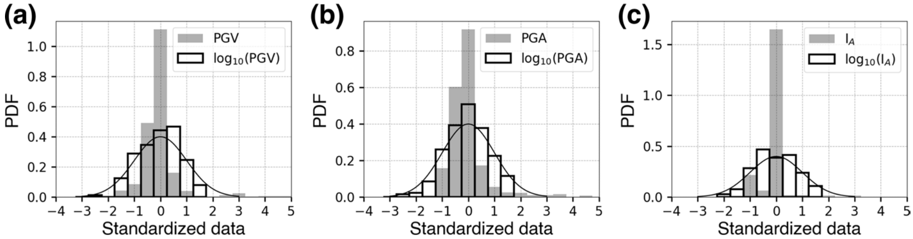


Fig. 4 Probability distribution functions for PGV **a**, PGA **b** and $\mathrm{I_A}$ **c**. In each graph the standardised data with zero mean and unit standard deviation are represented, in grey for the original data and as black boxes for the logarithm in base 10 of those. As a reference, the Gaussian normal distribution is depicted (solid black line)

corresponding intensity values I are also assigned an error σ_I to account for the dispersion of data. We tested different binning on the GMP data in order to check the corresponding discrete distribution of intensity values and decided to adopt a conservative value of $\sigma_I = 1.0$ as a common standard deviation associated with all intensity classes and all parameters.

For each parameter, GMICEs were thus calculated on nine data couples $(\mu_k^*, I_k = k)$, with associated errors (σ_{CSD}, σ_I) and intensity classes ranging from II to X (cf. Fig. SI4 for a graphic representation). Regression parameters are reported in Table 1. To allow for a qualitative comparison of our results, for each equation we calculated the R squared value (R^2), representing the proportion of explained variance of I, and the standard deviation of the bins (σ) and of the data (σ_d). The standard deviation of the bins is defined as:

$$\sigma = \sqrt{\frac{\sum_{k=2}^{10} \varepsilon_k^2}{9-2}}, \tag{4}$$

where $\varepsilon_k = I_k - \widehat{I_k}$ is the residual between predicted intensity value ($\widehat{I_k}$) and true intensity value (I_k) corresponding to μ_k^*. Due to the low sample population, we used a reduced form where the number of intensity points used in the regression (nine) is reduced by the number of fitted parameters (a and b in Eq. 1). Just like R^2, σ depends on residuals calculated on the binned dataset $\left(\varepsilon_k\right)$. For this reason, it does not fully catch the actual underlying variability in I, and its values are way lower than the prior ones assigned to the input data ($\sigma_I = 1.0$). Following Gomez-Capera et al. (2020), we also calculated the standard deviation of the data σ_d:

$$\sigma_d = \sqrt{\frac{\sum_{n=1}^{N_{tot}} \left(I_n - \widehat{I_n}\right)^2}{N_{tot} - 1}}, \tag{5}$$

where $I_n - \widehat{I_n}$ is the residual calculated for the n-th input point. Values of R^2, σ and σ_d are reported in Table 1. Obtained σ_d values are close to 1 and provide a better measure of the variability in I for a given GMP value. Even so, since I is an ordinal variable, they cannot be used as-are and require some degree of interpretation. One possibility is to define a probability associated to each $\widehat{I_n}$, in the form of a Gaussian distribution centred on the forecasted intensity and with σ_d as standard deviation (cf. Sect. 5.1).

Table 1 The resulting GMICEs using the ODR algorithm for each GMP studied with the associated R^2 and the standard deviation values of the bins (σ) and of the data (σ$_d$)

	a	b	R^2	σ	σ$_d$	#Records
GMP						
PGD	7.01 ± 0.17	2.33 ± 0.15	0.97	0.49	1.24	376
PGV	4.96 ± 0.17	2.65 ± 0.16	0.97	0.47	1.19	376
PGA	1.32 ± 0.35	2.85 ± 0.19	0.97	0.51	1.36	376
I_A	5.63 ± 0.23	1.46 ± 0.13	0.95	0.67	1.22	220
I_H	3.58 ± 0.30	2.46 ± 0.21	0.95	0.66	1.20	200
PSA03	0.65 ± 0.56	2.69 ± 0.25	0.94	0.73	1.32	220
PSA10	2.73 ± 0.35	2.41 ± 0.20	0.95	0.64	1.28	220
PSA30	4.78 ± 0.27	2.31 ± 0.22	0.94	0.74	1.31	195

4 Gaussian Naïve Bayes Classifiers

Linear regressions treat intensity as a continuous numerical value instead of an ordinal one; therefore, predicted outcomes are not directly meaningful and either have to be rounded to the nearest integer value, or be interpreted as reflecting an uncertainty between two intensity classes. To correctly handle ordinal data throughout the whole inversion process it is possible to apply a different method, the Gaussian Naive Bayes classification (GNB), which estimates a discrete conditional probability distribution $\Pr(I|X)$ linking the (ordinal) intensity I to any GMP X (Pedregosa et al. 2011). The goal is to obtain an alternative way to express forecasts in the form of an ordinal instrumental intensity value, with a known associated probability.

GNB is part of a set of supervised learning algorithms based on applying Bayes' theorem in the Naïve form, that is, with the assumption of conditional independence between every pair of features given the value of the class variable (Zhang 2004). In our particular case, having considered only one feature at a time, it coincides with the full Bayes theorem. We hereby give a synthetic overview of the procedure; for more details on the procedure itself and on the underlying statistics, we refer the reader to Lancieri et al. (2015) and references therein.

For any variable X taken among the eight selected GMPs, and the categorical variable I which is dependent on variable X, a Naïve Bayes classifier predicts the conditional probability distribution of I given $\log X$ by using Bayes' rule:

$$\Pr(I|\log X) = \frac{\Pr(\log X|I)\Pr(I)}{\Pr(\log X)}. \tag{6}$$

According to Bayes' rule, $\Pr(\log X|I)$ is the conditional probability of observing $\log X$ on class I, and $\Pr(I)$ and $\Pr(\log X)$ are the a priori probabilities for I and $\log X$, respectively. In this specific context, the probability of having intensity class k when the variable X takes the value x_i can be expressed as:

$$\Pr\left(I = k|\log X = \log x_i\right) = \frac{\Pr\left(\log X = \log x_i|I = k\right)\Pr(I = k)}{\sum_j \Pr\left(\log X = \log x_i|I = j\right)\Pr(I = j)}, \tag{7}$$

where summation over j covers the whole event space, i.e. all possible intensity classes. We should stress how the name Bayesian classifiers comes from the use of the Bayes' theorem, but does not automatically imply the use of Bayesian inference. In principle, it would be possible to define prior distributions and estimate the parameters using Bayesian inference; in fact, following Kuehn and Scherbaum (2010), all parameters were empirically estimated by maximum likelihood. $\Pr(I = k)$ was learnt from the data as the relative frequency of the classes observed on the dataset:

$$\Pr(I = k) = \frac{N_k}{N_{tot}}, \tag{8}$$

where N_k is the number of data in class $I = k$ and N_{tot} is the total number of data. As for the conditional probability $\Pr(\log X|I)$, we used the normal distribution already estimated from the dataset for each intensity class k, with a mean value of μ_k^* and common standard deviation σ_{CSD} (cf. Sect. 2):

$$\Pr(\log X = \log x_i | I = k) = \frac{1}{\sqrt{2\pi\sigma_{CSD}^2}} e^{-\frac{\left(\log x_i - \mu_k^*\right)^2}{2\sigma_{CSD}^2}}. \quad (9)$$

By using GNB, we fit the probabilities on the whole dataset to obtain a discrete conditional probability distribution on all intensity classes for each input GMP value. We then chose to select the class with the highest associated probability value as the best estimate of *I*.

4.1 Application

We applied the Python algorithm *pomegranate* (Schreiber 2018) to perform GNB classification on the whole dataset. As opposed to the regression procedure, we performed the calculations without binning the data and only used information from the binned database in the form of the parameters μ_k^* and σ_{CSD} to be inputted in Eq. 9. Figure 5 shows an example of the corresponding conditional probability distribution $\Pr(\log X = \log x_i | I = k)$ for each intensity class for the PGA parameter.

The resulting intensity predictions are plotted in Fig. 6 (cf. Fig. SI5 of the Online Resource), colour-coded from lower (white) to higher (black) associated probabilities. They are obtained by applying the model to a linear space covering all values of the input parameters; for each value on the x-axis, the corresponding colour-coded probability values along the vertical (intensity) axis sum up to one. For each parameter, the resulting ODR equation is reported for comparison.

5 Appraisal of the results and discussion

5.1 Performance on unseen data

The best way to assess the performance and reliability of the resulting intensity predictions, both from ODR and GNB, would be to test them on an ‘unseen’ dataset, different from the one used to extract them. In our case, the available database itself does not contain enough data to properly build both a training set and a testing set, so we resorted to Leave-One-Out cross-validation (LOOCV) as a proxy to assess the equations performance on unseen data.

LOOCV works by repeatedly dividing the whole dataset into two subset: the one used to train the equations, containing N-1 points, plus a single point which is left out to be used

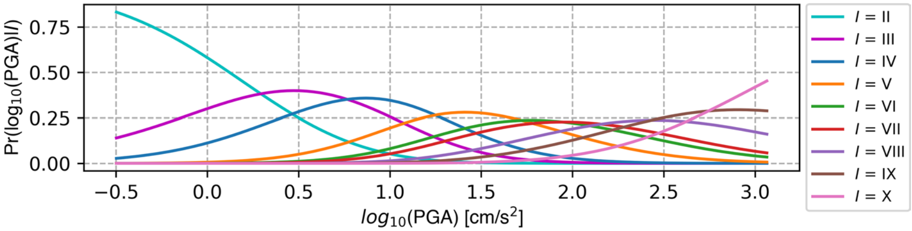


Fig. 5 Example of the conditional probability distribution $\Pr(\log X = \log x_i | I = k)$ for the PGA parameter, for each intensity class, used in the GNB Classification fitting

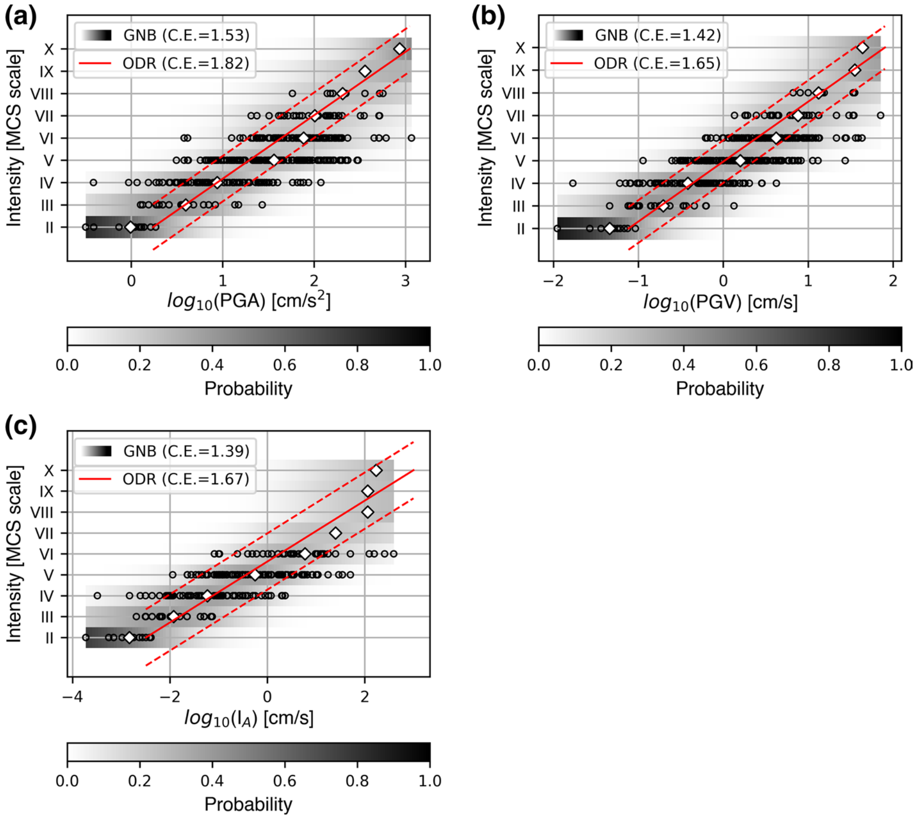


Fig. 6 Probability distributions obtained from GNB Classifiers (grey scale) for PGA **a**, PGV **b** and I_A **c**. The ODR equation with associated $\pm 2\sigma$ error is reported for comparison (red lines), together with the mean GMP values used to derive the equations (white diamonds) and the underlying dataset (black circles). For each value on the x-axis, the corresponding probability values along the vertical axis sum up to one

for validation. We constrained our LOO system so that only points associated to integer intensity values (i.e. weight equal to one) would be left out as a test case. We dropped each of these points in turn, performed the regressions on the remaining data, and used them to estimate intensity on the left-out data point ($\hat{I}$). The classification ability with respect to the actual values (I) was then scored using the Cross-Entropy loss function $\mathcal{L}$ (C.E.; also called log-loss):

$$\mathcal{L}[P] = -\frac{1}{N}\sum_{c} \delta_{o,c} \log\left(P_{o,c}\right), \tag{10}$$

where $\delta_{o,c}$ is 1 if the intensity value of observation o belongs to class c and 0 otherwise, and $P_{o,c}$ is the predicted probability that observation o has intensity class c.

We chose to use the Cross-Entropy loss as it takes into consideration the probability associated to each intensity class, which should reflect the intrinsic variability in intensity values for a given ground-motion input. This allows to compare models not only on their average classification ability, but also on how well they capture the uncertainty. C.E. loss score increases as the predicted probability deviates from the actual label; it would be 0 for a perfect model. Notice that, in the case of ODR forecasts, the predicted probability was estimated by integrating the normal density centred on $\hat{I}$ over the interval $[I - 0.5, I + 0.5]$.

The resulting C.E. loss scores are reported in Table 2. GNB classifier models score better than ODR regressions for all ground motion parameters, indicating an overall better performance of the GNB models. Note that among other parameters the equations regarding PGV (in agreement with Kuehn and Scherbaum 2010), PSA03, I_A and I_H provide the best performances.

5.2 Spectral parameters

We should stress that, even if both methodologies have a way to address the weakness arising from less populated classes, models obtained from spectral parameters are still resenting the lack of data in high intensity classes ($I >$ VII). As explained in Sect. 2, our database included many older, triggered waveforms, for which only the peak amplitude parameters could be extracted without risk of underestimation. This led us to using a less populated database for the case of spectral parameters. We can see from the distribution of such data (e.g. Fig. 6c) that it particularly lacks in high intensity classes, which could lead to inconsistencies in the related forecasts. This holds true for both methodologies, and is simply more evident in the case of GNB where it translates into not well resolved probability values. For this reason, we advise that only the resulting models for PGA and PGV should be used for forecasts.

5.3 Sensitivity study

We tested both GNB and ODR models on the training dataset (described in Sect. 2) by comparing the predicted classes with the observed ones, to check in which data ranges each performed better. Results are shown in form of weighted confusion matrices in Fig. 7: the elements distributed along the highlighted diagonal are the number of data correctly categorized, while the off-diagonal elements are the misclassified data. GNB models provide more realistic outcomes for all classes with respect to ODR models, which also tend to a class overestimation (more elements on the right side of the diagonal). In both cases, PGV-based classification is more robust than the PGA-based one.

5.4 Application of GNB forecasts

In order to be directly applicable into shaking intensity maps, GNB classification models have to be converted to GMICE-like objects. We assigned a single instrumental intensity value to each input ground motion parameter value in the database range. The forecast is

Table 2 The resulting Cross-Entropy (C.E.) loss scores for ODR regressions and for GNB Classifiers. Lower values correspond to better classification

GMP							
PGD	PGV	PGA	I_A	I_H	PSA03	PSA10	PSA30
C.E. for ODR							
1.70	1.65	1.82	1.67	1.71	1.80	1.75	1.77
C.E. for GNB							
1.49	1.42	1.53	1.39	1.38	1.42	1.44	1.50

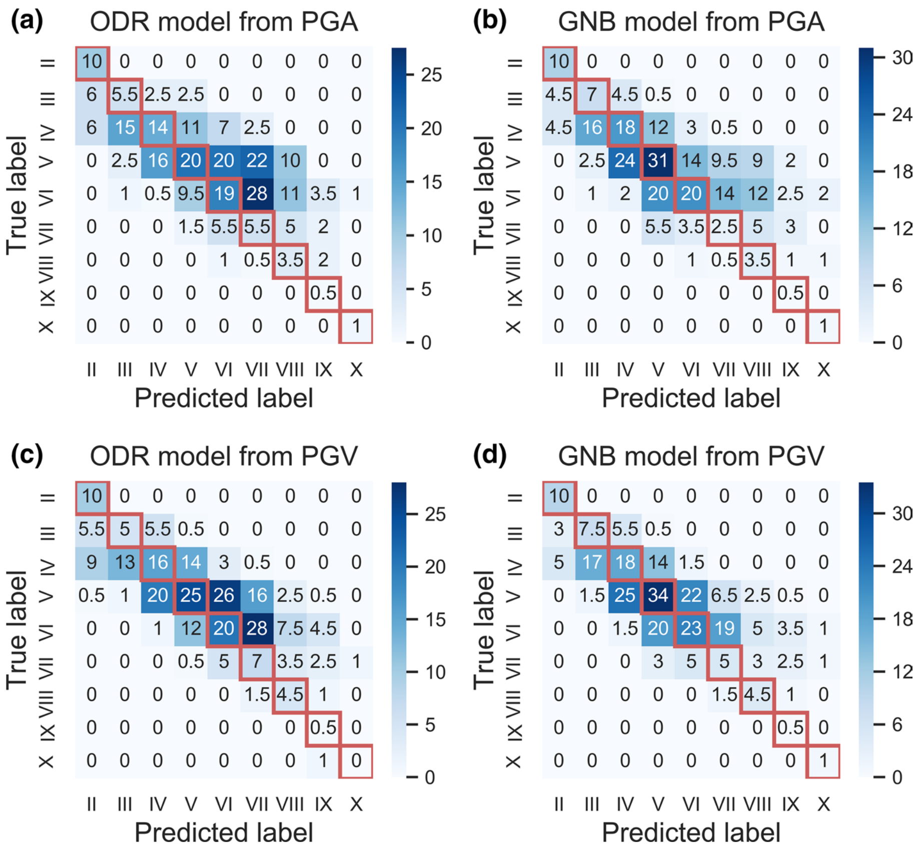


Fig. 7 Method classification on the training dataset in the form of weighted confusion matrices: on the y-axis the true (observed) class label and on the x-axis the predicted one are reported. The elements distributed along the highlighted diagonal are the number of data correctly categorized. Results refer to the ODR **a** and GNB **b** methods based on the PGA parameter (upper panels) and for the ODR **c** and GNB **d** methods based on the PGV parameter (lower panels)

chosen as the class with the highest associated probability (corresponding to the darkest colour in Fig. 6). The result is a linear trend that is comparable to the regression output and that can be used as a guide in defining parameter ranges for each instrumental intensity class (cf. Table 3). Results for the PGA and PGV cases are reported in Fig. 8.

5.5 Comparison of GMICEs for Italy

We compared the empirical GMICEs obtained in this study (which use integer classes only and are somehow more compliant to the MCS intensity scale) with the relationships reported by Gomez-Capera et al. (2020), Caprio et al. (2015), Faenza and Michelini (2010), and Faccioli and Cauzzi (2006). A summary of the characterizing parameters is reported in Tables 4 and 5, for the PGV and PGA cases respectively.

The relations are consistent to each other inside the common standard deviation values estimated with our dataset (Fig. 9). The main difference is in the reliability and range of

Table 3 Table of PGA and PGV value intervals for calculation of intensity *I*

	PGA_{min} (cm/s^2)	PGA_{max} (cm/s^2)	PGV_{min} (cm/s)	PGV_{max} (cm/s)
I_{MCS}				
II	0.32	1.91	0.01	0.10
III	1.91	6.31	0.10	0.28
IV	6.31	17.78	0.28	0.74
V	17.78	52.48	0.74	2.57
VI	52.48	85.11	2.57	5.75
VII	85.11	141.25	5.75	9.77
VIII	141.25	269.15	9.77	21.38
IX	269.15	575.44	21.38	39.81
X	575.44	1148.15	39.81	70.79

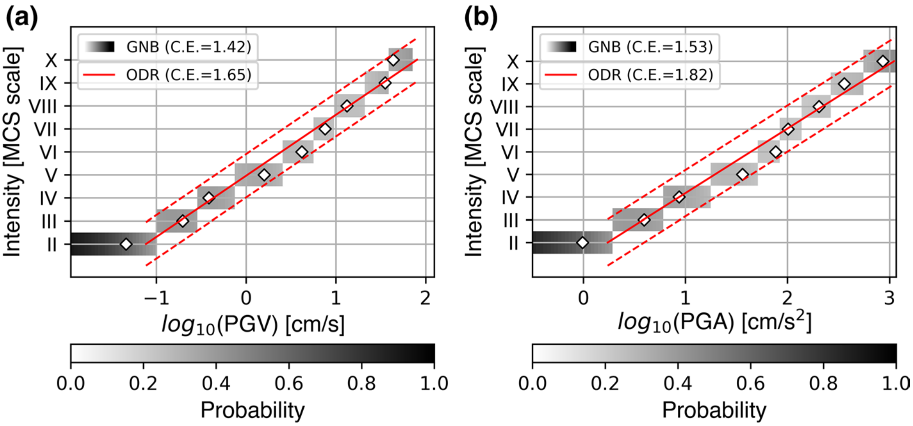


Fig. 8 Intensity classes with highest associated GNB probability (grey scale) for each PGV **a** and PGA **b** value in the database range. The ODR equation with associated ±2σ error is reported for comparison (red lines), together with the mean GMP values used to derive the equations (white diamonds)

Table 4 Comparison with PGV–Intensity relationships proposed by previous studies

Author	Relationship	R^2	σ	I range
This study (ODR)	$I_{MCS} = (4.96 \pm 0.17) + (2.65 \pm 0.16) * \log_{10}$ PGV	0.97	0.47	II to X
Faccioli and Cauzzi (2006)	$I_{MCS} = (5.09 \pm 0.22) + (1.80 \pm 0.17) * \log_{10}$ PGV	0.61	0.71	V to VIII–IX
Faenza and Michelini (2010)	$I_{MCS} = (5.11 \pm 0.07) + (2.35 \pm 0.09) * \log_{10}$ PGV	–	0.26	II to VIII
Caprio et al. (2015)	$I_{MCS} = (4.424 + 1.589 * \log_{10}$ PGV) *if* $\log_{10}$ *PGV* ≤ 0.3 $I_{MCS} = (4.018 + 2.671 * \log_{10}$ PGV) *if* $\log_{10}$ *PGV* > 0.3	–	–	II to VIII
Gomez-Capera et al. (2020)	$I_{MCS} = 4.514 * \exp(0.502 * \log_{10}$ PGV)	–	0.36	II to X–XI

Table 5 Comparison with PGA–Intensity relationships proposed by previous studies

Author	Relationship	R^2	σ	I range
This study (ODR)	$I_{MCS} = (1.32 \pm 0.35) + (2.85 \pm 0.19)*\log_{10}$ PGA	0.97	0.51	II to X
Faccioli and Cauzzi (2006)	$I_{MCS} = (2.62 \pm 0.10) + (1.96 \pm 0.29)*\log_{10}$ PGA	0.38	0.89	V to VIII–IX
Faenza and Michelini (2010)	$I_{MCS} = (1.68 \pm 0.22) + (2.58 \pm 0.14)*\log_{10}$ PGA	–	0.35	II to VIII
Caprio et al. (2015)	I_{MCS} = (2.270 + 1.647*$\log_{10}$ PGA) *if* $\log_{10}$ *PGA* < 1.6 $I_{MCS} = (-1.361 + 3.822*\log_{10}$ PGA) *if* $\log_{10}$ *PGA* > *1.6*	–	–	II to VIII
Gomez-Capera et al. (2020)	$I_{MCS} = 2.276*\exp(0.546*\log_{10}$ PGA)	–	0.31	II to X–XI

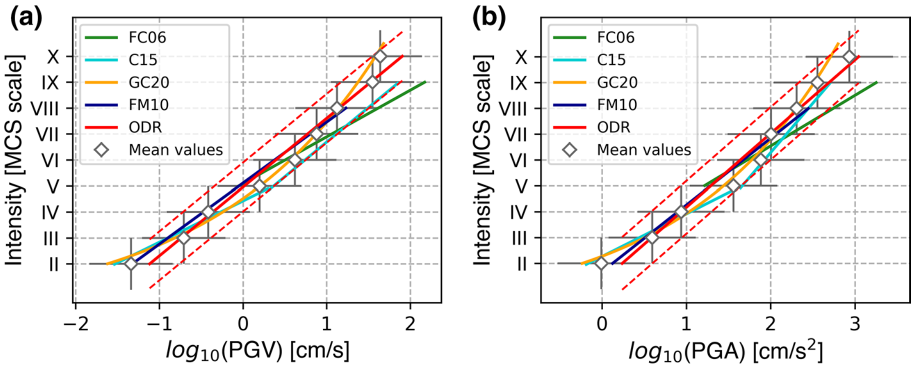


Fig. 9 **a** Comparison of the Intensity—PGV relationship obtained in this study with the ODR algorithm and four previous studies: Faenza and Michelini (2010), FM10; Faccioli and Cauzzi (2006), FC06; Caprio et al. (2015), C15; Gomez-Capera et al. (2020), GC20; **b** Same as (a), for the Intensity—PGA relationship. The dotted red lines are the ± 2σ error associated to the ODR GMICEs

validity of these laws, those estimated in this study having higher values of R^2 and a wider range of validity.

In fact, the resulting equations present high R^2 values for all the studied GMPs (over 0.90), rendering it impossible to indicate which of the parameters provides a better estimate of intensity. The lowest standard deviation of data is associated to the regression line for PGV ($\sigma_d = 1.19$). However, it should be kept in mind that the GNB models should be preferred to the ODR ones in any case, as results from cross validation and sensitivity test confirm.

6 Conclusions

The aim of this study was to provide an updated and more rigorous definition of instrumental macroseismic intensity for Italy. We used integer MCS intensity classes in the range II-X, together with high quality accelerometric data. Data was pre-processed in order to use integer classes only. For each investigated ground motion parameter (PGD, PGV, PGA, I_A, I_H, PSA03, PSA10 and PSA30) we provided both the GMICE formulation, which is more used but less appropriate, and the GNB formulation, which correctly treats intensity as an ordinal quantity.

Out of the eight tested parameters, models based on spectral parameters proved to be too unstable at higher intensity levels; PGA- and PGV-based equations should be used instead.

Overall, the GNB approach should be preferred as more rigorous in treating intensity as a discrete variable throughout its whole procedure. It goes beyond providing a single-valued intensity estimate, as it calculates a full discrete probability distribution for the MCS intensity classes. As a result, GNB-based models show better performance on unseen data and more capability in capturing the uncertainty than GMICEs. Overall, GNB models perform better than ODR ones on the whole considered intensity range, in terms of classification scores.

The possibility to increase the estimate accuracy with respect to the 'standard' GMICEs might be extremely useful in some applications, such as shaking intensity maps. In fact, GMICEs are the default choice in generating ShakeMaps with the USGS-ShakeMap software (Wald et al. 1999). We propose a conversion of GNB models to GMICE-like objects that can be substituted in the ShakeMap procedure.

The GNB-based methodology is a machine learning oriented procedure that can be easily updated as more data is collected. In the era of big data, it can be included in the effort to efficiently analyse incoming data in near-real time. Future work includes testing and calibration of this procedure for both the south-eastern Alps region, following Moratto et al. (2009), and for the Italian territory, as soon as new, independent intensity data on new events becomes available. In particular, it is fundamental to point out that in order to explain the damage in the near fault areas, a more focussed study is needed to expand the research to a combination of ground motion parameters.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10518-021-01064-6.

Acknowledgements This work was co-funded by the Italian Department of Civil Protection—Presidency of the Council of Ministers and supported by the Interreg V-A Italia-Austria 2014-2020 "ARMONIA" project. The authors wish to express their grateful thanks to the Italian Department of Civil Protection—Presidency of the Council of Ministers for the data used in this study. We are grateful to the technician team of the SeisRaM group of the DMG (Piero Falconer, Lorenzo Furlan, Enrico Magrin and Shaula Martinolli) for their good work in the maintenance of the Rete Accelerometrica del Friuli Venezia Giulia (RAF) stations within the CE3RN and the RAN networks. We thank the Interreg ARMONIA project team for its support on GMP selection. We would also like to express our gratitude to Dr. Stefano Parolai (OGS), Professor Claudio Asci (DMG, University of Trieste), Dr. Valerio De Rubeis, Dr. Marta Caprio and two anonymous reviewers, whose useful advice greatly helped us improve this work.

Author's contribution All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Laura Cataldi and Lara Tiberi. The first draft of the manuscript was written by Lara Tiberi; all subsequent writing was carried out by Laura Cataldi. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open access funding provided by Università degli Studi di Trieste within the CRUI-CARE Agreement. This study was partially funded by the Italian Department of Civil Protection—Presidency of the Council of Ministers (DPC) and by the European Regional Development Fund (ERDF) Interreg V-A Italia-Austria 2014–2020 "ARMONIA" project.

Data availability and material The first part of the accelerometric database is available either from the Italian Strong Motion Network (RAN, Rete Accelerometrica Nazionale) download website (http://ran.protezionecivile.it/IT/index.php), managed by the DPC, or from the ITalian ACcelerometric Archive (http://itaca

20.mi.ingv.it/itaca20/). The second part is available from the ORFEUS data centre web interface (https://orfeus-eu.org/webdc3/). Macroseismic data is available from the Italian Macroseismic Database (https://emidius.mi.ingv.it/CPTI15-DBMI15/).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Code availabilty Code is available upon request.

References

Agresti A (2013) Categorical data analysis, 3rd edn. Wiley, Hoboken

Azzaro R, D'Amico S, Mostaccio A, Scarfì L, Tuvè T (2016) RILIEVO MACROSISMICO DEL TERREMOTO IBLEO DELL'8 FEBBRAIO 2016 (Tech. rep.). INGV dell'Osservatorio Etneo—Sezione di Catania. http://www.questingv.it/index.php/rilievi-macrosismici/28-ibleo-08-02-2016-ml-4-6/file

Boggs PT, Spiegelman CH, Donaldson JR, Schnabel RB (1988) A computational examination of orthogonal distance regression. J Econ 38:169–201. https://doi.org/10.1016/0304-4076(88)90032-2

Bragato PL, Costa G, Gallo A, Gosar A, Horn N, Lenhardt W, Mucciarelli M, Pesaresi D, Steiner R, Suhadolc P, Tiberi L, Živčić M, Zoppè G (2014) The central and eastern European earthquake research network-CE3RN. Geophys Res Abstr. https://doi.org/https://doi.org/10.13140/RG.2.1.3507.7843

Caprio M, Tarigan B, Worden CB, Wiemer S, Wald DJ (2015) Ground Motion to Intensity Conversion Equations (GMICES): a global relationship and evaluation of regional dependency. Bull Seismol Soc Am 105(3):1476–4190. https://doi.org/10.1785/0120140286

Costa G, Moratto L, Suhadolc P (2010) The friuli venezia giulia accelerometric network: RAF. Bull Earthq Eng 8:1141–1157. https://doi.org/10.1007/s10518-009-9157-y

Costa G, Ammirati A, De Nardis R, Filippi L, Gallo A, Lavecchia G, Sirignano S, Suhadolc P, Zambonelli E, Nicoletti M (2015) The Italian Strong Motion Network (RAN), near-real time data acquisition and data analysis: a useful tool for seismic risk mitigation. Proceedings of 2ECCES: Special Sessions. https://doi.org/https://doi.org/10.13140/2.1.3513.4722

Faccioli E, Cauzzi C (2006) Macroseismic intensities for seismic scenarios, estimated from instrumentally based correlations. Proc 1st ECEES. Red Hook, NY, pp 3064–3073

Faenza L, Michelini A (2010) Regression analysis of MCS intensity and ground motion parameters in Italy and its application in ShakeMap. Geophys J Int 180:1138–1152. https://doi.org/10.1111/j.1365-246X.2009.04467.x

Galli P, Peronace E, Tertulliani A (2016) Rapporto sugli effetti macrosismici del terremoto del 24 Agosto 2016 di Amatrice in scala MCS (Tech. rep.). DPC, CNR-IGAG, INGV. https://doi.org/10.5281/zenodo.161323

Gallo A, Costa G, Suhadolc P (2014) Near real-time automatic moment magnitude estimation. Bull Earthq Eng 12:185–202. https://doi.org/10.1007/s10518-013-9565-x

Gomez-Capera AA, D'Amico M, Lanzano G, Locati M, Santulin M (2020) Relationships between ground motion parameters and macroseismic intensity for Italy. Bull Earthq Eng 18:5143–5164. https://doi.org/10.1007/s10518-020-00905-0

Gorini A, Nicoletti M, Marsan P, Bianconi R, De Nardis R, Filippi L, Marcucci S, Palma F, Zambonelli E (2010) The Italian strong motion network. Bull Earthq Eng 8:1075–1090. https://doi.org/10.1007/s10518-009-9141-6

Kuehn NM, Scherbaum F (2010) A naive bayes classifier for intensities using peak ground velocity and acceleration. Bull Seism Soc Am 100(6):3278–3283. https://doi.org/10.1785/0120100082

Lancieri M, Renault M, Berge-Thierry C, Gueguen P, Baumont D, Perrault M (2015) Strategy for the selection of input ground motion for inelastic structural response analysis based on naïve Bayesian classifier. Bull Earthq Eng 13:2517–2546. https://doi.org/10.1007/s10518-015-9728-z

Lesueur C, Cara M, Scotti O, Schlupp A, Sira C (2013) Linking ground motion measurements and macroseismic observations in France: a case study based on accelerometric and macroseismic databases. J Seismol 17:313–333. https://doi.org/10.1007/s10950-012-9319-2

Locati M, Camassi R, Rovida A, Ercolani E, Bernardini F, Castelli V, Caracciolo CH, Tertulliani A, Rossi A, Azzaro R, D'Amico S, Conte S, Rocchetti E (2016) DBMI15, the 2015 version of the Italian Macroseismic Database. Istituto Nazionale di Geofisica e Vulcanologia. https://doi.org/10.6092/INGV.IT-DBMI15

Luzi L, Hailemikael S, Bindi D, Pacor F, Mele F, Sabetta F (2008) Itaca (Italian accelerometric archive): a web portal for the dissemination of Italian strong-motion data. Seismol Res Lett 79(5):716–722. https://doi.org/10.1785/gssrl.79.5.716

Margottini C, Molin D, Serva L (1992) Intensity versus ground motion: a new approach using Italian data. Eng Geol 33(1):45–58. https://doi.org/10.1016/0013-7952(92)90034-V

Moratto L, Costa G, Suhadolc P (2009) Real-time generation of Shake Maps in the Southeastern Alps. Bull Seism Soc Am 99(4):2489–2501. https://doi.org/10.1785/0120080283

Pacor F, Paolucci R, Luzi L, Sabetta F, Spinelli A, Gorini A, Nicoletti M, Marcucci S, Filippi L, Dolce M (2011) Overview of the Italian strong motion database ITACA 1.0. Bull Earthq Eng 9(6):1723–1739. https://doi.org/10.1007/s10518-011-9327-6

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12(85):2825–2830

Schreiber J (2018) Pomegranate: fast and flexible probabilistic modeling in python. J Mach Learn Res 18(164):1–6

Tertulliani A, Azzaro R (eds) (2016) QUEST - Rilievo macrosismico in EMS98 per il terremoto di Amatrice del 24 agosto 2016 (Tech. rep.). INGV. http://www.questingv.it/index.php/rilievi-macrosismici/29-amatrice-24-08-2016-ml-6-0-rapporto-sul-rilievo-macrosismico-ems98/file

Tiberi L, Costa G, Jamšek Rupnik P, Cecić I, Suhadolc P (2018) The 1895 Ljubljana earthquake: can the intensity data points discriminate which one of the nearby faults was the causative one? J Seismol 22:927–941. https://doi.org/10.1007/s10950-018-9743-z

Wald DJ, Quitoriano V, Heaton TH, Kanamori H, Scrivner CW, Worden CB (1999) TriNet 'ShakeMaps': rapid generation of peak ground motion and intensity maps for earthquakes in southern California. Earthq Spectra 15(3):537–555. https://doi.org/10.1193/1.1586057

Zhang H (2004) The optimality of naive bayes. In: Barr V, Markov Z (eds) Proceedings of the seventeenth international florida artificial intelligence research society conference. AAAI Press, Menlo Park, California, pp 562–567.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.