Cameletti M., V. De Rubeis, C. Ferrari, P. Sbarra and P. Tosi (2017), An ordered probit model for seismic intensity data, Stoch. Environ. Res. Risk Assess., 31, 1593-1602, doi: 10.1007/s00477-016-1260-4.

Accepted version. The final publication is available at https://link.springer.com/article/10.1007/s00477-016-1260-4

An ordered probit model for seismic intensity data

Michela Cameletti · Valerio De Rubeis · Clarissa Ferrari · Paola Sbarra · Patrizia Tosi

the date of receipt and acceptance should be inserted later

Abstract Seismic intensity, measured through the Mercalli, Cancani, Sieberg (MCS) scale, provides an assessment of ground shaking level deduced from building damages, any natural environment changes and from any observed effects or feelings. Generally, moving away from the earthquake epicentre, the effects are lower but intensities may vary in space, as there could be areas that amplify or reduce the shaking depending on the earthquake source geometry, geological features and local factors. Currently, the Istituto Nazionale di Geofisica e Vulcanologia (INGV) analyzes, for each seismic event, intensity data collected through the online macroseismic questionnaire available at web-page `www.haisentitoilterremoto.it`. Questionnaire responses are aggregated at the municipality level and analyzed to obtain an intensity defined on an ordinal categorical scale.

The main aim of this work is to model macroseismic attenuation and obtain an intensity prediction equation, which describes the decay of macroseismic intensity as a function of the magnitude and distance from the epicentre. To do this we employ an ordered probit model, assuming that the intensity response variable is related through the link probit function to some predictors. Differently from what it is commonly done in the macroseismic literature, this approach takes properly into account the qualitative and ordinal nature of the macroseismic intensity as defined on the MCS scale. Using Markov chain Monte Carlo (MCMC) methods, we estimate the posterior probability of the intensity at each site. Moreover, by comparing observed and estimated intensities we are able to detect anomalous areas in terms of seismic SHAKING. This kind of information can be useful for a better assessment of seismic risk and for promoting effective policies to reduce major damages.

Keywords Bayesian modeling, Earthquakes, Intensity prediction equation, Macroseismic attenuation, Ordered probit model

M. Cameletti (corresponding author)
Department of Management, Economics and Quantitative Methods, Università degli Studi di Bergamo, Bergamo, Italy,
E-mail: michela.cameletti@unibg.it
Tel. 035 2052519

V. De Rubeis · P. Sbarra · P. Tosi
Istituto Nazionale di Geofisica e Vulcanologia, Roma, Italy,
E-mail: valerio.derubeis@ingv.it,paola.sbarra@ingv.it, patrizia.tosi@ingv.it

C. Ferrari
Unit of Statistics, IRCCS Fatebenefratelli, Brescia, Italy,
E-mail: cferrari@fatebenefratelli.eu

1 Introduction

Italy is one of the most earthquake prone country of Europe; its seismic network is composed by hundreds of seismograph stations (ITalian ACcelerometric Archive, ITACA http://itaca.mi.ingv.it/) used to estimate magnitude values and other seismological parameters (to this regard see the Italian Seismological Instrumental and parametric Data-basE, http://iside.rm.ingv.it). Even if these empirical data are reliable, obtaining a detailed definition and description of shaking is still a challenge, basically due to the high variability of ground motion. In addition to these instrumental data, there are also macroseismic data which refer to earthquake intensities measured by the Mercalli-Cancani-Sieberg scale (MCS; Sieberg, 1930) or the European Macroseismic Scale (EMS; Grünthal, 1998). In particular, macroseismic data regard earthquake effects on buildings, structures and people, and can be considered as a proxy of ground

shaking deduced from building damages, from any natural environment changes and from any observed effects or feelings. These macroseismic data are usually provided by expert operators who collect information from direct observation in each village and evaluate the intensity through a critical analysis. These data are then collected in historical catalogues which in Italy date back to 461 B.C.. It is worth to note that both the MCS and EMS intensity scales are qualitative and ordinal with categories ranging from I to XII. This means that we can surely say that the effects occurred in a municipality with intensity VIII are stronger than those associated with intensity IV, but there is not a well defined relation between intensity degrees and magnitude. In other words, we can not affirm quantitatively how intensity VIII relates to intensity IV as no precise numerical function is available to define the difference between intensity categories.

Concurrently with historical data, since 2007 INGV has been collecting macroseismic data through a web-survey available at `www.haisentitoilterremoto.it` ("hai sentito il terremoto?", hereafter HSIT, literally "did you feel the quake?"). This tool allowed to gather more than 700000 questionnaires regarding earthquakes widespread all over the Italian territory and felt by population. Even if derived from information provided by non-experts, the accuracy of the HSIT macroseismic intensities was assessed in Sbarra *et al.* (2010), Tosi *et al.* (2015) and Mak *et al.* (2015), where an agreement with values coming from traditional surveys and other internet-based datasets was found. Moreover, differently from historical macroseismic catalogue, the HSIT database includes a large amount of low degree intensity data, generally disregarded by traditional macroseismic investigation and analysis (Pasolini *et al.*, 2008). These data refer to areas far from the epicentre of high magnitude earthquakes or to areas at a short distance from low magnitude earthquakes.

The main aim of this work is the definition of a new intensity prediction equation (IPE) for Italian earthquakes using the macroseismic data available through the HSIT survey. The IPE describes the decay of macroseismic intensity as a function of the magnitude and distance from the epicentre and it is paramount in the analysis and interpretation of both recent and historical macroseismic intensity data. Moreover, it can be useful in seismic vulnerability assessment for prevention of damages, since it allows to compare expected (estimated by IPE) and observed intensities for detecting areas at major or minor risk to experience damages. In literature many IPEs (also named attenuation models or laws) have been proposed (see for example Gómez Capera, 2006 and Mak *et al.*, 2015), where the intensity (or its difference with the epicentral intensity) is a function of some covariates as epicentral intensity, quake depth and magnitude, site type, epicentral/hypocentral distance, etc.. However, these IPEs are based on historical databases which suffer from lack of accuracy for long distance and lack of data of low magnitude earthquakes.

The models for intensity decay can be specified using a deterministic (Atkinson and Wald, 2007) or a probabilistic approach (Magri *et al.*, 1994; Pasolini *et al.*, 2008) and, in the latter case, a statistical distribution is assumed for the response variable or the error term. Regardless of the adopted approach, so far intensities have been commonly treated as realizations of a quantitative distribution (continuous or discrete). As a result, numerical scores are (improperly) assigned to ordered intensity categories and least squares method are used to estimate the IPE parameters. Ignoring the ordinality of the response can yield predicted values which are not consistent with the ordinal nature of the intensity scale. More appropriate methodologies, which take into account the categorical nature of data, are proposed by Rotondi *et al.* (2008) and Zonno *et al.* (2009), even if applied on a small subset of data from the historical catalogue. Recently, a similar approach was adopted in Rotondi *et al.* (2015) for the large Italian macroseismic database DBMI11. Finally, Azzaro *et al.* (2013) proposed an anisotropic probabilistic model for the macroseismic intensity attenuation in the Mt. Etna region.

The novel contribution of this work consists in defining a new intensity prediction equation which takes properly into account the qualitative and ordinal nature of the macroseismic intensity, by using a large amount of data provided by HSIT web-survey. To do this, we adopt an ordered probit model (Agresti, 2010) where the intensity response variable is related through the probit link function to some predictors, such as the distance from the epicentre and the earthquake magnitude. Through this method, we are able to estimate the macroseismic intensity at all the desired locations, thus obtaining a new reliable IPE. Finally, an evaluation of anomalous areas, in terms of seismic shaking, is provided through ad-hoc residual analysis, i.e. by deriving the probability distribution of the difference between observed and expected intensities.

The paper is structured as follows: in Section 2 we introduce the web-based macroseismic survey of `www.haisentitoilterremoto.it`. In particular we describe the macroseismic questionnaire and the kind of data which are collected through it. The ordered probit model and the Bayesian estimation procedure via MCMC are detailed in Section 3, while Section 4 presents the results of the application with HSIT data. Section 5 con-

cludes the paper by summarizing the main findings, and includes some avenues for future research.

2 Macroseismic data from `www.haisentitoilterremoto.it`

The online macroseismic questionnaire, which is compiled by volunteers after having felt an earthquake, is composed by questions regarding the effects on the population and buildings evaluated following the MCS and EMS macroseismic scale (see Tosi *et al.* (2015) for a complete description). The questions regard: *i)* personal information and geographic location at the time of the earthquake; *ii)* transient effects evaluated through personal reactions, movement and/or fall of objects, and activity of the observer during the earthquake (sleeping, walking, being still); *iii)* building damages. In addition to volunteers, there exists also a permanent and constantly increasing group of compilers (approximately 25000), who are alerted via e-mail immediately after the occurrence of an earthquake near their municipality. Visiting the HSIT web-page of the considered event, they provide the location at the moment of the occurrence and declare if they felt or not the earthquake; in the first case, the macroseismic questionnaire can be filled in.

Using the procedure described in Tosi *et al.* (2015), an automated procedure controls the reliability of questionnaires and discharges those which either contain contradictory answers or insufficient information. Then, an algorithm is applied to the valid questionnaires in order to assign an unique intensity value (located on the centroid) for each municipality. Macroseismic intensity maps (both for MCS and EMS scales) are produced in real-time from the processing of the questionnaires and immediately displayed on the HSIT web-site (see Figure 1 for an example). Through the survey, thus, it is possible to obtain a real-time and widespread evaluation of earthquake intensities thanks to the amount of available data which is extremely larger than the one provided by direct observation of expert operators.

Note that the intensities provided by the HSIT procedure are given as real numbers, as a result of the algorithm described in Tosi *et al.* 2015, and then are approximated to the nearest integer value in accordance to the MCS and EMS degrees between II and VIII. Moreover, it is known that intensity web-based data collected for earthquakes very close in time could be affected by compilation errors. We thus excluded all aftershocks of magnitude lower than 4.5 occurred within 8 hours from each widely felt mainshock (identified as an earthquake of magnitude greater than or equal to 4.5 having more than 300 reports). Finally, we discarded the firstly felt earthquake before the mainshock, because, in case of a strong event, respondents often fail to choose the right event from the automatic list that appears on the HSIT web-site.

3 The ordered probit model

For ordinal data several multinomial models are available in literature and a comprehensive presentation can be found in Agresti (2010). Among those, a predominant role is played by the class of *cumulative link* models which link cumulative probabilities to a linear predictor. The most commonly used link functions are the *logit* and *probit*, the second one being the inverse of the standard Normal cumulative distribution function (cdf). The probit link was the most natural solution for this work as our model includes Gaussian distributions. Moreover, as specified in Albert and Chib (1993) and Cowles (1996), this choice gives rise to some computational benefits from the inferential point of view (see Section 3.1).

For municipality $i = 1, \dots, I$ and earthquake $c = 1, \dots, C$ let y_{ic} be the felt intensity estimated through the HSIT web-survey. The response y_{ic} is one of the values in the set $\{II, \dots, VIII\}$ of 7 intensity categories. The value y_{ic} can be defined as a realization of the Multinomial distribution Y_{ic} with 7 categories and one trial; we denote this as

$$Y_{ic} \sim \text{Multinomial}(1, \pi_{II}, \dots, \pi_{VIII})$$

with $\pi_j = p(Y_{ic} = j)$ for $j \in \{II, \dots, VIII\}$.

We introduce now a latent (i.e. non observable) continuous and normally distributed variable $Y^\star_{ic}$ defined as

$$Y^\star_{ic} = \boldsymbol{X}_{ic}\boldsymbol{\beta} + \epsilon_{ic}$$

where $\boldsymbol{X}_{ic} = (X_{ic1}, \dots, X_{ick}, \dots, X_{icK})$ is the vector of K covariates (i.e. explanatory variables) with coefficients $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k, \dots, \beta_K)^T$ and ϵ_{ic} is a Gaussian random variable defined as $\epsilon_{ic} \sim \text{N}(0, \sigma^2)$ independently for each i and c. The latent variable represents the actual strength of the ground shaking for which we can observe only the effects through y_{ic}.

The relationship between Y_{ic} and $Y^\star_{ic}$ is given by

$$\begin{cases} Y_{ic} = II & \text{if } Y^\star_{ic} \le \tau_1 \\ \dots & \dots \\ Y_{ic} = j & \text{if } \tau_{\text{r}(j)-2} < Y^\star_{ic} \le \tau_{\text{r}(j)-1} \qquad \text{for } j = III, \dots, VII \\ \dots & \dots \\ Y_{ic} = VIII & \text{if } Y^\star_{ic} > \tau_6 \end{cases} \tag{1}$$

where $\text{r}(\cdot)$ is the rank function (e.g. $\text{r}(II) = 2$) and $\boldsymbol{\tau} = (\tau_1, \dots, \tau_6)$ is the vector of ordered thresholds to

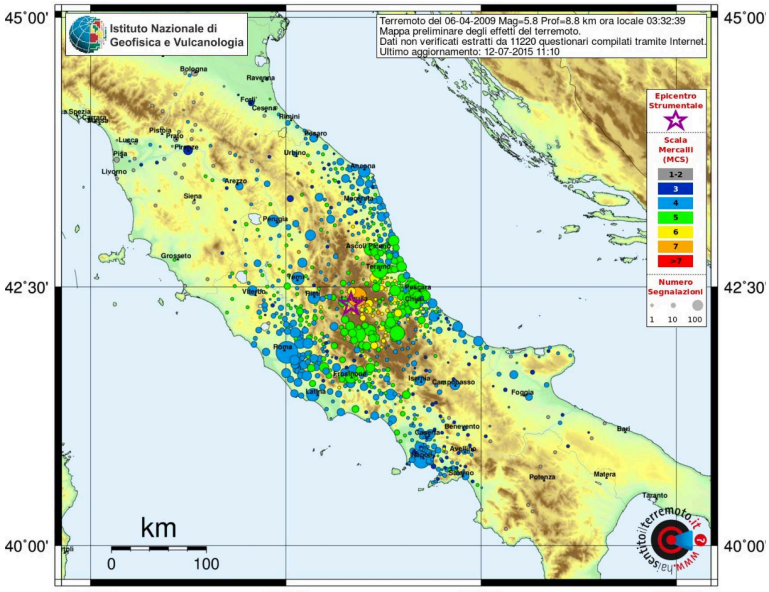

Fig. 1 Macroseismic intensity map from the HSIT web-site concerning the L'Aquila earthquake (April, 6th 2009, magnitude 5.8).

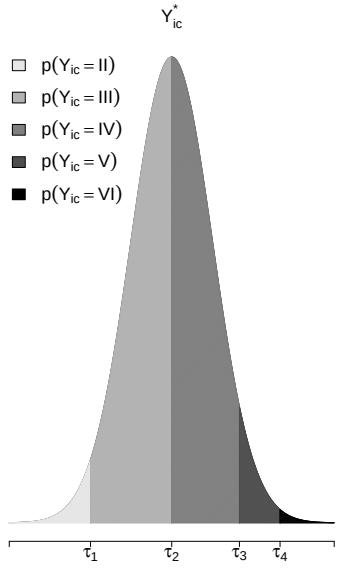

Fig. 2 Latent variable $Y_{ic}^{\star}$ and corresponding intensity probabilities.

be estimated. The number of thresholds is given by the number of intensity categories minus 1. To illustrate this relationship, we consider a simple example with a limited number of intensities ranging from II to VI, thus involving 5 categories and 4 thresholds $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_4)$. Figure 2 displays the distribution of the latent variable $Y_{ic}^{\star}$ and the corresponding intensity probabilities obtained using the relationship defined in Equation (1).

To compute the probability of having an intensity equal to II we proceed as follows:

$$p(Y_{ic} = II) = p(Y_{ic}^{\star} \leq \tau_1) = p(\boldsymbol{X}_{ic}\boldsymbol{\beta} + \epsilon_{ic} \leq \tau_1)$$
$$= p(\epsilon_{ic} \leq \tau_1 - \boldsymbol{X}_{ic}\boldsymbol{\beta}) = \Phi\left(\frac{\tau_1 - \boldsymbol{X}_{ic}\boldsymbol{\beta}}{\sigma}\right)$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard Normal distribution. In the same way the probability for a generic intensity $j \in \{III, \ldots, VII\}$ is given by

$$p(Y_{ic} = j) = p(\tau_{\mathrm{r}(j)-2} < Y_{ic}^{\star} \leq \tau_{\mathrm{r}(j)-1})$$
$$= p(\tau_{\mathrm{r}(j)-2} < \boldsymbol{X}_{ic}\boldsymbol{\beta} + \epsilon_{ic} \leq \tau_{\mathrm{r}(j)-1})$$
$$= \Phi\left(\frac{\tau_{\mathrm{r}(j)-1} - \boldsymbol{X}_{ic}\boldsymbol{\beta}}{\sigma}\right) - \Phi\left(\frac{\tau_{\mathrm{r}(j)-2} - \boldsymbol{X}_{ic}\boldsymbol{\beta}}{\sigma}\right).$$

Moreover, for the last intensity it holds that $p(Y_{ic} = VIII) = 1 - p(Y_{ic} \leq VII)$, where the cumulative probability for $j \in \{II, \ldots, VII\}$ is defined as

$$p(Y_{ic} \leq j) = \Phi\left(\frac{\tau_{\mathrm{r}(j)-1} - \boldsymbol{X}_{ic}\boldsymbol{\beta}}{\sigma}\right), \qquad (2)$$

with the property that $0 < p(Y_{ic} \leq II) < p(Y_{ic} \leq III) < \ldots < p(Y_{ic} \leq VIII) = 1$.

Following Agresti (2010), the *cumulative probit model* is defined as

$$\Phi^{-1}(p(Y_{ic} \leq j)) = \frac{\tau_{\mathrm{r}(j)-1} - \boldsymbol{X}_{ic}\boldsymbol{\beta}}{\sigma} \qquad (3)$$

for $j \in \{II, \ldots, VII\}$, where $\Phi^{-1}(\cdot)$ is the inverse of the Gaussian cdf which represents the so called *probit* function that links the cumulative probability to the linear predictor given by $\frac{\tau_{\mathrm{r}(j)-1} - \boldsymbol{X}_{ic}\boldsymbol{\beta}}{\sigma}$.

For identifiability reason[1], for probit models it is common to fix the first threshold τ_1 equal to 0. Moreover, as mentioned in Agresti (2010), since the observed ordinal scale provides no information about the variability of the latent variable $Y_{ic}^{\star}$, without loss of generality, we can set its standard deviation σ equal to 1. So Equation (3) becomes

$$\Phi^{-1}(p(Y_{ic} \leq j)) = \text{probit}(p(Y_{ic} \leq j)) = \tau_{\text{r}(j)-1} - \boldsymbol{X}_{ic}\boldsymbol{\beta}$$

for $j \in \{II, \ldots, VII\}$.

To illustrate the cumulative probit model and the interpretation of the covariate coefficients, we get back to the simple example introduced before with 5 categories (from II to VI) by assuming to have just one explanatory variable (thus $K = 1$ and $\boldsymbol{X}_{ic}$ is a scalar simply denoted by x_{ic}) which can take real values in the interval $[-6, +6]$. Moreover, we assume that the covariate coefficient β is positive. The top plot in Figure 3 depicts the cumulative probabilities $p(Y_{ic} \leq j)$ for different values of the covariate. It is worth noting that each curve (each one corresponds to a differ intensity) has the same shape since the coefficient β is common to all the categories, i.e. the covariate effect does not change according to the intensity. Moreover, it can be observed that for a given intensity j, when x_{ic} increases, the corresponding cumulative probability decreases, hence Y_{ic} is less likely to assume a value lower or equal to category j (and therefore values greater than j are more likely to occur). In fact, the bottom plot in Figure 3, which displays the probability $p(Y_{ic} = j)$ for different values of the covariate, shows that for small values of x_{ic} the lowest category occurs with the highest probability and the highest category happens for high values of x_{ic}. Note that for a given value of x_{ic} the sum of the 5 probabilities is equal to 1. For the case $\beta < 0$ (not reported here) the opposite happens: the cumulative probabilities increase as the covariate increases and the lowest category is more likely to happen for high values of x_{ic}.

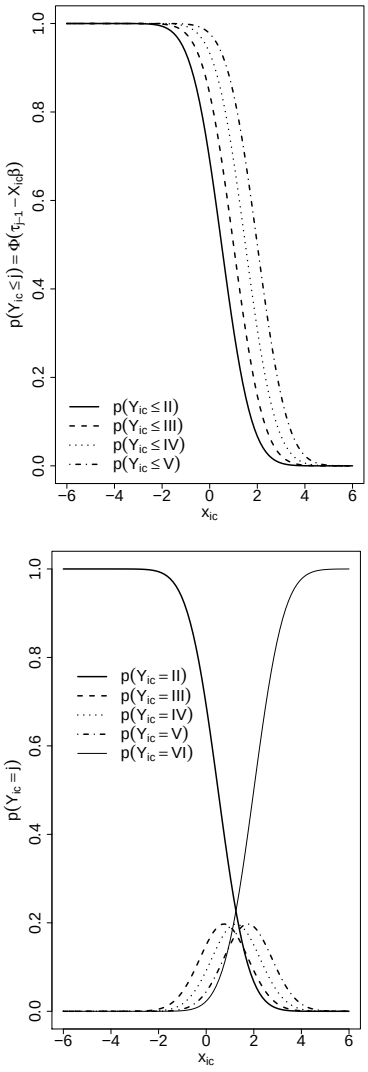

Fig. 3 Cumulative probabilities (top) and category probability distribution (bottom) for different values of the covariate when considering 6 categories and $\beta > 0$.

3.1 Estimation procedure in a Bayesian framework

The parameter vector for the cumulative probit model defined in the previous section is given by $\boldsymbol{\theta} = (\boldsymbol{\tau}, \boldsymbol{\beta})$. Bayesian inference via MCMC is carried out following the approach of Albert and Chib (1993) which is based on the data augmentation method (Tanner and Wong, 1987) that treats the latent variable $Y^{\star}$ as an additional parameter.

Let $\boldsymbol{X} = \left(\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_n^T, \ldots, \boldsymbol{X}_N^T\right)^T$ be the $(N \times K)$ covariate matrix, $\boldsymbol{y} = (y_1, \ldots, y_n, \ldots, y_N)^T$ the $(N \times 1)$ vector of observations and $\boldsymbol{Y}^{\star} = (Y_1^{\star}, \ldots, Y_n^{\star}, \ldots, Y_N^{\star})^T$ the $(N \times 1)$ latent variable vector. Note that the total number of cases $N \leq I \times C$ (I and C being the n. of municipalities and earthquakes respectively) since not all the earthquakes are felt in all the municipalities. The index $n = 1, \ldots, N$ refers to the case identified by the couple (i, c) with $i \in \{1, \ldots, I\}$ and $c \in \{1, \ldots, C\}$.

Given this notation and following the Gibbs sampler algorithm described in Albert and Chib (1993), the following full conditionals are derived when diffuse prior for $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ are used:

1. $p\left(\boldsymbol{\beta} \mid \boldsymbol{Y}^{\star}, \boldsymbol{y}\right) = \text{N}\left((\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y}^{\star}, (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\right)$,

[1]The model parameters $\boldsymbol{\beta}, \boldsymbol{\tau}, \sigma$ are not identified as any change in the scale parameter σ can be offset by changes in $\boldsymbol{\tau}$ and $\boldsymbol{\beta}$.

2. $p\left(Y_{ic}^{\star} \mid \boldsymbol{\beta}, \boldsymbol{\tau}, y_{ic}\right) = \mathrm{N}(\boldsymbol{X}_{ic}\boldsymbol{\beta}, 1)$ truncated at the left by $\tau_{\mathrm{r}(j)-1}$ and at the right by $\tau_{\mathrm{r}(j)}$ (with $j \in \{II, \ldots, VII\}$),

3. $p\left(\tau_{\mathrm{r}(j)} \mid \boldsymbol{Y}^{\star}, \boldsymbol{\beta}, \boldsymbol{y}, \{\tau_{\mathrm{r}(k)}, k \neq j\}\right) = \mathrm{Unif}\left(\gamma, \delta\right)$, where $\gamma = \max\{\max\left\{Y_{ic}^{\star} : Y_{ic} = j\right\}, \tau_{\mathrm{r}(j)-2}\}$, $\delta = \min\{\min\left\{Y_{ic}^{\star} : Y_{ic} = j+1\right\}, \tau_{\mathrm{r}(j)}\}$, where $j \in \{II, \ldots, VII\}$, $\tau_0 = -\infty$ and $\tau_7 = +\infty$.

To simulate values from the joint posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ the Gibbs sampler draws values iteratively from all the conditional distributions. For implementing such a procedure we resort to the `MCMCoprobit` function of the `MCMCpack R` package (R Core Team, 2015), whose details are reported in Andrew *et al.* (2011).

4 Application

4.1 Data and model specification

The considered data refer to $C = 1917$ earthquakes occurred in the Italian territory from January 2009 to August 2015 with magnitude (M_L, measured by Richter scale) ranging from to 2 to 5.9 and depth lower than 35 km. Most of the events had M_L between 2 and 4 (about 95%) while the percentage of earthquakes with a magnitude greater than 5 is about 0.5%. The intensities (on the MCS scale) range from II to VII with the modal intensity II occurring in 46% of the cases.

In order to have more reliable data, we selected the macroseismic intensities of the municipalities having more than ten questionnaires for each seismic event, resulting in $I = 945$ municipalities. Each municipality may have experienced more than one seismic event, so that the finals database consists of $N = 6723$ cases. Besides intensity, for each municipality and earthquake, the log_{10}-hypocentral distance (logD) and the magnitude are available. Thus, the covariate vector for each case is given by $\boldsymbol{X}_{ic} = (1, M_{Lic}, \mathrm{logD}_{ic})$, where the term 1 refers to the intercept with β_0 coefficients. As there are 6 intensity categories (from II to VII) we have 5 thresholds $\boldsymbol{\tau} = (\tau_1 = 0, \tau_2, \ldots, \tau_5)$ and the vector of unknown parameters is $\boldsymbol{\theta} = (\boldsymbol{\tau}, \beta_0, \beta_{M_L}, \beta_{\mathrm{logD}})$.

In order to ensure convergence of the Gibbs sampler, we ran chains of 2500000 iterations, with a burn-in of 500000 and a thin interval of 200. Convergence was assessed by monitoring the mixing of the chains, through trace plots, together with the Gelman-Rubin and Geweke diagnoses (Gelman and Rubin, 1992, Geweke, 1992).

Table 1 Posterior parameter estimates of the ordered probit model: mean, stand deviation (Sd) and 95% highest posterior density interval (HPD).

	Mean	Sd	Lower HPD	Upper HPD
β_0	-0.837	0.070	-0.976	-0.700
β_{M_L}	2.464	0.038	2.391	2.541
β_{logD}	-5.229	0.089	-5.408	-5.059
τ_2	1.385	0.030	1.330	1.444
τ_3	4.981	0.091	4.804	5.160
τ_4	6.628	0.143	6.356	6.911
τ_5	7.912	0.220	7.497	8.348

4.2 Results

Convergence diagnoses indicated a good chain mixing for all parameters. In particular, the Geweke z statistics (in absolute value) range from 0.24 to 0.78 with p-values bigger than 0.43, thus confirming the convergence achievement. Similarly, the Gelman-Rubin diagnoses, computed by running two independent chains for each parameter, produce a potential scale reduction factor lower than 1.1 for all parameters.

The posterior parameter estimates are reported in Table 1. It can be seen that all the parameters are significantly different from zero (95% credible intervals do not include zero). Moreover, the magnitude coefficient β_{M_L} is positive with posterior mean equal to 2.464. This means that (keeping all the other covariates fixed) a change in the magnitude of 1 degree causes an increase in the latent variable $Y^{\star}$ of 2.464 (on average). Concerning the influence of M_L on the response variable (i.e. the intensity), we can conclude that when the magnitude increases the cumulative probability of observing an intensity lower than or equal to the generic category decreases (and the probability of observing a higher intensity increases). This is a situation similar to the one plotted in Figure 3. Differently, the posterior mean of the distance coefficient β_{logD} is equal to -5.229. This means that (keeping all the other covariates fixed) a change distance of 1 (on the kilometer logarithmic scale) causes an average change in the latent variable of -5.229. As expected, with respect to the response variable, when the distance increases, the cumulative probability of observing a given intensity, or one lower, increases (and the probability of observing a higher intensity decreases). This is reasonable and in line with the nature and the geophysical characteristic of the phenomenon under study (Schubert, 2015).

Once the model parameters have been estimated, it is possible to compute, for any desired value of the log-hypocentral distance and of the magnitude, the intensity posterior distribution, i.e. the posterior probabilities of occurrence for every intensity value $j \in$

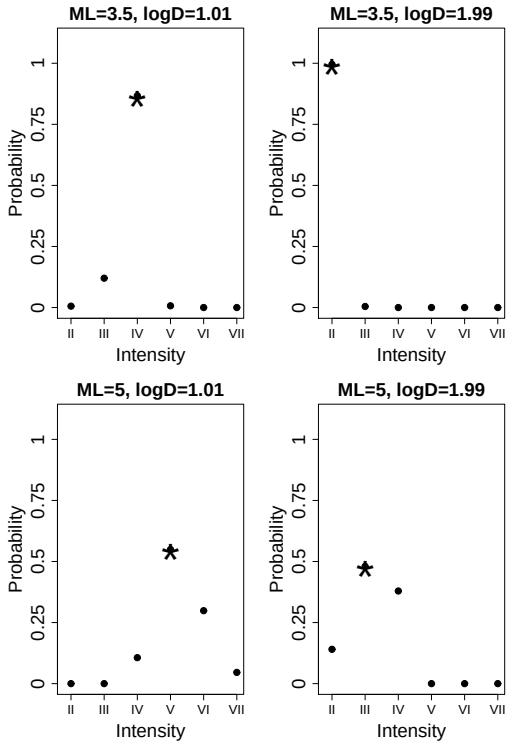

Fig. 4 Probability distribution of intensity given some values of magnitude (3.5 and 5) and logD (1.01, 2). The star denotes the highest probability intensity (modal intensity).

$\{II, \ldots, VII\}$ using the following formula:

$$p(Y_{ic} = j) = \Phi\left(\frac{\hat{\tau}_{\mathrm{r}(j)-1} - (\hat{\beta}_0 + \hat{\beta}_{M_L} \cdot M_{Lic} + \hat{\beta}_{\mathrm{logD}} \cdot \mathrm{logD}_{ic})}{\sigma}\right) - \Phi\left(\frac{\hat{\tau}_{\mathrm{r}(j)-2} - (\hat{\beta}_0 + \hat{\beta}_{M_L} \cdot M_{Lic} + \hat{\beta}_{\mathrm{logD}} \cdot \mathrm{logD}_{ic})}{\sigma}\right), \quad (4)$$

where the hat notation is used to denote the posterior parameter mean. As mentioned in Section 3, here σ is fixed equal to 1 and for the first and last category the formula is adapted accordingly.

Figure 4 displays the intensity probability distribution for two given values of M_L (3.5, 5) and logD (1.01, 2); the category with the highest probability (i.e. the modal intensity) is depicted by a star. As we can see, with a moderate magnitude (M_L=3.5) and with a short distance (logD=1.01) the modal intensity is IV (with a probability of about 0.85); instead, when logD=2, the modal intensity becomes II. Coherently, with a higher magnitude (M_L=5) the modal intensity (with a probability of about 0.5) is V for the shorter distance and decreases down to III at the longer distance.

We focus now on the main objective of this work, i.e. the definition of a new intensity prediction equation based on the intensity probability distribution. In particular, we analyze the effect of a distance change on the intensity distribution (computed using Equation (4)) by determining the modal intensity for different values of M_L (3.5 and 5) and logD (100 values between 0.5 and 3). Figure 5 displays the modal intensity according to distance (i.e. the estimated IPE). Each point represents the modal intensity with its corresponding probability (using the classes [0,0.5], (0.5,0.8], (0.8,1]). In particular, with M_L=3.5 and logD$\simeq$1 the intensity IV (dark gray point) has a probability of occurrence in $(0.8, 1]$; the same probability class is reached for distance larger than 1.7 by intensity II. Notably, we consider the probability associated to each intensity as a measure of uncertainty. The segments departing from each point indicate which intensities have to be accounted, together with the modal one, for reaching an occurrence probability of at least 0.8. Looking, for example, at the bottom panel of Figure 5, with logD=2 the modal intensity is III with a probability lower than 0.5 (light gray point). From this point a light gray segment departs toward intensity IV, that has an occurrence probability of about 0.38 (see bottom-right panel of Figure 4): this means that the probabilities of degree III and IV together reach at least 0.8. In this sense, points with no segment refer to very reliable intensities (i.e., probability bigger than 0.8), whereas point with one or two segments refer to more uncertain modal intensity.

By comparing several IPEs (note reported here), we can conclude that with lower magnitudes the IPEs decrease more rapidly and show less uncertainty with respect to those generated from earthquakes with higher magnitudes.

4.3 Analysis of residuals

Once IPE is defined for a given magnitude, it can be used as an operative tool to compute expected intensities (and probabilities) as a function of the hypocentral distance. Computing the residuals between observed and expected intensities is paramount in defining anomalous areas, with positive (negative) residuals possibly associated with seismic waves amplification (attenuation) (Papoulia and Stavrakakis, 1990).

Considering the range of observed macroseismic intensities (from II to VII), in order to improve the seismic interpretability of residuals, we exclude from the analysis the cases with estimated modal intensity equal to the minimum and maximum values (II and VII) because they would give rise to residuals which are always positive/negative or equal to zero. For each municipality i, where a number of m_i earthquakes were felt, m_i observed intensities I^{Obs} are available from

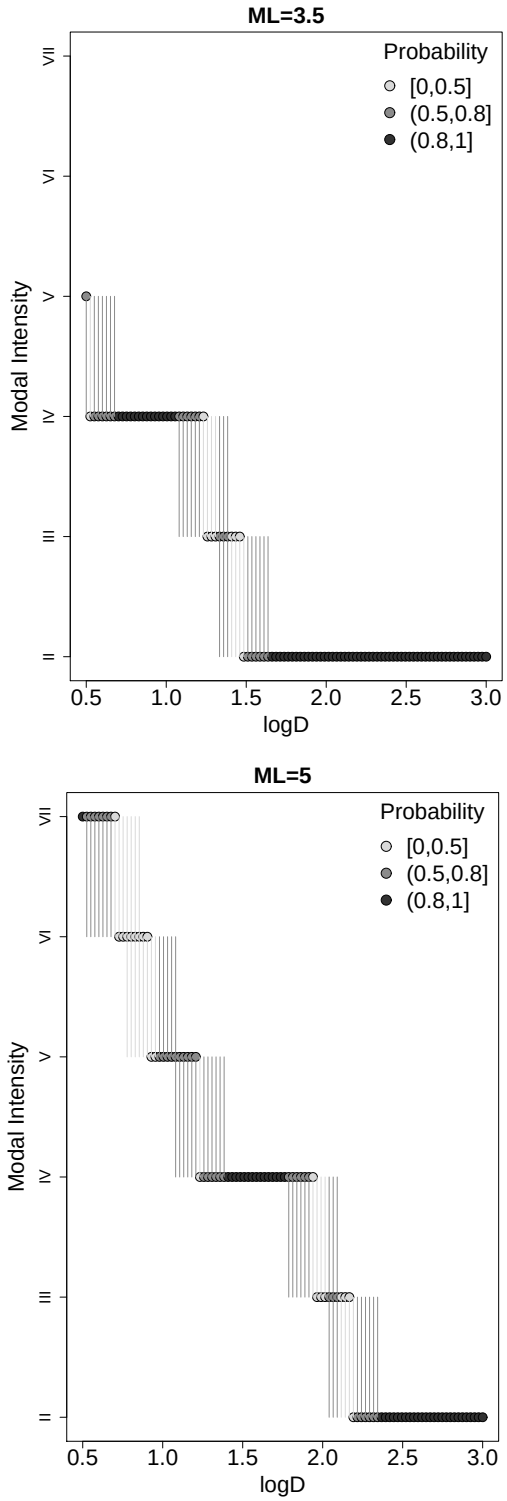

Fig. 5 IPE for some values of magnitude (3.5 and 5) and logD (100 values between 0.5 and 3). The color of points represents the probability (in classes) associated with the modal intensity. The segments of each point indicate which intensities have to be accounted, together with modal intensity, for reaching an occurrence probability of at least of 0.8.

the HSIT web-site, together with m_i intensity probability distributions with modal category $\hat{I}$ obtained by the ordered probit model. It is thus possible to derive for each municipality m_i residual probability distributions, each being a discrete random variable defined as $(I_{ic}^{Obs} - \hat{I}_{ic})$ with probabilities p_{ic} obtained by Equation (4) and $c = 1, \ldots, m_i$. Then we calculate the random variable *sum of residuals* denoted by R_i, obtained by summing the m_i residual probability distributions. Finally, for each municipality we are able to estimate the mean residual and its corresponding variance as the expected value and variance of the random variable *mean of residuals* R_i/m_i.

Figure 6 shows the obtained mean residuals and standard deviations. Blue circles correspond to municipalities where observed intensities are lower than estimated intensities, suggesting seismic attenuation; while red points, with positive residuals, point out seismic amplification. Gray shaded areas correspond to high values of the residual standard deviation. Interestingly, positive (red) and negative (blue) values tend to be spatially aggregated, whereas the areas of high standard deviation correspond to a greater uncertainty of the municipality intensity data. The prevalence of orange circles in North Italy highlights an amplification area localized in between the Alps chain and the Padana plain. This could be caused by the presence of sedimentary basin trapping and amplifying seismic waves. The greater part of municipalities near the North Apennines have negative residuals revealing an attenuation area. Furthermore, we can highlight other two areas with prevalence of positive mean residuals: one in central Italy with a North-South elongation, and the second one located at the south of Naples.

5 Discussion

Italy, as one of the most seismically active countries, needs an effective and reliable analysis of seismic risk. The possibility of defining zones of high seismic shaking is a crucial goal for promoting effective policy to prevent major damages. In this regard, a reliable IPE definition is the necessary step. It offers an operative way to calculate the expected intensity given the earthquake magnitude and the hypocentral distance. On the other hand, it is worth to note that the intensity evaluation based only on the IPE is not complete, due to several factors which may significantly change the observed shaking. For this reason a common procedure consists in performing a residual analysis with observed and estimated intensity data. If these residuals are spatially homogeneous they can be caused by the influence of regional geological condition or by the predominating

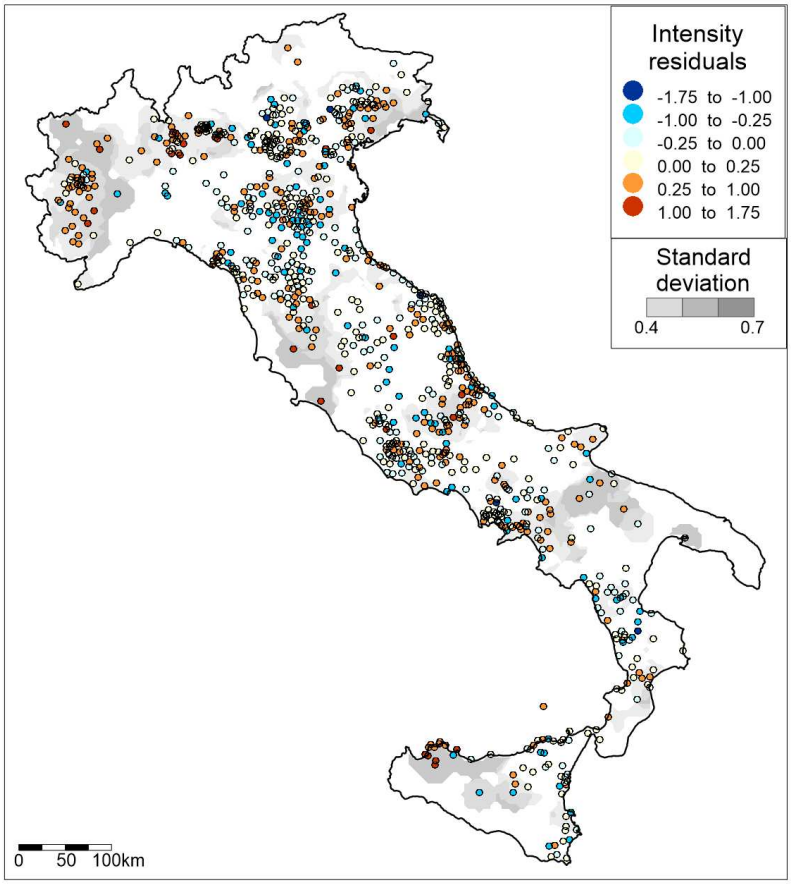

Fig. 6 Map of the mean and standard deviation of residuals. Geographic coordinates are between 6.6° and 20° East longitude and between 36.6° and 46.7° North latitude. Colored circles represent the municipality and the gray shaded contours represent the corresponding standard deviations.

source mechanism (Sbarra *et al.*, 2012). Analyzing several events for each municipality, the source mechanism contribution to the intensity is reduced, thus evidencing a possible local effect related, for example, to geological characteristics which could be included in the model as covariates. However, note that for low magnitude earthquakes (which constitute the major part of the HSIT dataset), the focal mechanism solutions are almost unknown.

Further confirmation are necessary to validate our findings because of the short length of the HSIT data series (2009-2015) which could not be fully representative of a seismicity of long period. However, our results are consistent with those found in previous works (see e.g. Albarello *et al.*, 2002) and, at the same time, provide an interesting new benchmark for comparison with any other risk maps carried out for these kind of data. In this sense, our work can be considered as a first step to detect local responses to seismic shaking in Italy.

From a methodological point of view, we employed the ordered probit model using a Bayesian approach. Although this model is well established in the statistical literature, its application to a large amount of macroseismic intensity data is original and unavoidable for defining a reliable IPE which takes properly into account the ordinal nature of data.

A possible extension of this work could deal with the spatial structure of the data, by including a spatial process in the model equations. In literature, models for spatially correlated ordered categorical data are relatively new (Brewer *et al.*, 2004, Higgs and Hoeting, 2010). The main obstacle to implement such models concerns the computational burden that can negatively impact the performance of MCMC model-fitting algorithms (Berrett and Calder, 2012), making the estimation procedure for large data set unfeasible. Unfortunately, as far as we know, not even other more efficient algorithms alternative to MCMC, such as the

Integrated Nested Laplace Approximation (INLA, Bangiardo and Cameletti, 2015), can be applied as they are not available for ordinal response data. Thus, the development of computationally effective ad-hoc algorithms needs to be addressed in the future research for analyzing the complete HSIT dataset through a spatial model. Another possibility would consist in restricting the analysis to a small target area - identified for example using the residual map in Figure 6 - in order to apply the algorithm proposed by Higgs and Hoeting (2010).

6 Acknowledgements

This research was partially founded by INGV-DPC agreement S2-2012 - Constraining Observations into Seismic Hazard. The authors would like to thank two anonymous reviewers who provided valuable comments.

References

1. Agresti, A. (2010). *Analysis of Ordinal Categorical Data*. Wiley.
2. Albarello, D., Bramerini, F., D' Amico, V., Lucantoni, A., and Naso, G. (2002). Italian intensity hazard maps: a comparison of results from different methodologies. *Bollettino di Geofisica teorica ed applicata*, **43**, 249–262.
3. Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**, 669 – 679.
4. Andrew, D. M., Kevin, M., and Jong Hee, P. (2011). MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software*, **42**, 1–21.
5. Atkinson, G. and Wald, D. (2007). "Did you feel it?" intensity data: A surprisingly good measure of earthquake ground motion. *Seismological Research Letters*, **78**, 362–368.
6. Azzaro, R., D'Amico, S., Rotondi, R., Tuvé, T., and Zonno, G. (2013). Forecasting seismic scenarios on etna volcano (italy) through probabilistic intensity attenuation models: A bayesian approach. *Journal of Volcanology and Geothermal Research*, **251**, 149 – 157.
7. Bangiardo, M. and Cameletti, M. (2015). *Spatial and Spatio-temporal Bayesian Models with R-INLA*. John Wiley and Sons, Ltd.
8. Berrett, C. and Calder, C. (2012). Data augmentation strategies for the Bayesian spatial probit regression model. *Computational Statistics and Data Analysis*, **56**, 478–490.
9. Brewer, M. J., Elston, D. A., Hodgson, M., Stolte, A. M., Nolan, A. J., and Henderson, D. J. (2004). A spatial model with ordinal responses for grazing impact data. *Statistics and Probability*, **4**, 127–143.
10. Cowles, M. (1996). Accelerating Monte Carlo Markov Chain convergence for cumulative-link generalized linear models. *Journal of the American Statistical Association*, **6**, 101–111.
11. Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, pages 457–472.
12. Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian Statistics 4, (Bernardo, J.M., Berger, J.O., Dawid, A.P. and Smith, A.F.M. eds.) Oxford: Oxford University Press*, pages 169–193.
13. Gómez Capera, A. A. (2006). Seismic hazard map for the Italian territory using macroseismic data. *Earth Sciences Research Journal*, **10**, 67 – 90.
14. Grünthal, G. E. (1998). European macroseismic scale 1998. *Cahiers du Centre Européen de Geodynamique et de Séismologie*, **15**, 1–99.
15. Higgs, M. D. and Hoeting, J. A. (2010). A clipped latent-variable model for spatially correlated ordered categorical data. *Computational Statistics and Data Analysis*, **54**, 1999–2011.
16. Magri, L., Mucciarelli, M., and Albarello, D. (1994). Estimates of site seismicity rates using ill-defined macroseismic data. *Pure Applied Geophysics*, **143**, 617–632.
17. Mak, S., Clements, R. A., and Schorlemmer, D. (2015). Validating Intensity Prediction Equations for Italy by Observations. *Bulletin of the Seismological Society of America*, **105**(6), 2942–2954.
18. Papoulia, J. E. and Stavrakakis, G. N. (1990). Attenuation laws and seismic hazard assessment. *Natural Hazards*, **3**, 49–58.
19. Pasolini, C., Albarello, D., Gasperini, P., D'Amico, V., and Lolli, B. (2008). The attenuation of seismic intensity in italy, part ii: Modeling and validation. *Bulletin of the Seismological Society of America*, **98**, 692–708.
20. R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
21. Rotondi, R., Tertulliani, A., Brambilla, C., and Zonno, G. (2008). The intensity attenuation of Colfiorito and other strong earthquakes: the viewpoint of forecasters and data gatherers. *Annals of Geophysics*, **51**, 499–507.
22. Rotondi, R., Varini, E., and Brambilla, C. (2015). Probabilistic modelling of macroseismic attenuation and forecast of damage scenarios. *Bulletin of Earth-*

quake Engineering, pages 1–20.

23. Sbarra, P., Tosi, P., and De Rubeis, V. (2010). Web-based macroseismic survey in Italy: method validation and results. *Natural Hazards*, **54**(2), 563–581.
24. Sbarra, P., De Rubeis, V., Di Luzio, E., Mancini, M., Moscatelli, M., Stigliano, F., Tosi, P., and Vallone, R. (2012). Macroseismic effects highlight site response in Rome and its geological signature. *Natural Hazards*, **62**, 425–443.
25. Schubert, G. (2015). *Treatise on Geophysics*. Elsevier Science.
26. Sieberg, A. (1930). Geologie der erdbeben. *Handbuch der Geophysik*, **2**, 552–555.
27. Tanner, T. and Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, **82**, 528–550.
28. Tosi, P., Sbarra, P., De Rubeis, V., and Ferrari, C. (2015). Macroseismic intensity assessment method for web questionnaires. *Seismological Research Letters*, **86**(3), 985–990.
29. Zonno, G., Rotondi, R., and Brambilla, C. (2009). Mining macroseismic fields to estimate the probability distribution of the intensity at site. *Bulletin of the Seismological Society of America*, **99**, 2876–2892.