Geophysical Journal International

Geophys. J. Int. (2010) **180,** 1138–1152 doi: 10.1111/j.1365-246X.2009.04467.x

Regression analysis of MCS intensity and ground motion parameters in Italy and its application in ShakeMap

Licia Faenza and Alberto Michelini

Istituto Nazionale di Geofisica e Vulcanologia, Via di Vigna Murata, 605, *Rome* 00143, *Italy. E-mails: licia.faenza@ingv.it; alberto.michelini@ingv.it*

Accepted 2009 November 27. Received 2009 November 26; in original form 2009 April 1

SUMMARY

In Italy, the Mercalli–Cancani–Sieberg (MCS) is the intensity scale in use to describe the level of earthquake ground shaking, and its subsequent effects on communities and on the built environment. This scale differs to some extent from the Mercalli Modified scale in use in other countries and adopted as standard within the USGS-ShakeMap procedure to predict intensities from observed instrumental data. We have assembled a new PGM/MCS-intensity data set from the Italian database of macroseismic information, DBMI04, and the Italian accelerometric database, ITACA. We have determined new regression relations between intensities and PGM parameters (acceleration and velocity). Since both PGM parameters and intensities suffer of consistent uncertainties we have used the orthogonal distance regression technique. The new relations are

$$I_{\rm MCS} = 1.68 \pm 0.22 + 2.58 \pm 0.14 \log PGA, \sigma = 0.35$$

and

$$I_{\rm MCS} = 5.11 \pm 0.07 + 2.35 \pm 0.09 \log PGV, \sigma = 0.26.$$

Tests designed to assess the robustness of the estimated coefficients have shown that single-line parametrizations for the regression are sufficient to model the data within the model uncertainties. The relations have been inserted in the Italian implementation of the USGS-ShakeMap to determine intensity maps from instrumental data and to determine PGM maps from the sole intensity values. Comparisons carried out for earthquakes where both kinds of data are available have shown the general effectiveness of the relations.

Key words: Earthquake ground motions; Seismicity and tectonics.

1 INTRODUCTION

The use of intensity scales is historically important because no instrumentation is necessary, and useful measurements on the level of shaking can be made by an unequipped observer (e.g. Musson 2002). To some extent, the mid-years of the 20th century saw a decline in interest of macroseismic investigations, since large improvements were made in instrumental monitoring. However, since the mid-1970s there has been a resurgence in the subject since macroseismic data are essential for revision of historical seismicity and are of great importance in seismic hazard assessments. It follows that macroseismic studies of modern earthquakes are still crucial for (i) assessing the size of historical earthquakes; (ii) studying local ground-motion attenuation and (iii) investigations of vulnerability, seismic hazard and seismic risk.

Since the late 1990s, the software package ShakeMap (Wald *et al.* 1999b) which seeks to estimate rapidly (few minutes) the level of ground shaking resulting from an earthquake has been proposed and implemented in several parts of the world (e.g. USA, Canada, Iceland, Italy and at local scales, for the city of Seattle). ShakeMap is a seismologically based interpolation algorithm that combines observed data and seismological knowledge to produce maps of peak ground motion (PGM). The shaking is represented through maps of peak ground acceleration (PGA), peak-ground velocity (PGV), response spectral acceleration (SA), and ground-motion shaking intensity. The 'instrumental intensity' values are derived from the conversion of PGM into intensity values (e.g. Wald *et al.* 1999a). These maps have become adopted worldwide to provide quantitative, first order assessments of the level of shaking and of the extent of potential earthquake damage. In particular, intensities have been found informative by non-expert audiences unfamiliar with instrumental ground motion parameters. The intensity values are derived from the ground-motion recorded values, using a correlation relationship. For the USGS-ShakeMap standard distribution this calibration has been performed using California earthquakes ground-motion data and the Mercalli Modified (MM) intensity scale (e.g. Wald *et al.* 1999a).

In Italy, the software ShakeMap has been operational at the 'Istituto Nazionale di Geofisica e Vulcanologia' since 2006 (Michelini *et al.* 2008) and the intensity maps of peak ground

motion shaking adopt the California relationship of Wald *et al.* (1999a). In Italy, however, the analysis of historical seismicity through the use of the macroseismic studies has a long tradition. The Mercalli–Cancani–Sieberg (MCS) Scale (Sieberg 1930), is the scale adopted in Italy. MCS combines an earlier ten-degree scale proposed by Mercalli (1902), the evolution of this scale with additional two-degree introduced for dealing with very strong earthquakes by Cancani (1904) and the successive remodulation by Sieberg (1912). To this regard, Musson *et al.* (2009) provide a thorough assessment of the various scales and of their evolution through time.

There are two main reasons that have lead us to re-calibrate the conversion scale between peak ground motion and the reported MCS intensity data. The first follows from the fact that the MM instrumental intensity adopted within the implementation of ShakeMap (Michelini *et al.* 2008) can be misleading as the MCS representation is customary in Italy. Consequently, differences between the two scales can cause confusion. The second follows from the large number of macroseismic data available for past events in Italy (i.e. Stucchi *et al.* 2007). These data have been also used to generate scenarios for seismic hazard analysis.

The aim of this work is to develop a new correlation relationship between recorded peak ground motions and reported MCS intensities for Italy. The derived MCS instrumental intensity relation is intended to be introduced for the calculation of shakemaps in Italy. To this regard, the intensity maps are the most viewed output from non-specialist audience when consulting, for example, the INGV ShakeMap portal (Michelini *et al.* 2008). For this reason, it is important to maintain the same intensity scale between the shakemaps and the other products that represent intensities throughout the Italian territory (i.e. the Italian database of macroseismic information, DBMI (Stucchi *et al.* 2007), and the 'Did You Feel It' maps (http://terremoto.rm.ingv.it/). In addition, access to a relation that allows conversion between MCS intensities and PGM's allows for the calculation of PGM's ground estimates for historical events which can be of high relevance when attempting to reconstruct past ground motion scenarios.

2 CORRELATION BETWEEN INTENSITIES AND PGMs

The problem of the correlation between the reported intensity and the ground motion parameters has been debated at length in the literature. Although it is largely accepted that there is a 'relation' between intensity and the logarithm of the peak ground motions, either in PGA, or in PGV (e.g. Cancani 1904; Gutenberg & Richter 1942; Kawasumi 1951; Hershberger 1956; Ambraseys 1975; Margottini *et al.* 1992; Wald *et al.* 1999a; Faccioli & Cauzzi 2006; Gómez Capera *et al.* 2007, and see references therein), it has not yet been proposed a physical relation capable to represent it, and the empirical regressions proposed are mainly statistical. We also note that, being the intensity scale based on observations and not on instrumental values, there is no guarantee that a logarithmic relation is effectively applicable. This has been long recognized by several authors (e.g. Hershberger 1956; Ambraseys 1975) who recommended much caution in using these relations. Among all the works available in literature, it seems that the principal differences consist in the selection of the data base. Recently, a good overview of this topic at the global scale, and for Italy in particular, has been prepared by Gómez Capera *et al.* (2007).

In general, the relations are obtained at regional scales, with the exception of the studies by Ambraseys (1975) who proposes a single regression for Europe and the Middle East, and Decanini *et al.* (1995) who adopt a unique regression for Italy, West USA and South America. This implies that each work relies on its own regional data base.

Apart from some exceptions [Theodulidis & Papazachos (1992), that include soil classification for the Greek territory; Atkinson & Kaka (2007) and Tselentis & Danciu (2008) that include magnitude, epicentral distance and soil classification for Greece, and Souriau (2006) that includes only the epicentral distance], all the regressions adopt the same functional form—a linear regression between intensity and the logarithm of the peak ground motion. The foremost difference stays instead in the processing of the data. In general, some works (mainly those of the U.S. researchers) use the geometric mean value of the recorded ground motion for each intensity class (e.g. Hershberger 1956; Trifunac & Brady 1975; Murphy & O'Brien 1977; Wald *et al.* 1999a) while others, mainly Italians (e.g. Chiaruttini & Siro 1981; Margottini *et al.* 1992; Faccioli & Cauzzi 2006; Gómez Capera *et al.* 2007) have chosen not to group the peak values for each intensity value. We note that by using data grouped into intensity classes obviates the problems of the large scatter of the peak ground motion data for each intensity unit—for each intensity unit a single value of peak ground motion is determined (usually through the geometric mean and in the Appendix, we address the role that different data pre-processing have on the results). Furthermore, and with the notable exception of Gómez Capera *et al.* (2007), all adopted regressions neglect the errors of the independent variable, and this may be at the origin of some bias in the resulting regressions. Lastly, a factor that makes difficult the comparison between the different regression, and the determination of a general regression formula, follows from the use of different macroseismic scales throughout the world (i.e. the MM for USA, the MKS and MCS for Europe, and the JMA for Japan).

Our analysis starts by considering the studies performed on Italian data by Margottini *et al.* (1992), Faccioli & Cauzzi (2006) and Gómez Capera *et al.* (2007). Margottini *et al.* (1992) obtained first an empirical correlation between PGA and intensity for $I > 5$. The remaining two works used and modified the data base compiled earlier by Margottini *et al.* (1992). Faccioli & Cauzzi (2006) developed a relation for intensity versus PGA and PGV using least-squares fitting. Gómez Capera *et al.* (2007) used only PGA data and adopted the orthogonal distance regression technique, ODR, (Fuller 1986; Boggs *et al.* 1988).

3 DATA

Intensity can be defined as a classification of the strength of shaking at any place during an earthquake, in terms of its observed effects on buildings and human beings. The fact that it is essentially a classification, rather than a physical parameter, leads to some special conditions on its use. Principal among these is its being a discrete scale, and therefore caution is needed to correlate continuous (i.e. ground motion) and a discrete (i.e. intensity) scales.

Margottini *et al.* (1992) are the first to provide a data base that relates peak ground motions and MCS intensities for the entire Italian territory. [In fact, the earlier study by Chiaruttini & Siro (1981) focussed only to earthquakes primarily in NE Italy and it is not representative of the whole Italian territory]. In Margottini *et al.* (1992), the intensities were directly assigned by the authors after gathering the data of the strongest instrumental Italian earthquakes

since 1980. The intensities were divided into 'local' and 'general'. While the former (i.e. local) refers to the damage of the buildings located few hundreds of metres from the accelerograph station, the latter classifications (i.e. general) are associated to the damage of the town or village closest to the station. A total of 56 data points from nine earthquakes constituted the final Margottini *et al.* (1992) data base. A revision and integration of this data base was performed by Faccioli & Cauzzi (2006) who considered only the points with 'general' intensity, and integrated it with other non-Italian earthquakes, for a total of 26 earthquakes and 75 data points. Although the criterion adopted to associate instrumental and intensity data was not specified by Faccioli & Cauzzi (2006) (i.e. distance between the stations and the intensity points), this data base is the most recent and complete currently available for intensities larger than $I = \mathrm{IV}-\mathrm{V}$ in Italy.

Recently, the results of the project ITACA—the Italian accelerometric database—have been made available (Luzi *et al.* 2008). ITACA contains 2182 three component waveforms generated by 1004 earthquakes with a maximum magnitude of 6.9 (1980 Irpinia earthquake) covering the period range from 1972 to 2004. The project aims to collect, homogenize and distribute the data acquired over the time period 1972–2004 in Italy by different Italian institutions, namely 'Ente Nazionale per l'Energia Elettrica' (ENEL, Italian electricity company), 'Ente per le Nuove tecnologie, l'Energia e l'Ambiente' (ENEA, Italian energy and environment organization) and the 'Dipartimento della Protezione Civile' (DPC, Italian Civil Protection) (see http://itaca.mi.ingv.it for additional detail).

As previously noted, in Italy, there is a large and homogeneous macroseismic intensity data base—the DBMI data base (Stucchi *et al.* 2007)—available at http://emidius.mi.ingv.it/DBMI04/, with a revised release 1900–2008 (i.e. DBMI08). This database is a revised collection of all the macroseismic analysis done for the Italian peninsula. It includes a total of almost 60 000 observations from 12 000 earthquakes at more than 14 000 localities. Although it is well known that local conditions can affect the amplitude (and duration) of the wave field, we have made no attempt to subdivide further the pair association according to the different recording sites since the intensity values reported in DBMI04 represent already average values. The reported intensities follow the MCS scale in classes spaced by 0.5 intensity units (e.g. 4, 4.5, 5, ...).

The possibility to access and cross-match these two sources of data gave us the opportunity to assemble a new, homogeneous database consisting of intensity and peak ground motion values (see Table S1). To this purpose, we have extracted all the localities reporting intensity data which are located within 3 km from the accelerograph stations that recorded the data. This was performed for all the events within ITACA.

Fig. 1 shows the spatial distribution of the selected events and the location of the stations. 66 earthquakes in the time span 1972–2004 ($3.9 \leq M_{\mathrm{w}} \leq 6.9$) and intensity MCS ≤ 8 have been analysed, for a total of 266 pairs Intensity-PGM (see Table S1). Fig. 2 plots the distribution of the data versus the distance from the epicentre to the station. Overall, the database is well distributed although we note that there are few intensity data at closer distances for small intensity values (i.e. in the range $2 \leq \mathrm{MCS} \leq 3.5$). This follows from the DBMI08 data being compiled for damaging events (i.e. medium-large magnitude earthquakes producing macroseismic damage). Perhaps more importantly, the assembled data set does not provide intensity-PGM pairs at intensity levels larger than 8. Unfortunately, this is an inherent limitation of the assembled data set and to some extent it prevents to constrain tightly the largest intensity values in terms of observed PGM values.

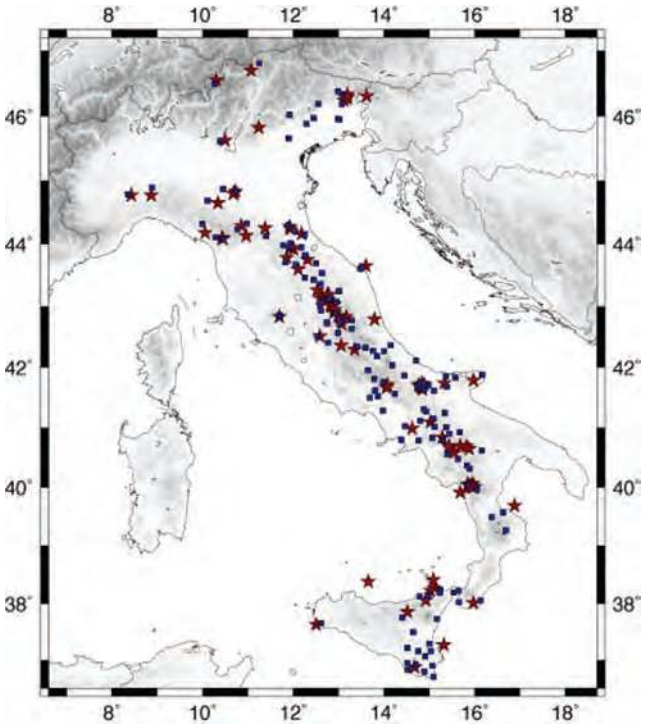

Figure 1. Map showing the location of the analysed events (red stars) and of the stations (blue solid squares) used to assemble the intensity–PGM pair data set.

As mentioned in Section 2, there are two distinct procedures to use the data in the regression. The first consists of binning the data (BID hereafter) into classes at 0.5 intensity intervals and calculating for each class the PGM mean and its standard deviation. The second procedure does not involve any averaging and adopts the whole data set although some robust statistics can be applied (e.g. remove the tails of the data distribution) to remove the influence of the outliers. In the following we adopt the geometric mean approach (see also the Appendix). The geometric mean, μ_g, is calculated as

$$\mu_g = \frac{1}{n}\sum_{1}^{n} \log PGM_i, \tag{1}$$

where n is the number of data points for each intensity class.

The use of the geometric mean is motivated by the PGM data distribution about the arithmetic and logarithmic means as shown in Fig. 3. The expected normal distribution curves are also shown for reference purposes and it is evident that the deviations from the arithmetic mean are not approximated by a normal distribution. For both PGA and PGV the distributions about the arithmetic means are skewed to the lower side of the mean value where the great majority of the residuals fall. In contrast, the distributions computed using the logarithmic mean agree well with the theoretical normal distribution curve. To test the likelihood of the normal distribution we have performed the 1-sample Kolmogorov–Smirnov test. We can reject the null-hypothesis of a normal distribution for the PGA and PGV with an α-value less than 1 per cent. Conversely, we cannot reject the null-hypothesis for log PGA and log PGV with an α-value equal to 95 per cent and 45 per cent, respectively. This all indicates that the data appear to be nearly log-normally distributed

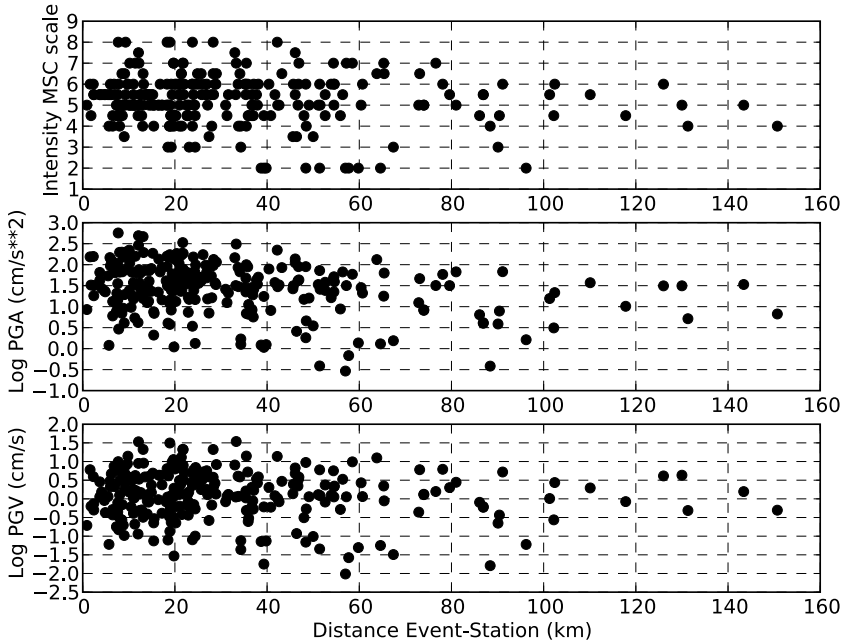

Figure 2. Distance coverage of the assembled PGM–MCS intensity pairs data set.Top panel: MCS Intensity; middle panel: log PGA and bottom panel: log PGV. The distance is calculated using the epicentral location of the events. Adoption of this distance for large events, rather than the fault distance, will introduce some differences in the diagrams but it is inconsequential to the analysis carried out here.

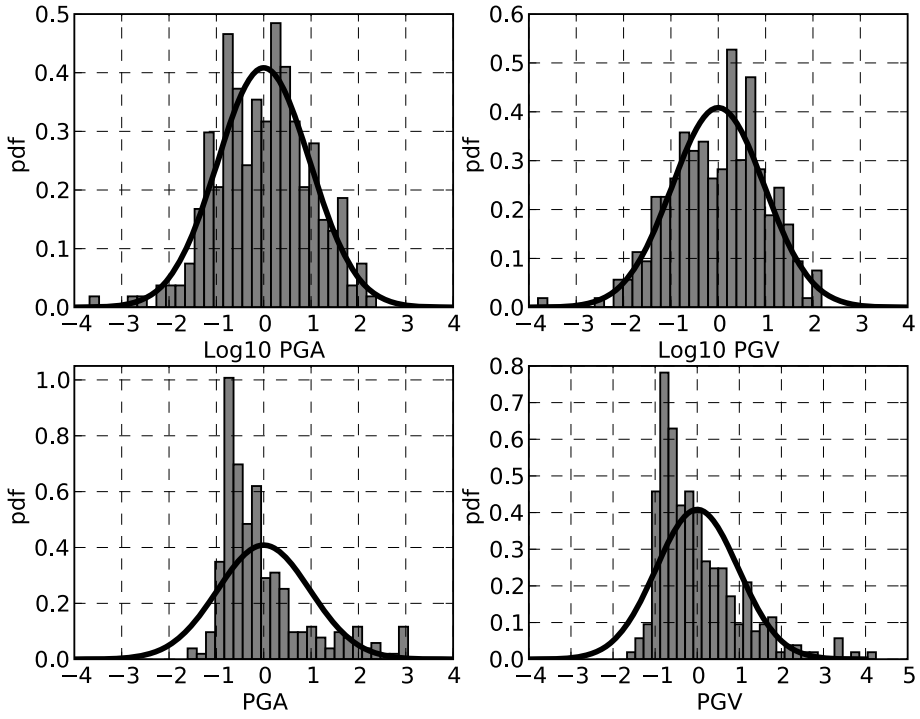

Figure 3. PGM data distribution. Original data (bottom panel) and after application of the logarithm in base-10 (top panel). (PGA: left-hand panel and PGV: right-hand panel). For each intensity bin, the data set is normalized to obtain standardized values, having zero mean and unit standard deviation. To the purpose of reference, the expected normal distribution curves are also shown as thick solid lines.

and will be treated as such in the following analyses. Our results are very similar to those presented by Murphy & O'Brien (1977).

For what concerns the standard deviation associated to the measurements, a value of 0.5 for the intensity seems a conservative but reasonable value. For the ground motion data, we use, for each class, the sample geometrical standard deviation, σ_g, defined as

$$\sigma_g = \exp\left[\sqrt{\frac{\sum_1^n (\log PGM_i - \mu_g)^2}{n}}\right]. \quad (2)$$

In summary, 12 pairs of intensity and PGM data are used to fit using BID. The PGM values are calculated using the geometric mean average. The intensity standard deviations have been set equal to the conservative value of $\sigma_I = 0.5$ while the corresponding PGM value is determined from the geometric standard deviation (see eq. 2).

4 METHOD: ORTHOGONAL DISTANCE REGRESSION

The ordinary least-squares (OLS) fitting is the most commonly applied criteria for fitting data to models and for estimating parameters of the models. The mathematical and statistical validity of this method is based on the stringent, important constraint that the independent variable must be known to a much greater accuracy then the dependent variable. It follows that this regression can never be inverted, that is, the regression of y against x can not be inverted to derive the regression of x against y.

The orthogonal distance regression (Fuller 1986; Boggs *et al.* 1988; Castellaro & Bormann 2007; Gómez Capera *et al.* 2007) is a more appropriate technique in problems where dependent and independent variables are both affected by uncertainty. ODR extends least square data fitting to problems with independent variables that are not know exactly (Boggs *et al.* 1988) and it can be used for fitting linear and non-linear models. The data fitting problems arise by considering a data set (x_i, y_i; $i = 1, \ldots, n$) and a model that is purported to explain the relationship of $y_i \in \mathcal{R}^1$ versus $x_i \in \mathcal{R}^m$. Assuming errors in both variables, with ϵ_i the error for the dependent variable y_i and δ_i that for the independent variable x_i, the functional to be satisfied is

$$y_i = f(x_i + \delta_i; \beta) - \epsilon_i, \tag{3}$$

where $\beta \in \mathcal{R}^p$ is the parameters vector, $f()$ is a smooth function that can be either linear or non-linear in x_i and β.

While OLS resolves the parameters vector, β^{OLS}, for which the sum of the squares of the n vertical distances from the curve $f(x_i; \beta)$ to the n data points is minimal, ODR minimizes the weighted orthogonal distances from the curve. Thus, the parameter vector, β^{ODR}, is found by minimizing the following problem

$$min_{\beta,\delta} \sum_{i=1}^{n} \left[\left(f(x_i + \delta_i; \beta) - y_i \right)^2 + \delta_i^T D_i^2 \delta_i \right], \tag{4}$$

where $D_i \in \mathcal{R}^{m \times m} (i = 1, \ldots, n)$ is a set of positive diagonal matrices that allow ϵ_i and δ_i to have different variance (Boggs *et al.* 1987b, 1988). Problem (4) is non-linear even if $f(x_i;\beta)$ is linear in both x and β, that is, the model is a straight line. When eq. (3) is satisfied, and $\epsilon, \delta_1, \ldots, \delta_n$ are independent and normally distributed, then eq. (4) results in the maximum likelihood estimator of β (Britt & Luecke 1973; Boggs *et al.* 1988). In the simplest use of ODR, it is assumed that each $D_i = dI$ where d is the ratio of the standard deviation of the errors in the y and x data, that is, $d = \sigma_\epsilon / \sigma_\delta$. In this work, we used the algorithm developed by Boggs *et al.* (1987a)—a FORTRAN code wrapped within the SciPy Python module (http://www.scipy.org).

5 APPLICATION

We fit the data using a linear relation between the intensity (I) and the logarithm in base 10 of the peak-ground motion, PGM (i.e. PGA or PGV)

$$I = a + b \log PGM. \tag{5}$$

Use of the ODR technique allows also for the direct inversion between PGM and I so that the calculated coefficients can be used to express PGM as function of I. This is a nice property of ODR since it allows, using the same coefficients, for prompt conversion between the sought variables.

5.1 PGA

We fit the data using ODR using both a single- and a double-line parametrization. With the single-line regression, we have obtained $a = 1.68 \pm 0.22$ and $b = 2.58 \pm 0.14$, with a standard deviation of the regression line of $\sigma_{\mathrm{singleline}} = 0.35$.

The data, however, seem to show some different scaling between low and high intensity values and, as in Wald *et al.* (1999a) (see also Atkinson & Kaka 2007), the data set is subdivided into two parts—intensities less than 5 and intensities greater or equal to 5. The resulting coefficients from application of ODR using the double-line regression are $a_{I \geq 5.0} = -0.21 \pm 1.12$, $b_{I \geq 5.0} = 3.54 \pm 0.57$ (7 data out of 12 belong to this group), and for the data with intensity less than 5, the parameters are $a_{I<5.0} = 2.02 \pm 0.09$, $b_{I<5.0} = 2.02 \pm 0.06$. The standard deviation of the double-line fitting is $\sigma_{\mathrm{doubleline}} = 0.28$ (see Fig. 4a).

The decrease of the value of the standard deviation with the double-line regression when compared to that of the single-line may suggest it more appropriate a regression with two lines. However the standard deviations associated to our estimates for the $I \geq 5.0$ coefficients are quite large to indicate the indeterminacy that arises when attempting to fit with a double-line the available data set. This aspect will be analysed more thoroughly below using synthetic tests.

5.2 PGV

The procedure described for PGA has been also applied to PGV. The parameter for the single-line regression using our binned data set are $a = 5.11 \pm 0.07$ and $b = 2.35 \pm 0.09$, with a standard deviation of the model as $\sigma_{\mathrm{singleline}} = 0.26$.

The value of the coefficients of the double-line regression are $a_{I \geq 5.0} = 4.68 \pm 0.22$, $b_{I \geq 5.0} = 2.93 \pm 0.30$, and for the data with intensity <5.0, the parameters are $a_{I<5.0} = 4.79 \pm 0.01$, $b_{I<5.0} = 1.94 \pm 0.10$. The standard deviation of the model is $\sigma_{\mathrm{doubleline}} = 0.26$. The comparable values of the standard deviation between single- and double-line ODR fitting and the relatively small values of the uncertainties of the coefficients would suggest the former to be adequate to fit the intensity-PGV data set (Fig. 4b).

5.3 Appraisal of the results

The results shown for PGA and PGV in the previous sections do leave some ambiguities on which of the regression results should be chosen.

First, we have verified whether our results depend on the values of the standard deviation assigned to the PGM. To this regard, we have repeated the analysis using the standard deviation of the mean [i.e. $\sigma_g / \sqrt{n}$, in eq. 2] as uncertainty and found results in agreement to those shown in Sections 5.1 and 5.2.

Secondly, studies similar to those presented here but carried out on different data sets (e.g. Wald *et al.* 1999a; Atkinson & Kaka 2007) evidence an apparent change in slope at intensity 5 whereas our data set does not seem to replicate clearly the same behavior (see Fig. 4). The reason for this could be, however, attributed to the differences of the MCS scale when compared to the MM (and other

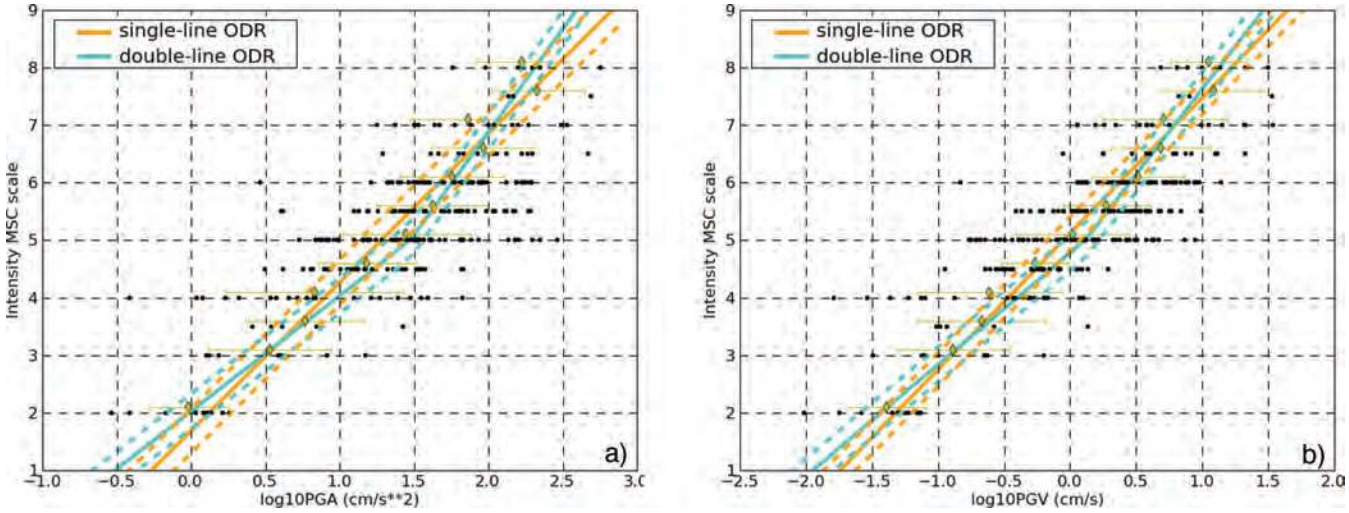

Figure 4. MCS Intensity versus PGM for BID (i.e. the PGM geometric mean binned data set). Data (black solid circles), data geometric mean (yellow diamonds) and standard deviations (yellow error bars), single-line ODR (solid orange line) and double-line ODR (solid cyan line). The associated 1σ standard deviations (dashed lines) are shown on each regression line. The diamonds and the error bars are slightly shifted for plotting purposes. Left-hand panel is for PGA, right-hand panel for PGV.

scales) in the range of intensities between 5 and 7 (e.g. Margottini *et al.* 1992) or, more simply, to lack of resolving power of the data set employed. To test this latter hypothesis, we have used synthetic data sets generated to replicate the statistical features (and range of values) of the observed data set. We restrict the analysis to PGA although analogous conclusions can be drawn from PGV.

In practice, we have generated two log-normal distributed, data sets consisting of PGA-Intensity pairs for a single- and a double-line data distribution. The data sets consist of 500 PGA-Intensity pairs for each intensity level class. A conservative value of 0.5 has been assigned to the standard deviation of the intensity values. The classes range between 1 and 10 at 0.5 interclass intervals. The true values of the coefficients for the single-line in eq. (5) are $a = 1.82$ and $b = 2.40$. The double-line data set was generated using the following values for the coefficients in eq. (5) $a_{I\geq 5.0} = 0.214$, $b_{I\geq 5.0} = 3.54$ and $a_{I<5.0} = 2.024$, $b_{I<5.0} = 2.023$. These values are all comparable to those of the observed data. We refer to these data sets including all the values (i.e. 18 class values times 500 PGA points each) as the 'whole' data set.

We first test the accuracy of the coefficient estimates using the 'whole' data set; we have generated 1000 synthetic data sets using the coefficient values above and the purpose is to investigate the robustness of the parameters estimates using the BID data processing. The results show the coefficient estimates to be accurate and tightly distributed (see Fig. 5 as example for the single-line estimates of the intercept and slope coefficients). In practice the results obtained with the 'whole' data set indicate that with a large data set featuring the same statistical properties of our data set it would be possible to estimate accurately the parameter vector in both single- and double-line parametrization.

Our goal is, however, to verify the robustness of the estimates obtained with the observed data. Therefore, we repeat the analysis using different sampling of the single- and double-line synthetic data sets. Each sampled subset matches, in terms of number of data points drawn for each intensity level, that of the observed data set. We refer to these (re)sampled data sets (i.e. 266 PGA-intensity data points each) as the 'sampled' data set.

In Fig. 6, we present an example, for one of the data sets, of the BID set regressions for both the single- and the double-line data sets. The BID set has been determined for both the 'whole' data set and for one of the 'sampled' data sets drawn from the selected 'whole' synthetic data set. As anticipated, we see that the determined regression lines for the 'whole' data set match very closely those used to construct the synthetic PGA-Intensity pairs. In contrast, this is not the case when fitting the data for one of the 'sampled' cases (see Fig. 6).

To test the accuracy of our estimates obtained with the observed data, we need to construct enough replications of the 'sampled' synthetic data set to then compute some adequate statistics. To this end the sampling was repeated for 1000 times on one of the 'whole' data sets above. To provide a better perception of the uncertainties associated to the coefficient estimates, we present the results of the investigation using cumulative distributions.

In Fig. 7, we show the distribution of the single- and double-line slope regression coefficient of the 'sampled', BID processed, data set. We note that the true values do match closely the median value of the cumulative distribution. However, there is a remarkable difference for the distribution of the b value of eq. (5) for the single- and the double-line fits. The single-line distribution is very tight around the median value (i.e. the 80 per cent of the sampled outcomes lies in the range 2.38–2.42) whereas the slope coefficients of the double-line regressions display a much larger scatter. This is particularly apparent for the $I > 5$ line which relies on a very small number of data points at the higher intensity values (Fig. 7b) and the 80 per cent of the estimates falls in the broad range 2.2–5.0, approximately.

Our final step has been to investigate the distribution of the single- and double-line model standard deviations. We want to assess the significance of the relatively small value of the standard deviation found when fitting the intensity values using the double-line parametrization to the observed data (i.e. $\sigma_{\text{doubleline}} = 0.28$) when compared to that of the single-line (i.e. $\sigma_{\text{singleline}} = 0.35$). An uncritical examination of these data may in fact lead to the conclusion that the observed smaller values of the double-line regression are significant. To this purpose, we have determined the mean average of the synthetic standard deviations from the 'sampled' data sets. The results are summarized in Table 1. We see first that the mean standard deviation from the 'sampled' single-line synthetic data sets,

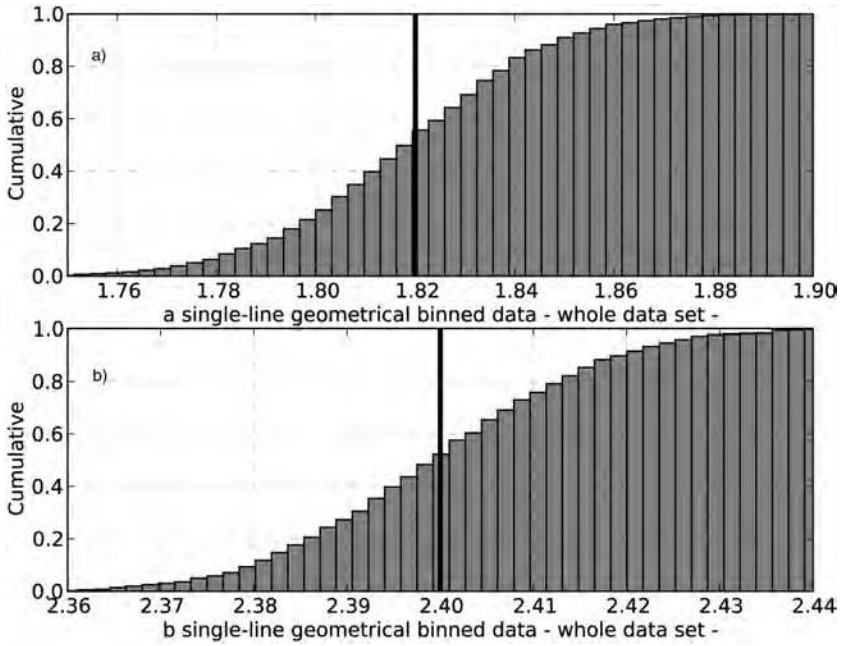

Figure 5. Synthetic test addressing the robustness of the whole data set. Cumulative distribution of the ODR regression on 1000 whole data sets for the single-line case. The true values of 1.82 and 2.40 for the *a* and *b* coefficients are shown as vertical, thick solid lines. The figure shows the results for PGA but similar results are found also for PGV.

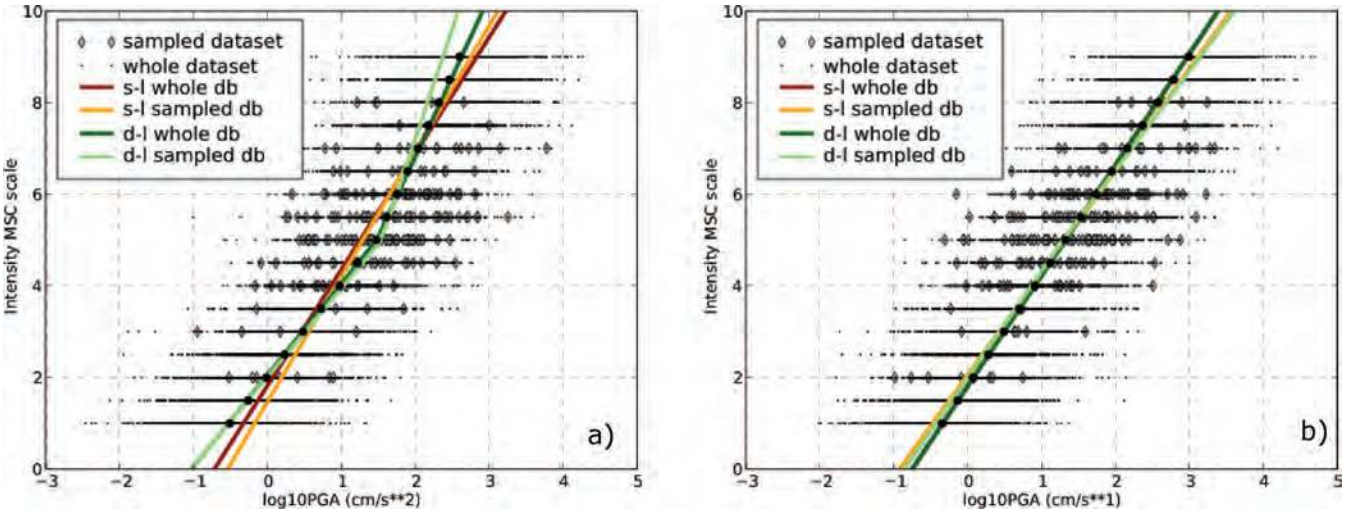

Figure 6. Regression results for the synthetic data. The 'whole' data set includes 500 PGA points at each intensity level. Double-line synthetic data set (a) and single-line synthetic data set (b). The 'sampled' data (grey solid diamonds) indicate an instance of synthetic data sampling replicating that of the observed data (i.e. 266 PGM-MCS pairs). The solid black circles indicate the true values determined from the adopted regression lines. See legend for detail on the symbols and line coding used.

fitted through a single-line, does not differ from that obtained from the 'sample' double-line synthetic data fitted also through a single-line (see second column in Table 1). The values obtained from the synthetic tests are very similar to those found from the observed data. Similarly, the mean standard deviations obtained from fitting, through a double-line, the single- and the double-line 'sampled' synthetic data sets also display very similar values (≈0.4; see third column in Table 1). In this latter case, however, the values obtained from the synthetic analysis differ to some extent from the observed value although the latter still lies within the $\pm\sigma$.

In conclusion, we do not feel of significance that the observed standard deviation is lower when using the double-line parametrization (see Fig. 7b) and our data set does not allow to discriminate between single- and double-line regression parametrization. Since this all follows also from the limited resolving power of the data set used, it seems that inclusion of additional degrees of freedom in the regression (e.g. epicentral distance or magnitude) would most likely increase the indeterminacy of the analysis.

5.4 Discussion

In Fig. 8(a) we summarize the results for PGA obtained in Section 5.1 using the regression

$$I_{\mathrm{MCS}} = 1.68 + 2.58 \log PGA \qquad (6)$$

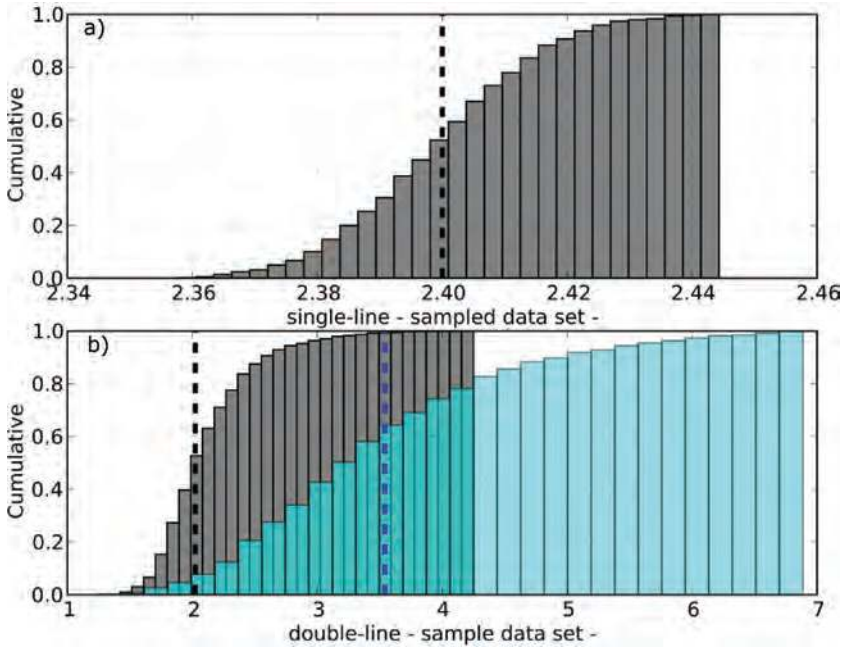

Figure 7. Synthetic tests. (a) cumulative distributions showing the distribution of the single-line regression slope coefficient. The thick dashed line marks the true value. (b) same as the top panel but for the double-line regression. In grey, the cumulative distribution of the slope coefficients of the data set for $I < 5.0$; the black thick dashed line marks the true value; in light blue the same kind of distribution but for the data set with $I \geq 5.0$; the dark blue thick dashed line marks the true value.

Table 1. ODR standard deviation values.

Data type observed	Single-line proc. $\sigma_{\rm sl} = 0.35$	Double-line proc. $\sigma_{\rm dl} = 0.28$
Synthetic: single line	$\bar{\sigma}_{\rm sl} = 0.38 \pm 0.13$	$\bar{\sigma}_{\rm dl} = 0.39 \pm 0.16$
Synthetic: double line	$\bar{\sigma}_{\rm sl} = 0.36 \pm 0.16$	$\bar{\sigma}_{\rm dl} = 0.41 \pm 0.19$

($\sigma_a = 0.22$ and $\sigma_b = 0.14$) together with the regressions obtained by Margottini *et al.* (1992), Faccioli & Cauzzi (2006) and Gómez Capera *et al.* (2007) for Italy, and the regression of Wald *et al.* (1999a) currently in use in the generation of maps of shaking in Italy (Michelini *et al.* 2008). The uncertainties expressed as $\pm\sigma$ bounds associated to each regression are also shown. Similarly, in Fig. 8(b) we also present the results for PGV and we compare the results of the determined single-line regression (with the $\pm\sigma$ bound) with those of Faccioli & Cauzzi (2006) and Wald *et al.* (1999a).

In general and in the range of values $I_{\rm PGA} \geq 5$, we find that our regression line features a slope coefficient intermediate between that found by Faccioli & Cauzzi (2006) (i.e. smaller value) and those obtained by Margottini *et al.* (1992) and Gómez Capera *et al.* (2007) (i.e. larger values). Specifically, at intensities between 5 and 6, the regression line determined in this study matches closely the results of Faccioli & Cauzzi (2006) and at intensities between 7 and 8, our regression predicts PGA values (and viceversa) consistent to those

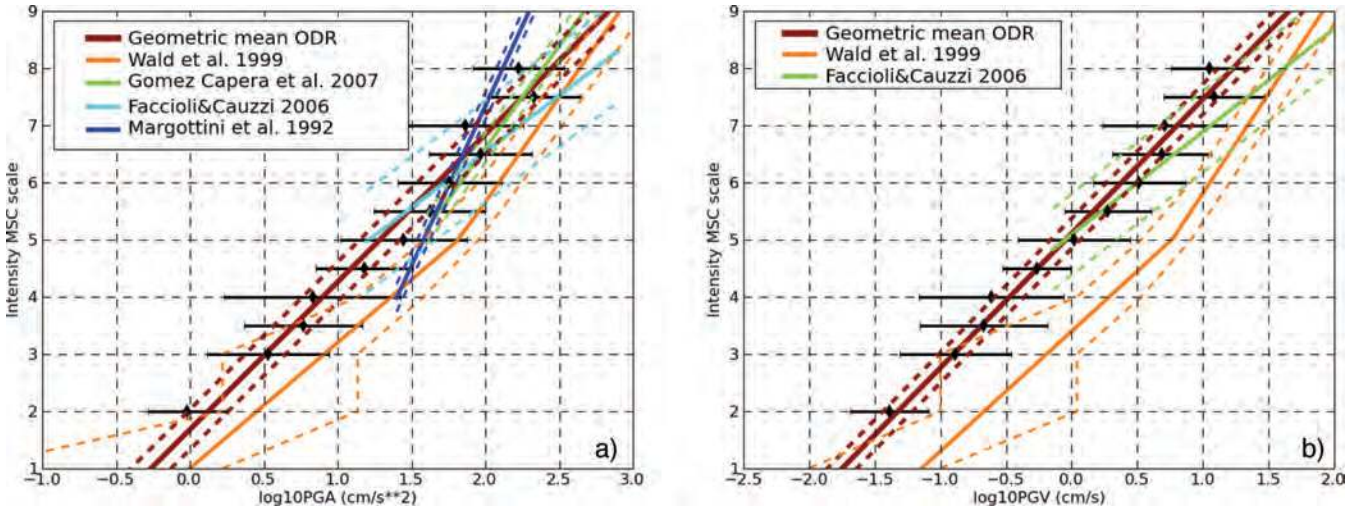

Figure 8. Comparison between the intensity versus PGM regressions obtained using the BID set and the ODR technique. PGA and PGV in (a) and (b) panels, respectively. For comparison, the recently published regressions of Gómez Capera *et al.* (2007), Faccioli & Cauzzi (2006), Margottini *et al.* (1992) and Wald *et al.* (1999a) are also shown (see legend).

of Gómez Capera *et al.* (2007). For $I_{PGA} \geq 7.5$, our regression predicts *PGA* values intermediate between those of Faccioli & Cauzzi (2006), that seem to overestimate the values of PGA at larger intensities, and those of Margottini *et al.* (1992) that, conversely, seem to overestimate the level of intensity at relatively smaller PGAs. The observed differences can be accounted by the different data sets used, the range of intensity values, the criteria adopted to pair the intensity values with the recorded ground motion and by the regression technique adopted. For example, Faccioli & Cauzzi (2006) do not describe the criteria used and do not comment on the different and inhomogeneous intensity scales grouped together—MCS and MM—in their data set. Furthermore, in this study we adopt a different regression technique, which takes into account explicitly the uncertainties in both dependent and independent variables, and that we do bin the data whereas Faccioli & Cauzzi (2006) do not.

The considerations made for PGA are in part applicable to PGV. As for the single-line PGA fit, we find that the regression

$$I_{MCS} = 5.11 + 2.35 \log PGV \quad (7)$$

($\sigma_a = 0.07$ and $\sigma_b = 0.09$) displays a slope coefficient larger than that of Faccioli & Cauzzi (2006). Thus, while both our regression and that of Faccioli & Cauzzi (2006) feature a very similar I_{PGV}–PGV pair values at $I_{PGV} \approx 5$, they do differ progressively at increasing intensities (or PGV). This results in almost one intensity unity difference at $PGV \approx 10^{1.5}$ cm s^{-1}, that is, $I_{PGV} \approx 8$ and $I_{PGV} \approx 9$, for Faccioli & Cauzzi (2006) and this study, respectively. When our regression is compared to that proposed by Wald *et al.* (1999a) for the MM scale, we find that the two regressions differ between one and two intensity units up to $PGV \approx 10^{1.5}$ cm s^{-1}. The maximum difference occurs at $PGV = 10^{0.75}$ cm s^{-1}. These differences can originate significant differences in the values of the instrumentally derived intensities when compared to those obtained, for example, from the 'Did You Feel It' questionnaire (e.g. 'hai sentito il terremoto', http://www.haisentitoilterremoto.it/) or from more thorough macroseismic post-earthquake investigations.

Finally, in the range of values $I_{PGA<5}$ and $I_{PGV<5}$ it is not possible to compare the obtained regressions because both Faccioli & Cauzzi (2006) and Gómez Capera *et al.* (2007) confined their analysis to intensities larger than 5.

6 APPLICATION TO SHAKEMAP

One of the main goals that motivated this study was the determination of a reliable, instrumentally derived, MCS intensity scale which can be adopted in the USGS-ShakeMap procedure (Wald *et al.* 1999b) for the Italian territory (Michelini *et al.* 2008) to provide rapid MCS intensity maps following $M > 3$ earthquakes. In addition, correct calibration of the intensity conversion gives the opportunity to generate maps of PGM parameters (PGA and PGV) exploiting the very large intensity database for past earthquakes available in Italy (Stucchi *et al.* 2007). This reverse approach is important when attempts are made to provide first-order estimates of the ground shaking of historical earthquakes without relying on sophisticated and costly waveform modeling techniques, or the creation of earthquake scenarios that use just peak ground motion attenuation relations without any constraint provided by observed data.

In defining the conversion we have followed Wald *et al.* (1999b); we first compute the instrumental intensity adopting the PGA regression and if the instrumental intensity is larger than six, we adopt the instrumentally derived intensity from PGV. This choice follows from the observation that near-source strong ground-motions are often dominated by short-duration, pulse-like peaks and therefore PGV appears to be a more robust measure of intensity for strong shaking (Wald *et al.* 1999b, 2006).

To show the validity of the regressions determined in this study, we have applied eqs (6) and (7) to the data of all the earthquakes with at least 4 instrumental records used in this study. For each earthquake, the shakemaps that adopt the observed PGM data are compared to those obtained after conversion from I_{MCS} to PGM. In the S2 online supplement (see Supporting Information section), we provide all the shakemaps expressed both in terms of MCS intensity and of PGA and PGV for the 25 earthquakes selected. In the following, we show two significant examples (M4.6 and M6.4 in Molise and Friuli, respectively) drawn form the calculated shakemaps that are explicative of the results of our study. These two earthquakes have been chosen to show application to earthquakes representative of the seismicity occurring in Italy. In fact, about ten M4+ earthquakes occur annually and are widely felt although they generally induce only much awareness without causing damage; M6+ earthquakes take place only a few per century but result in extensive damage and large number of fatalities.

In Fig. 9 top panels, we show the intensity shakemap for the M4.6, 2002 November 12, Molise event. We see a remarkable similarity between the strong motion data and the intensity derived maps of MCS intensity. The only notable difference between the two maps lies in the level of local resolution that depends on the number of observations. The standard shakemap that relies on PGM data alone has been determined using many fewer data (yellow triangles in the left panel of Fig. 9) and this results in a much smoothed local shaking distribution when compared to that obtained using the much larger number of intensity data (yellow triangles in the right-hand panel of Fig. 9). In Fig. 9 (middle and lower panels), we compare the PGM data shakemaps with those obtained after converting the MCS intensities into PGM using the relations of this study. Again, we note a remarkable similarity in the PGA and PGV shakemaps obtained directly from the data and from the intensity to PGM conversion. This result corroborates that the regressions found in this study can be adopted to provide first order, maps of peak ground motion although in these examples the level of local resolution is hampered by the paucity of observations when using the PGM data in the standard ShakeMap manner.

In Fig. 10, we show the results obtained for the May, 6, 1976 Friuli main shock. This earthquake caused very extensive damage and nearly one thousand fatalities. The PGM and intensity derived shakemaps (Fig. 10 – top panel) are similar although there seems to be some slight overestimation of intensities with the PGM data derived intensity; in terms of PGA the two maps are remarkably similar whereas in terms of PGV the instrumental, data derived shakemap has PGV values somewhat larger than that inferred using the relationships of this study. Nevertheless we feel that, to first order, the PGV shakemap obtained from the MCS intensities does provide, within the limitations imposed by a relationship calibrated using earthquakes throughout all Italy, a rather faithful representation of the level of shaking experienced in the area. These conclusions are confirmed by the maps shown in S2, which shows an overall agreement between the Intensity, PGA, and PGV maps based either on instrumental records or on macroseismic data.

Finally and in order to summarize concisely the differences between the shakemaps determined using recorded data and those derived from the macroseismic surveys using the relations found here, we have calculated the per cent differences for all the shakemaps shown in S2 and in Figs 9 and 10. The points used to determine the differences include the phantom grid points of USGS-ShakeMap

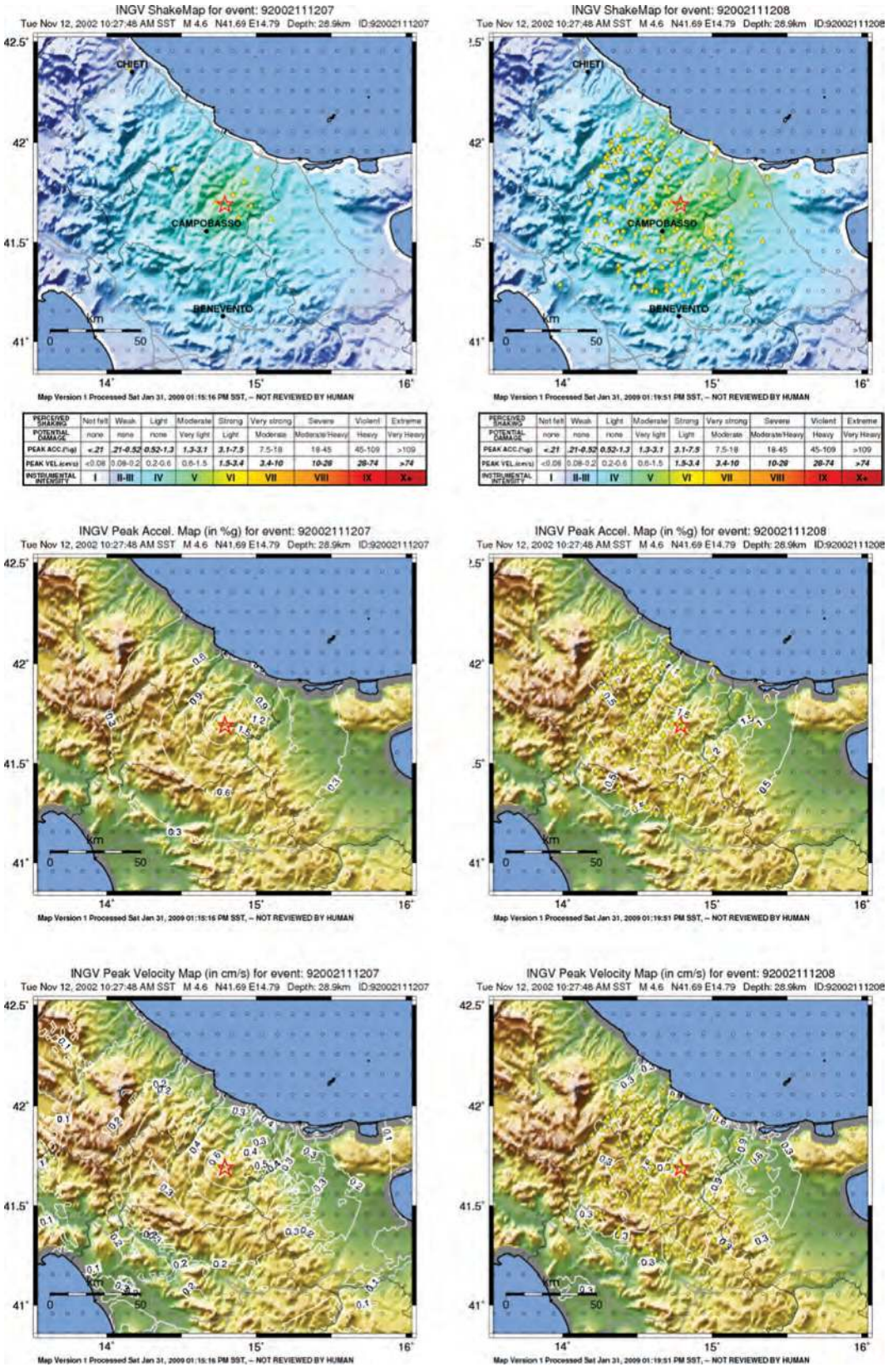

Figure 9. Shakemap of the $M_L = 4.6$, 2002 November 12, earthquake in the Molise area in Southern Italy. We have applied the site conditions derived from the geological VS30 and the regionalized Italian ground motions equation (see Michelini *et al.* 2008), with standard deviation $\sigma_{PGA} = 1.698$ and $\sigma_{PGV} = 1.940$. Top panel: shakemaps expressed in terms of MCS Intensity; Middle panel: shakemaps expressed in terms of PGA (in per cent g); Bottom panel: shakemaps expressed in terms of PGV (in cm s^{-1}). PGM and MCS intensity derived shakemaps are shown in the left- and right-hand columns, respectively. The yellow triangles are the stations (left-hand panels) and intensity site (right-hand panels) used as input in the analysis.

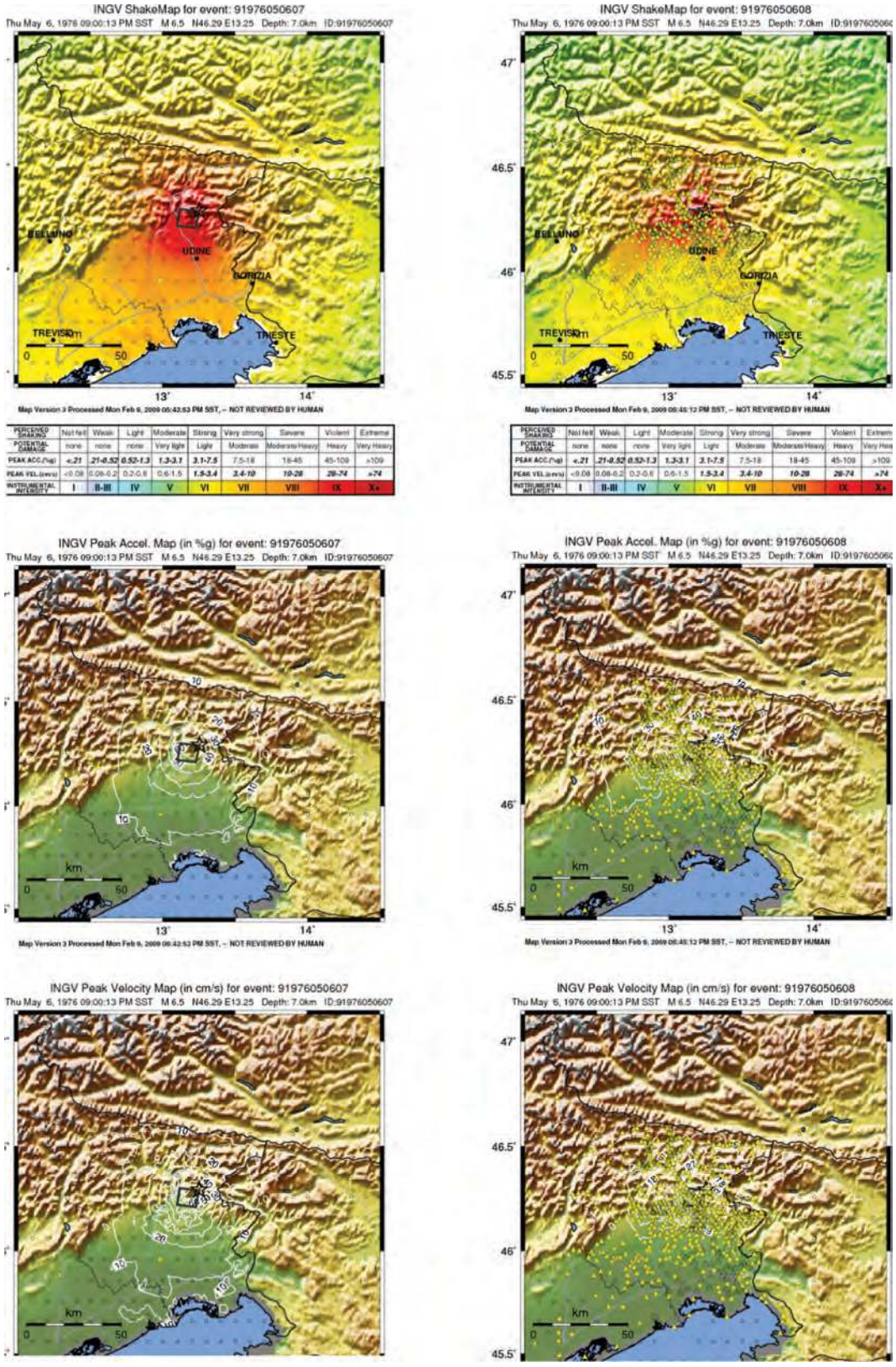

Figure 10. Shakemap of the $M_L = 6.4$, 1976 May 6, Friuli main shock in Northern Italy. We have applied the site conditions derived from the geological VS30 and the Akkar and Bommer PGM relations, see Michelini *et al.* (2008), with standard deviation $\sigma_{PGA} = 1.779$ and $\sigma_{PGV} = 1.862$. (Same format as Fig. 9).

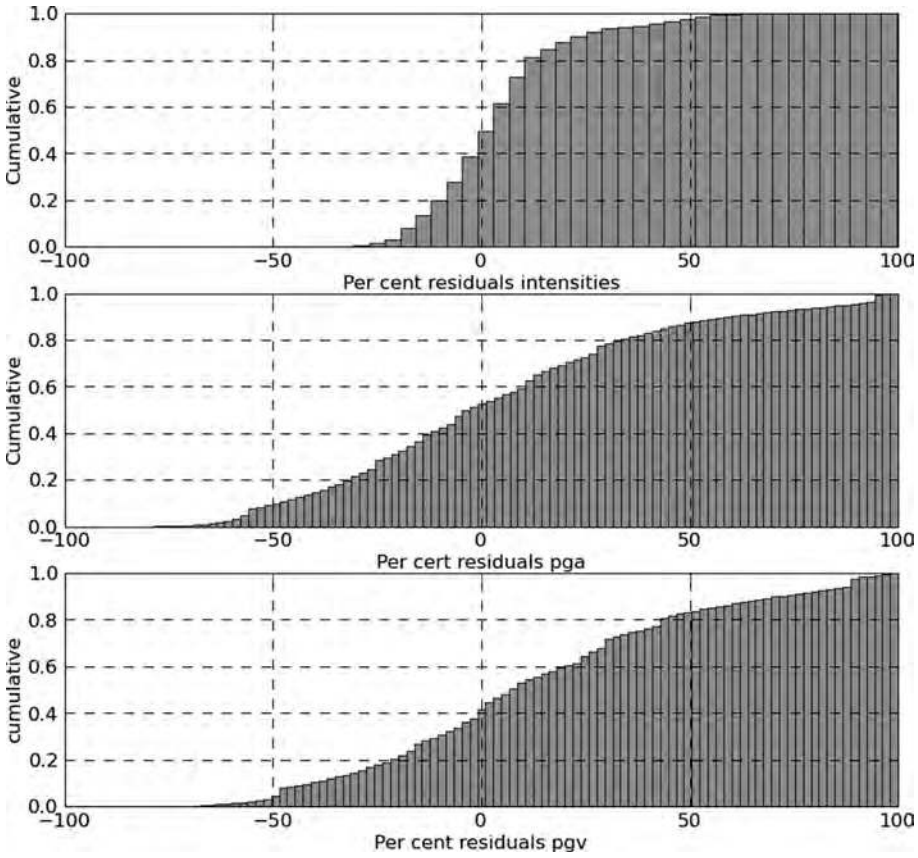

Figure 11. Comparison between MCS intensity, PGA and PGV values determined from instrumental (inst) data and from macroseismic (macro) surveys (e.g. $(PGA_{\rm inst} - PGA_{\rm macro})/PGA_{\rm macro} \times 100$) using cumulative distributions. More than 250 000 data points are used in each graph. Mean, standard deviation and median values are (4.63, 16.54, 2.97) for intensity, (3.37, 40.21, −2.49) for PGA, and (12.59, 39.17, 8.39) for PGV.

(e.g. Wald *et al.* 1999b; Michelini *et al.* 2008) within a radius of 140 km from the epicentre for a total of more than 250 000 geographical points. Fig. 11 shows the residuals for intensity, PGA and PGV. The three cumulative distributions show that the per cent differences for all three parameters are centred around zero. In particular, we find that 90 per cent of the intensity values are comprised within ±30 per cent. For PGA and PGV, we find that 80 and 70 per cent of the values, respectively, lie within ±50 per cent differences.

Finally, we have verified whether a correlation of the residuals with distance and magnitude occurs in our analysis. To this end, we have determined 2-D histogram of the residual distribution as function of magnitude and epicentral distance for all the data available. The results shown in Fig. 12 do not seem to support the existence of such dependencies although we cannot exclude them given the scatter of the data used in the analysis.

7 CONCLUSIONS

In this study, we have performed regression analysis between MCS intensities and instrumentally recorded peak ground motion data expressed in terms of PGA and PGV. The data set has been assembled for earthquakes that have occurred in Italy in the time period 1972–2004. The work has been driven by the need to represent intensities using the MCS scale within the implementation of ShakeMap for the Italian territory. This should insure improved interconsistency between the rapid shakemaps obtained from application of the USGS-ShakeMap procedure (Wald *et al.* 1999b; Michelini *et al.* 2008) using observed PGM data, and the characterizations of ground motion shaking that rely on either 'Did You Feel It' analysis (http://www.haisentitoilterremoto.it/) and/or macroseismic data in general (Stucchi *et al.* 2007).

Because both the intensity and the PGM data are affected by inherent uncertainties, we have adopted the ODR technique which explicitly takes into account the uncertainties in dependent and independent variables. In order to apply the technique, we have chosen to bin the data using the geometric mean. This is motivated by the PGM data conforming to a log-normal distribution.

The data set used in the analysis has been assembled from two thoroughly verified data sources—the database of the Italian strong motion recordings, ITACA (Luzi *et al.* 2008) and the Macroseismic Database of Italy 2004 (Stucchi *et al.* 2007). Compilation of the data set resulted in 266 PGM-$I_{\rm MCS}$ data pairs, which are two to three times larger than those analysed in previous similar studies for Italy.

The results show that with the data available a single-line regression is sufficient to fit the data without introducing two regression lines, that is, for low and high intensities (or PGM), respectively. Adoption of the single- rather than the double-line parametrization has been explored thoroughly using synthetic tests for data distributions replicating the observed data.

Finally, we have tested the determined relations by inserting them in the USGS-ShakeMap procedure currently in use at INGV (Michelini *et al.* 2008) to find (i) the instrumentally derived MCS intensity maps do match closely the reported macroseismic data and maps and (ii) the regression relations can be used to predict PGM maps which we have found to be generally consistent with those

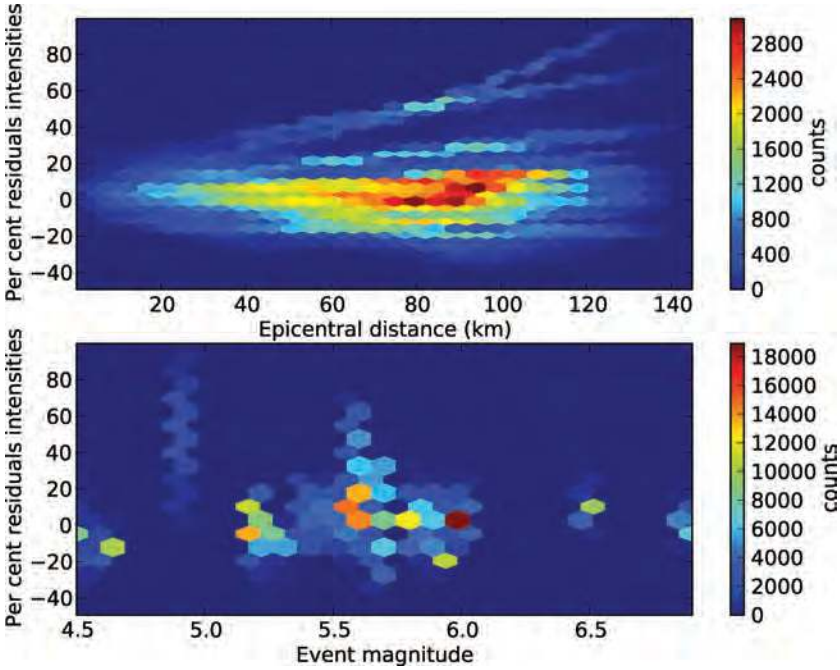

Figure 12. Analysis of the dependencies of the per cent residual intensities (cf Fig. 11) versus epicentral distance (top panel) and magnitude (bottom panel). The panels show that no significant trend of the residuals against magnitude and distance occur for our data set.

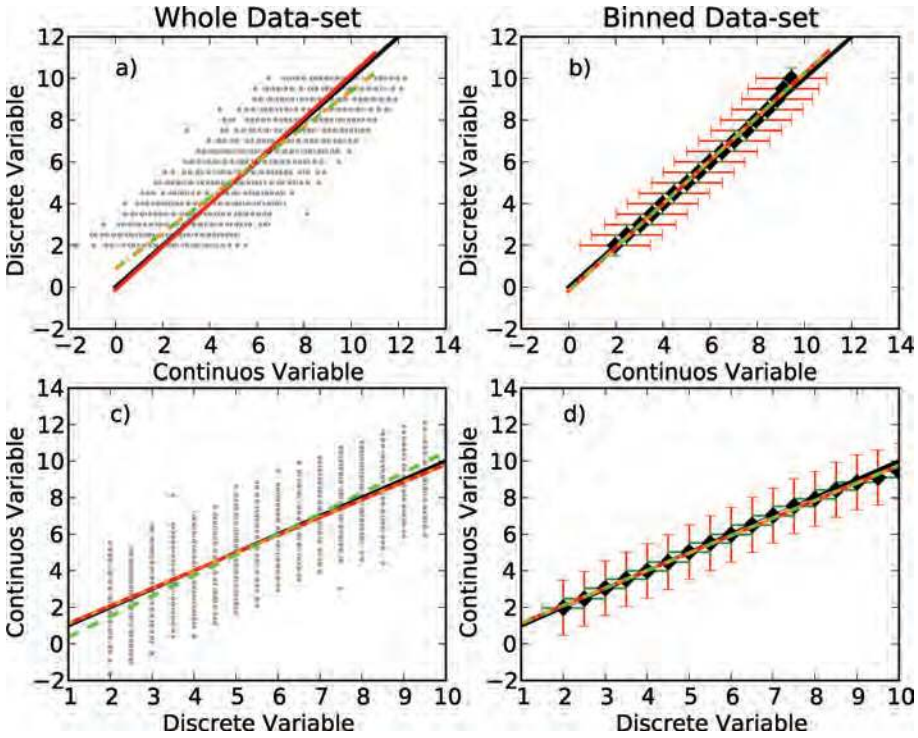

Figure 13. Regression tests: small black dots are the synthetic whole data set; black diamonds are the mean values for each class of the discrete variables; black solid line is the true regression—the bisector; orange dashed line is the least-squares regression without errors in the variables; green dash–dotted line is the ODR regression with much smaller errors in the continuous variables than in the discrete one (ODR_{case1}); red solid line is the ODR line with much smaller errors in the discrete variables then in the continuous one (ODR_{case2}). Left-hand panels (*a,c*) show the results of the analysis when no data binning is applied (i.e. whole data set). Conversely, data binning is applied on the right-hand side panels (b and d). In the context of the work (a) replicates the case of the regression of $I = f(PGM)$ without binning; (b) replicates the case of the regression of $I = f(PGM)$ using the binned data set; (c) replicates the case of the regression of $PGM = f(I)$ without binning; (d) replicates the case of the regression of $PGM = f(I)$ using the binned data set. The green error bar (b and d) is relative to ODR_{case1} while the red one to ODR_{case2}. The small differences observed between the regression lines and the bisector used to generate the synthetic data set are to be attributed to the manner the data of the discrete variable have been constrained to the hard upper- and lower-bounds.

from observed instrumental data. The residuals analysis made on the shakemaps shown in this work appear to prove the consistency of our regression equations, both for intensity versus PGM, and PGM versus intensity. In addition, we have verified that the found regressions do not depend on either magnitude or distance.

Overall, we find that the results obtained from application of the regressions determined in this study do provide an improved representation of the level of ground shaking in terms of the adopted MCS intensity scale in Italy or, alternatively, the regressions can be used to predict realistic ground motions from intensity data alone.

ACKNOWLEDGMENTS

This research has been funded by the SAFER project (EC contract n. 036935) and by the Italian Civil Protection 2007-2009 DPC S3 contract. The authors are grateful to A. Amato and M. Stucchi for the careful reading of a preliminary version of the manuscript. We would like to thank Yehuda Ben-Zion, Trevor Allen and John Douglas for reviewing the manuscript and providing constructive criticism. We are also particularly grateful to R. Camassi and P. Gasperini for their useful advices and suggestions for the analysis of the macroseismic data.

REFERENCES

Ambraseys, N., 1975. The correlation of intensity with ground motion, in *Proc. 14th Conf. Europ. Seism. Comm.*, Trieste, Bull. Europ. Comm. Earthq. Eng., Vol. 1, pp. 335–341.

Atkinson, G.M. & Kaka, S.I., 2007. Relationships between felt intensity and instrumental ground motion in the Central United States and California, *Bull. seism. Soc. Am.,* **97**(4), 1350–1354.

Boggs, P.T., Byrd, R.H., D.J.R. & Schnabel, R., 1987a. Odrpack - software for weighted orthogonal distance regression, Tech. Rep. CU-CS-360-87, University of Colorado, Department of Computer Sciences, Boulder, CO.

Boggs, P.T., Byrd, R.H. & Schnabel, R.B., 1987b. A stable and efficient algorithm for nonlinear orthogonal distance regression, *SIAM J. Scient. Stat. Comput.,* **8,** 1052–1078.

Boggs, P.T., Spiegelman, C.H., Donaldson, J.R. & Schnabel, R.B., 1988. A computational examination of orthogonal distance regression, *J. Econometr.,* **38,** 169–201.

Britt, H.L. & Luecke, R.H., 1973. The estimation of parameters in nonlinear implicit models, *Technometrics,* **15,** 233–247.

Cancani, C., 1904. Sur l'emploi d'une double echelle seismique des intesites, empirique et absolue, *Gerlands Beitrage Geophysik,* **2,** 281–283.

Castellaro, S. & Bormann, P., 2007. Performance of different regression procedures on the magnitude conversion problem, *Bull. seism. Soc. Am.,* **97**(4), 1167–1175.

Chiaruttini, C. & Siro, L., 1981. The correlation of peak ground horizontal acceleration with magnitude, distance, and seismic intensity for Friuli and Ancona, Italy, and the Alpide belt, *Bull. seism. Soc. Am.,* **71**(6), 1993–2009.

Decanini, L., Gavarini, C. & Mollaioli, F., 1995. Proposta di definizione delle relazioni tra intensità macrosismica e parametri del moto del suolo, in *Atti del 7 Convegno Nazionale Ingegneria sismica in Italia,* Vol. 1, pp. 63–72.

Dowdy, S. & Wearden, S., 1991. *Statistics for Research,* 2nd edn, John Wiley & Sons, New York, Chichester, Brisbane, Toronto, Singapore.

Faccioli, E. & Cauzzi, C., 2006. Macroseismic intensities for seismic scenarios, estimated from instrumentally based correlations, in *Abstract Book 1st ECEES*, http://www.ecees.org/abstracts_book.pdf, p. 125.

Fuller, W.A., 1986. *Measurement Error Models,* John Wiley & Sons, Inc., New York, NY, USA.

Gómez Capera, A.A., Albarello, D. & Gasperini, P., 2007. Aggiornamento relazioni fra l'intensità macrosismica e PGA, Tech. rep., Convenzione INGV-DPC 2004-2006.

Gutenberg, B. & Richter, C.F., 1942. Earthquake magnitude, intensity, energy, and acceleration, *Bull. seism. Soc. Am.,* **32**(3), 163–191.

Hershberger, J., 1956. A comparison of earthquake accelerations with intensity ratings, *Bull. seism. Soc. Am.,* **46**(4), 317–320.

Jorgensen, B., 1997. *The Theory of Linear Models,* Chapman & Hall, New York, NY.

Kawasumi, H., 1951. Measures of earthquake danger and expectancy of maximum intensity throughout Japan as inferred from the seismic activity in historical times, *Bull. Earthq. Res. Inst.,* **1951.10.5**, 469–482.

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, *Seism. Res. Lett.,* **79**(5), 716–722.

Margottini, C., Molin, D. & Serva, L., 1992. Intensity versus ground motion: a new approach using Italian data, *Eng. Geol.,* **33**(1), 45–58.

Mercalli, G., 1902. Intensity scales, *Bollettino della Societ Sismologica Italiana,* **8,** 184–191.

Michelini, A., Faenza, L., Lauciani, V. & Malagnini, L., 2008. Shakemap implementation in Italy, *Seism. Res. Lett.,* **79**(5), 689–698.

Murphy, J.R. & O'Brien, L.J., 1977. The correlation of peak ground acceleration amplitude with seismic intensity and other physical parameters, *Bull. seism. Soc. Am.,* **67**(3), 877–915.

Musson, R., 2002. Intensity and Intensity Scales, *IASPEI New manual of seismological observatory practice,* ch. 12, pp. 653–672, ed. Bormann, P., GeoForschungsZentrum Potsdam, Potsdam.

Musson, R.M.W., Grunthal, G. & Stucchi, M., 2009. The comparison of macroseismic intensity scales, *J. Seismol.,* doi:10.1007/s10950-009-9172-0.

Sieberg, A., 1912. Über die makroseismische Bestimmung der Erdbebenstrke, *Gerlands Beitrage Geophysik,* **11,** 227–239.

Sieberg, A., 1930. Geologie der Erdbeben, *Handbuch der Geophysik,* **2**(4), 552–555.

Souriau, A., 2006. Quantifying felt events: a joint analysis of intensities, accelerations and dominant frequencies, *J. Seismol.,* **10**(1), 23–38.

Stucchi, M., et al., 2007. DBMI04, il database delle osservazioni macrosismiche dei terremoti italiani utilizzate per la compilazione del catalogo parametrico CPTI04., *Quaderni di Geofisica,* **49,** http://portale.ingv.it/produzione-scientifica/quaderni-di-geofisica/archivio/resolveUid/0c549ba6165e5d96636aba24f3677c17.

Theodulidis, N. & Papazachos, B., 1992. Dependence of strong ground motion on magnitude-distance, site geology and macroseismic intensity for shallow earthquakes in Greece. I: peak horizontal acceleration, velocity and displacement, *Soil Dyn. Earthq. Eng.,* **11**(7), 387–402.

Trifunac, M.D. & Brady, A.G., 1975. On the correlation of seismic intensity scales with the peaks of recorded strong ground motion, *Bull. seism. Soc. Am.,* **65**(1), 139–162.

Tselentis, G.-A. & Danciu, L., 2008. Empirical relationships between modified mercalli intensity and engineering ground-motion parameters in Greece, *Bull. seism. Soc. Am.,* **98**(4), 1863–1875.

Wald, D.J., Quitoriano, V., Heaton, T.H. & Kanamori, H., 1999a. Relationships between peak ground acceleration, peak ground velocity, and modified mercalli intensity in California, *Earthq. Spectra,* **15**(3), 557–564.

Wald, D.J., Quitoriano, V., Heaton, T.H., Kanamori, H., Scrivner, C.W. & Worden, C.B., 1999b. Trinet 'shakemaps': rapid generation of peak ground motion and intensity maps for earthquakes in southern California, *Earthq. Spectra,* **15,** 537.

Wald, D.J., Worden, C.B., Quitoriano, V. & Pankow, K.L., 2006. Shakemap manual, technicalmanual, usersguide, and software guide, Tech. rep., U.S. Geological Survey.

Zar, J.H., 1999. *Biostatistical Analysis,* 4th edn, Prentice Hall, NJ.

APPENDIX

In this Appendix, we present some issues that should be taken into account when analysing data sets composed of data defined at discrete intervals. In the following, we cannot deal exhaustively the topic of regression strategies regarding continuous and discrete variables but we rather focus on some features that we have found of

great interest when performing the analysis object of this work. In particular, we have found of importance (i) the need for a biunique regression (i.e. correspondence between the two sets of data is one-to-one along both directions); (ii) the specific definition of the uncertainties for both variables and (iii) the data binning before processing the data.

We start by discussing the least-squares technique. In general, regression analyses are widely used in research since they are used to explain a given variable (the dependent variable, y) in terms of a combination (linear or not) of a given explanatorily variable (the independent variable, x). If y and x are inter-related, a model relationship can be used to predict the dependent variable given the independent one. Application of the least-squares, LS, method for a simple linear regression model, where 'simple' indicates here that there is only one independent variable, and 'linear' indicates that the model consists of a straight line, is based on four conditions (Dowdy & Wearden 1991).

(i) The x values have negligible errors.

(ii) For each x value there is a normal distribution of y values—this assumption is necessary for inference.

(iii) The distribution of y for each x has the same variance, that means that the variance around the trend line is the same irrespective to the value of x.

(iv) The expected values of y for each x lie on a straight line.

From the first point, it is obvious that this regression technique is not biunique, unless we suppose that our variables have both negligible errors, which is not the case. This constraint suggests the use of a different regression method—the ODR—which allows for the inclusion of errors in the variables along both axes making the analysis more realistic, and biunique.

Some statistics books (e.g. Dowdy & Wearden 1991), define the analysis where both variables are affected by uncertainties as correlation rather than regression models. The characteristics of a correlation model are:

(i) Both x and y contain sampling variability.

(ii) For each value of x there is a normal distribution of y, and for each value of y there is a normal distribution of x.

(iii) The x distributions have the same variance; the y distributions have the same variance.

(iv) The joint distribution of x and y is the bivariate normal distribution.

The ODR fits fully these requirements.

The last point we need to discuss regards the importance of data binning before carrying out the regression analysis. We note that in the literature it cannot be found a general agreement on a standard methodology to apply to a given data set before regression. For example, some authors discourage the binning since it causes loss of information (e.g. Zar 1999) whereas others encourage its use (e.g. Jorgensen 1997).

To test these different perspectives on the matter, we have performed a numerical experiment adopting an ideal synthetic data set featuring the same characteristics of our intensity-PGM data set (i.e. with one discrete variable and one continuous) but consisting of many more data points.

The data set belongs to a 2-D normal distribution, with mean values centred at the bisector and $\sigma = 1$ uncertainties for both variables (Fig. 13). For each value of the discrete variable 1000 pairs are generated, for a total of 18 000 pairs. For the discrete variable, hard bounds were set at the upper- and lower-most values of 2 and 10, respectively.

The test consists of applying three different regression models to the whole and the binned data sets. The regressions applied are the LS without uncertainties in both variables (orange dashed line in Fig. 13), the ODR technique with much smaller uncertainties for the continuous variable than in the discrete one (green dash–dotted line in Fig. 13; hereinafter ODR_{case1}), and, lastly, the ODR with much smaller uncertainties in the discrete variable than in the continuous one (red solid line in Fig. 13; hereinafter ODR_{case2}). As anticipated, the aim of this numerical experiment is to verity (i) the applicability of LS in our analysis; (ii) the role of the uncertainties linked to both variables in the ODR technique; (iii) the robustness and accuracy of the results depending on binning (or not-binning) the data set.

When the three methods of analysis are applied to the binned data set, we have found that all provide proper fits to the data regardless of the choice of the independent variable (Figs 13b and d). Whereas ODR_{case1} and ODR_{case2} regressions are biunique, indicating that the line in Fig. 13(b) is the inverse of Fig. 13(d), LS is not.

The results change when the whole data set (i.e. without binning) is used for the regressions. When the continuous variable is used as independent, the LS regression method introduces some bias on both slope and intercept (see the orange dashed line in Fig. 13a). Conversely, the fit does not show any bias when the discrete variable is used as the independent one (see the orange dashed line in Fig. 13c). This result is not surprising since the LS regression minimizes the vertical distance. As remarked earlier, the LS regression is not biunique.

Similarly, we have found that caution must be paid in the assignment of the uncertainty to the variables when the ODR is applied to the whole data set. Only the ODR_{case2} provides correct fits regardless of the choice of the independent variable (red lines in Figs 13a and c), whereas the ODR_{case1} analysis introduces some bias (green dash lines in Figs 13a and c).

In summary and recalling the constraints posed by our analysis (i.e. a biunique regression), and by our data set (i.e. a mix of continuous and discrete variables) the results of this test would indicate that the preferential procedure to be adopted consists of using the binned data set and the ODR regression approach.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Table S1. Data set used for determining the regressions. The supplement table is provided in both pdf and comma separated value csv formats.

Table S2. Shakemaps for the 25 events analysed. Site conditions have been derived from the geological VS30 and the regionalized Italian ground motions equation (see Michelini *et al.* 2008). For each event, the shakemaps are expressed in terms of MCS Intensity (top), PGA (middle) and PGV (bottom). PGM and MCS intensity data derived shakemaps are shown in the left and right columns, respectively. The yellow triangles represent the strong motion stations (left panels) and the intensity sites (right panels) used as input for the analysis.

Please note: Wiley-Blackwell are 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 article.