Bulletin of Earthquake Engineering manuscript No.
(will be inserted by the editor)

Probabilistic modelling of macroseismic attenuation and forecast of damage scenarios

Renata Rotondi · Elisa Varini · Carla Brambilla

Received: date / Accepted: date

Abstract According to the idea now widespead that macroseismic intensity should be expressed in probabilistic terms, a beta-binomial model has been proposed in the literature to estimate the probability of the intensity at site in the Bayesian framework and a clustering procedure has been adopted to define learning sets of macroseismic fields required to assign prior distributions of the model parameters. This article presents the results concerning the learning sets obtained by exploiting the large Italian macroseismic database DBM1I11 [5] and discusses the problems related to their use in probabilistic modelling of the attenuation in seismic regions of the European countries partners of the UPStrat-MAFA project [17], namely South Iceland, Portugal, SE Spain and Mt Etna volcano area (Italy). Anisotropy and the presence of offshore earthquakes are some of the problems faced. All the work has been carried out in the framework of the Task B of the project.

Keywords Macroseismic intensity · Beta-binomial probability model · Anisotropy · Seismic scenario · Source models

CNR - Institute of Applied Mathematics and Information Technology *Enrico Magenes*,
Via Bassini 15, Milano, Italy
Fax: +39-02-23699538

Renata Rotondi
Tel: +39-02-23699528
E-mail: reni@mi.imati.cnr.it

Elisa Varini
Tel: +39-02-23699527
E-mail: elisa@mi.imati.cnr.it

Carla Brambilla
Tel: +39-02-23699523
E-mail: carla@mi.imati.cnr.it

1 Introduction

The capability to forecast seismic scenarios in terms of macroseismic intensity at a site is of great importance and the issue has largely been analysed in the past by modelling the intensity attenuation according to a deterministic point of view. Nowadays, however, the idea that the intensity at a site, as well as the intensity decay, must be expressed in probabilistic terms in order to obtain a more complete treatment of its intrinsic uncertainty is widespread [8], [6], [16]. According to this idea, [13] proposed to estimate the probability distribution of the intensity at a site, conditioned on the epicentral intensity and on the epicentre-to-site distance, by using a beta-binomial model. The estimation process is carried out according to the Bayesian paradigm, exploiting a learning set of macroseismic fields to assign prior distributions of the model parameters. The model was at first tested on the Camerino (28/07/1799) and Colfiorito (1997/9/26) earthquakes by using, as learning set, a set of macroseismic fields from seismogenetic zones of the zonation ZS4 [7] judged homogeneous from the viewpoint of kinematic context and expected rupture mechanism to the zone which the epicentres of the two earthquakes belong to.

The definition of a suitable learning set is a key point for an extensive use of the model. [19] suggested to apply a clustering procedure to macroseismic fields chosen from Italian macroseismic databases to derive classes of fields internally homogeneous from the attenuation point of view, and to use these classes as potential learning sets in subsequent studies aimed at forecasting damage scenarios in terms of macroseismic intensity. The idea underlying this proposal is that the estimation of the model parameters improves if it is possible to choose a learning set whose attenuation trend fits as much as possible the one characterizing the situation under study. Moreover, since the difference in the decay trend depends on many geological characteristics, not all available or easily measurable, they launched the innovative idea of describing the macroseismic fields through summaries of the spatial distribution of the intensity decay and to base the clustering procedure on them. This approach for constructing potential learning sets has been first applied to the macroseismic fields of 55 hearthquakes of epicentral intensity MCS $\geq VII$, selected from the DBMI04 Italian database [15] and judged to be representative of spatial and temporal distribution of the Italian seismicity.

In the UPStrat-MAFA project [17] the strategy above mentioned has been refined and the proposed model has been applied to macroseismic fields of European seismic regions of Iceland, Portugal, Spain, and in Mt Etna volcano area. The aim is that of contributing to implement common strategies to assess seismic hazard in terms of intensity through probabilistic modelling of the different attenuation trends.

We want to highlight that the estimation of the intensity at site allows to forecast the severity of the damages that a future earthquake can cause and the expected extension of the area hit by the event so as to expedite the selection of the strategies of post-seismic intervention. From a long-term perspective this knowledge allows to grade and prioritize the interventions addressed to risk mitigation and protection of urban settlements from an earthquake [9].

This article is divided in three main parts. In the first part we explain the methodological aspects of the strategy in detail; in the second one we introduce the potential

learning sets we derived on the basis of the most updated database of Italian macroseismic fields; in the third part we present the answers given to some issues arisen in the analysis of the test areas.

2 Some statistical methods and a probability model in seismic attenuation

2.1 Construction of learning sets

Given an as large as possible reference set of N macroseismic fields produced by earthquakes that occurred in a wide region, it is necessary to examine whether the attenuation trends are considerably different and, in this case, to apply a clustering procedure which identifies groups of fields homogeneous from the attenuation point of view. Each clustering procedure rests on the description of the objects to be clustered by means of a set of attributes; in the case of macroseismic fields, an obvious choice is to use summaries of the spatial distribution of the intensity decay as attributes. More precisely, each macroseismic field may be characterized by the mean, the median and the 3^{rd} quartile of each set of distances between the epicentre and the sites where the same intensity I_s, or, equivalently, the same ΔI, was observed. This results in characterizing each macroseismic field by a vector of $q = 3$ I_0 attributes, where I_0 is the value of the epicentral intensity. The $N \times 3$ *max* I_0 matrix obtained by considering all of the N fields is the basis for the computation of a dissimilarity matrix among the macroseismic fields, that is the next step required for the application of a clustering method. It has to be pointed out that this representation of the macroseismic fields is likely to produce missing values since some of the intensities I_s in the range $(1, I_0)$ may have not been recorded in some of the macroseismic fields and, moreover, the value I_0 itself is not the same for all of the macroseismic fields. However, this is not a problem because the methods used in the following are able to handle missing data. For details we refer the interested reader to [4].

As a measure of dissimilarity we use the so-called Manhattan distance, that is

$$d(i,j) = \sum_{k=1}^{q} \mid x_{ik} - x_{jk} \mid,$$

where $(x_{i1}, x_{i2}, \ldots, x_{iq})$ and $(x_{j1}, x_{j2}, \ldots, x_{jq})$ are the attributes of the two macroseismic fields i and j. This measure is not particularly sensitive to outliers and this is reason for which it has been chosen.

The clustering method used to group the macroseismic fields belongs to the class of the hierarchical agglomerative methods and it is known as Ward method [4]. We choose this class of methods since we think it allows a more thorough understanding of the clustering process than the methods designed for a fixed number of groups, and in particular the Ward method since it does not suffer from the drawbacks typical of other agglomerative methods, such as, for example, the chaining effect typical of the single linkage method. According to the agglomerative methods each object, in our case each macroseismic field, is initially considered as a separate cluster; therefore at the beginning of the procedure one has as many clusters as fields to be clustered. Subsequently, at each step, the number of clusters is reduced by one by merging the

two clusters whose combination optimizes a given objective function. The merging process continues until only one cluster is left. In particular, Ward method minimizes the loss of information associated with each merger. Let $\bar{x}_R$ denote the centroid of a cluster R, defined by the value $\bar{x}_k = 1/m_R \sum_{i \in R} x_{ik}$, $k = 1, \ldots, q$, where m_R is the number of macroseismic fields (objects) in the cluster. The loss of information associated with the cluster is defined in terms of the sum of the distances from its centroid:

$$E_R = \sum_{i \in R} d(x_i, \bar{x}_R),$$

where, in our case, d is the Manhattan distance above defined. At each step the union of every possible pair of clusters R and S, generating a new cluster T, is considered and the two clusters whose merger results in the minimum increase ΔE are combined, where ΔE is given by:

$$\Delta E = \min_{R,S} \Delta E_{RS} = E_T - E_R - E_S.$$

Several quantitative and graphical tools can provide insight into the hierarchy of clusters so obtained. The most used graphical tool is the dendrogram, that depicts the whole merging process, step by step; another is the so-called silhouette plot [14]. Examples of dendrogram and of silhouette plots are provided in section 3, as well as an explanation of this last plot. See also [19].

2.2 Probabilistic analysis

2.2.1 Beta-binomial model

Conditioned on the epicentral intensity I_0, and on a fixed epicentral distance, the intensity at a given site I_s is assumed to have a binomial distribution with parameter p:

$$Pr(I_s = i \,|\, I_0 = i_0, p) = \binom{i_0}{i} p^i (1-p)^{i_0 - i} \ ,$$

which is equivalent to assume that also the intensity decay ΔI has a binomial distribution with parameter p since

$$Pr(\Delta I = i_0 - i \,|\, I_0 = i_0, p) \ = Pr(I_s = i \,|\, I_0 = i_0, p).$$

The choice of the binomial distribution is predicated on respecting as far as possible the ordinal nature of the intensity scale applied. The parameter p, in its turn, is taken as a random variable in order to account for the variability in ground shaking even among sites at the same epicentral distance, and it is assumed to have the beta distribution:

$$Be(p; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^p x^{\alpha - 1} (1-x)^{\beta - 1} dx$$

with hyperparameters α and β. The beta distribution is chosen because of its great flexibility and tractability within the Bayesian framework. Mean and variance are given, respectively, by:

$$E(p) = \frac{\alpha}{\alpha+\beta} \qquad \sigma^2(p) = \frac{\alpha\beta}{(\alpha_+\beta)^2(\alpha+\beta+1)}. \tag{1}$$

First we assume that the decay is isotropic, i.e., we assume to have a point source and circular isoseismal lines bounding the points of equal intensity. In section 4.1 we shall drop this assumption and also consider the anisotropic case. In the present case we draw J circular bins around the epicentre and suppose that in all of the sites within each j-th bin, I_s - so as ΔI - has the same binomial distribution with parameter p_j, i.e.:

$$Pr(I_s = i \,|\, I_0 = i_0, p_j) = Pr(\Delta I = I_0 - i \,|\, I_0 = i_0, p_j) = \binom{i_0}{i} p_j^i (1-p_j)^{i_0-i}. \tag{2}$$

In its turn, each p_j has a beta distribution with hyperparameters α_j and β_j. The width of the bins may vary depending on the situation under study; in the analysis of the macroseismic fields of the Italian database DBMI11 [5] we have taken bins of 10 km.

2.2.2 Parameters estimation

In the Bayesian framework, the estimation procedure of the model parameters consists of the following main steps: elicitation of *prior* distributions of the parameters from our beliefs on the phenomenon, computation of the *posterior* distributions given the current observations, and evaluation of the parameter estimator. In the present context we proceed in different way depending on whether we have sufficient information "to learn from the past" or nor. In Italy, the availability of the database DBMI11 allows us to build classes of macroseismic fields with similar intensity decay (section 3) and to use them as sources of prior information; on the contrary, as for other European countries where such a database is missing, in section 4.2 we describe how and with which of those classes the macroseismic fields of the area under study may be matched, by using those classes as potential learning sets for all the case studies carried out inside the UPStrat-MAFA project [12].

After selecting an attenuation class $\mathscr{C}$, we assume that our aim is to analyse the attenuation of earthquakes, belonging to this class, of fixed epicentral intensity I_0. If, for instance, I_0 = VII, then we will have $\mathscr{C} = \mathscr{C}_0 \cup \mathscr{C}_1$, where $\mathscr{C}_0$ is the set of the fields generated by earthquakes of intensity VII and $\mathscr{C}_1$ the set of the remaining fields. In each bin drawn around the epicentre, on the basis of the information provided by the macroseismic fields which constitute the set $\mathscr{C}_1$, we assign the hyperparameters $\alpha_{j,0}$ and $\beta_{j,0}$ of the prior beta distribution of the parameter p_j in the following way. The probability that the decay is equal to 0 is given by:

$$Pr(I_s = i_0 \,|\, I_0 = i_0, p_j) = Pr(\Delta I = 0 \,|\, I_0 = i_0, p_j) = p_j^{i_0} \tag{3}$$

and it can be roughly estimated by using the relative frequency of null decay $N_j(i_0)/N_j$, where $N_j(i_0)$ is the number of the N_j sites in the jth bin where the intensity at site is not smaller than the epicentral intensity. In this way the initial mean value for p_j will be $p_{j,0} = (N_j(i_0)/N_j)^{1/i_0}$. To overcome the problem that this procedure is not applicable in the bins where there is no report of null decay, the available values $p_{j,0}$ are approximated by the smoothing inverse power function $g(d) = [c_1/(c_1+d)]^{c_2}$, whose coefficients c_1, c_2 are estimated by the method of least squares. In this way we are able to obtain initial values $p_{j,0}$ for every jth bin. The variance $\sigma^2(p_j)$ must satisfy some inequalities of the form $b_{1j} < \sigma_j^2 < b_{2j}$, suggested by the significance of p_j in the decay process, as indicated in [13]; therefore, we have chosen to set $\sigma^2(p_j) = b_{1j} + 0.99 \times (b_{2j} - b_{1j})$. After assigning the variance $\sigma^2(p_j)$, we invert (1) to obtain the values of the hyperparameters $\alpha_{j,0}$ and $\beta_{j,0}$.

Being the beta distribution a conjugate prior for the binomial model, the posterior distribution of p_j is again a beta distribution. Then, on the basis of the macrosesmic fields belonging to $\mathscr{C}_0$, we update the parameters of each posterior beta distribution according to the Bayesian approach, and estimate the parameters p_j through their posterior mean:

$$\hat{p}_j = \frac{\alpha_{j,0} + \sum_{n=1}^{N_j} i_s^{(n)}}{\alpha_{j,0} + \beta_{j,0} + i_0 \cdot N_j}, \tag{4}$$

where N_j is the total number of sites that are in the jth bin and $i_s^{(n)}$ is the intensity at the n-th site. We point out that the estimates really updated are only those associated with bins where data points were observed; therefore we again smooth these values $\hat{p}_j$ through a new inverse power function $g(d) = [c_1/(c_1+d)]^{c_2}$ by obtaining, in this way, a binomial distribution of I_s, $Pr(I_s|I_0; g(d))$, conditioned on I_0, at any distance d from epicentre.

2.2.3 Forecasting

Whenever we have new observations, once estimated the parameters p_j through their posterior mean (4), and smoothed these estimates $\hat{p}_j$ by using a specific inverse power function $g(d) = [c_1/(c_1+d)]^{c_2}$, we are able to forecast, in terms of macroseismic intensity I_s at site, the damage scenario that a future earthquake of given intensity I_0 could cause by the *smoothed* binomial probability distribution:

$$Pr_{smooth}(I_s = i \,|\, I_0 = i_0; g(d)) = \binom{i_0}{i} g(d)^i \, (1 - g(d))^{(i_0 - i)} \tag{5}$$

and by using the mode i_{smooth} of this distribution as forecast value of the intensity I_s at any site distant d from the epicentre.

Having an entire probability distribution for the variable I_s to predict, instead of just its estimate as in the deterministic attenuation laws, is a great potentiality of our probability model; indeed this allows to better express the uncertainty of the phenomenon by computing, for instance, the probability that I_s exceeds a given value at a fixed site.

2.2.4 Validation

To validate the results three criteria are proposed. The first one is the so-called logarithmic scoring rule [18], based on the logarithm of the likelihood function

$$score_{smooth} = -\frac{1}{N'} \log \prod_{n=1}^{N'} \binom{i_0}{i_s^{(n)}} g(d_n)^{i_s^{(n)}} \left(1 - g(d_n)\right)^{(i_0 - i_s^{(n)})} , \tag{6}$$

where N' is the total number of the observed intensities at site, $i_s^{(n)}$ is the intensity at the n-th site and d_n is the distance of the n-th site from the epicentre.

The second criterion is based on the $p(O)/p(F)$ ratio between the probability that the fitted model assigns to an observation O and the probability of the forecast value F, that is

$$odds_{smooth} = -\frac{1}{N'} \log \prod_{n=1}^{N'} \frac{Pr_{smooth}(i_s^{(n)})}{Pr_{smooth}(i_{smooth}^{(n)})} ,$$

where $i_{smooth}^{(n)}$ is the estimate of the intensity at the n-th site provided by the mode of the *smoothed* binomial distribution. The idea behind this measure is based on a consideration of how much is gained from having predicted F when O occurs and is borrowed from the concept of deviance [11]. Of course the gain is maximum when we have predicted what really occurs.

The third and last criterion is based on the absolute discrepancy between observed and estimated intensities at site

$$diff_{smooth} = 1/N' \sum_{n=1}^{N'} \left| i_s^{(n)} - i_{smooth}^{(n)} \right| . \tag{7}$$

On the basis of the absolute discrepancy (7) comparisons have been performed with the best deterministic attenuation relationships proposed in the literature for the various test areas; for the complete results the reader is referred to the final report of the UPStrat-MAFA project [12].

3 The learning sets

The data set considered in the UPStrat-MAFA project [17] as a basis for the clustering procedure is composed of 298 macroseismic fields drawn from the most recent Italian macroseismic database, DBMI11 [5], that have at least 40 data points and correspond to the earthquakes of MCS $\geq V$ that occurred in Italy from 1500. The spatial locations of these events cover all the Italian territory. Table 1 shows the number N of the macroseismic fields which compose our data set for each value of the epicentral intensity. On the whole the data points are 43350.

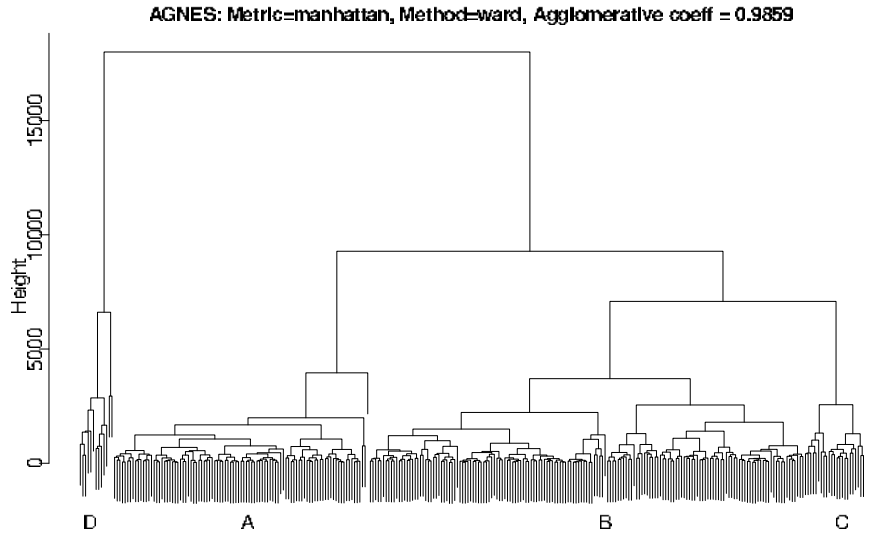


Fig. 1 Dendrogram obtained by applying the Ward method to the part of the DBMI11 Italian database used in the UPStrat-MAFA project. A, B, C and D denote the four classes we selected.

Table 1 Number of macroseismic fields analysed, subdivided by epicentral intensity.

I_0	V	V-VI	VI	VI-VII	VII	VII-VII	VIII	VIII-IX	IX	IX-X	X	X-XI	XI	tot.
N	28	42	55	17	40	25	26	7	22	6	18	3	9	298

The dendrogram produced by applying the Ward method, described in section 2.1, to this set of data is depicted in Figure 1. Background knowledge about the Italian sismicity, visual inspection of the dendrogram and the analysis of the silhouette plots (explained in the following) drove us to select the four classes A, B, C and D highlighted in Figure 1, of size 97, 165, 23 and 13, respectively.

The Agglomerative Coefficient (AC) whose value is shown in Figure 1 concerns the strength of the clustering structure obtained. For each macroseismic field i, $i = 1,\ldots,298$, we denote by $d(i)$ its dissimilarity to the first cluster it is merged with, divided by the dissimilarity of the merger in the final step of the process; AC is the average of all $(1-d(i))$. Its possible value ranges between 0 and 1, and the higher the AC value, the clearer the clustering structure. As it is seen, in our case the value is very high, signifying the sharpness of the clustering structure.

As for the silhouette plots that we used to support the choice of the classes, the silhouette value $s(i)$ of each object, in our case each macroseismic field, is computed as follows. Let A be the cluster to which object i belongs and $a(i)$ the average dissimilarity of i to all other objects in A. Then let C be any cluster different from A and $d(i,C)$ the average dissimilarity of i to all objects of C. After identifying the cluster B such that

$$b(i) = d(i,B) = min_{C \neq A} d(i,C),$$

$s(i)$ is defined as

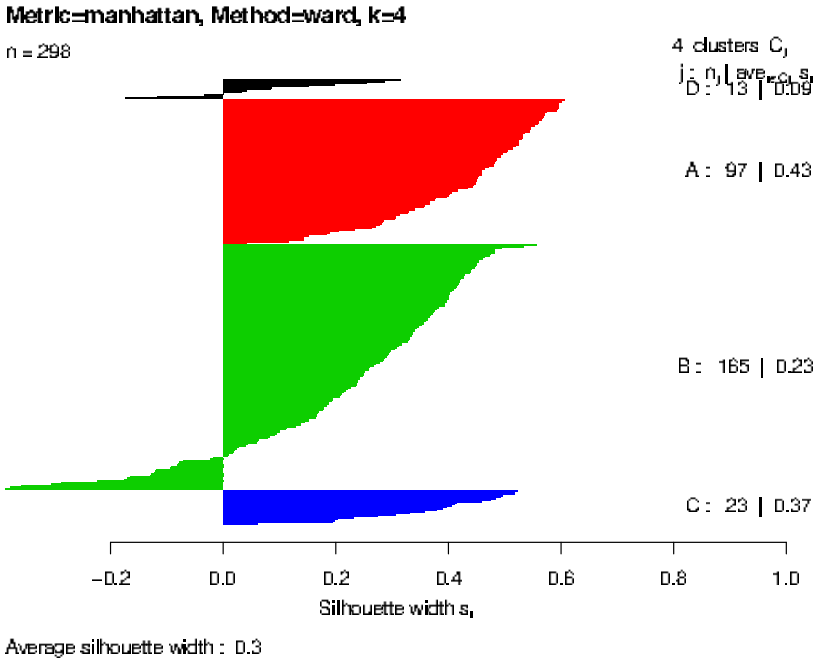


Fig. 2 Silhouette plot corresponding to the four classes A, B, C, and D.

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}.$$

B is called neighbour of object i and the second-best cluster for object i. The value $s(i)$ always lies between -1 and +1, and it is clear that the more $s(i)$ approaches 1, the better the object i is classified, whereas if $s(i) \approx -1$, object i is badly classified. The overall average silhouette width is then defined as the average of the $s(i)$ over all objects i in the data set. The silhouette plot is the graphic representation of this quality index. It is a plot showing the silhouettes of all clusters next to each other, where the silhouette of a cluster is a plot of the $s(i)$, ranked in decreasing order, of all its objects i. Figure 2 provides the silhouette plot corresponding to the four classes A, B, C and D we selected. The average silhouette width is 0.3, which is the best value compared to those obtained by considering 3 and 5 classes.

The software used to perform the clustering procedure and to derive the silhouette plots is the free software R [10].

A summary of the data set analysed is given in Table 2. We remark that in the probability model proposed the intensity is considered as an integer variable, hence the intensities I_0 assessed by half degrees have to be approximated. We have chosen to approximate them by below, that is, for instance, the earthquakes with I_0 = V-VI are grouped together with those of I_0 = V, and so on. We stress that this is the unique

Table 2 Summary of the 298 macroseismic fields analysed, subdivided by epicentral intensity I_0 and attenuation classes A, B, C, D.

	class				
I_0	A	B	C	D	
V	7	20	1	-	28
V-VI	6	35	1	-	42
VI	15	29	9	2	55
VI-VII	1	9	3	4	17
VII	15	20	3	2	40
VII-VIII	10	11	3	1	25
VIII	7	16	2	1	26
VIII-IX	1	5	-	1	7
IX	14	5	1	2	22
IX-X	4	2	-	-	6
X	9	9	-	-	18
X-XI	3	-	-	-	3
XI	5	4	-	-	9
	97	165	23	13	298

approximation we have done; throughout the algorithm the intensity values are used as they are in the data set, respecting their uncertainty.

Figure 3 shows the epicentre distribution of the 298 earthquakes which constitute the four attenuation classes A, B, C, and D among which we chose the learning set for each of the case studies, and those available, at present, for potential future studies. The classes have attenuation trends decreasing in steepness, that is class A contains the macroseismic fields with the steepest attenuation trend, whereas class D gathers macroseismic fields with the flattest attenuation trend. This is exemplified in Figure 4, that also shows the global internal homogeneity of the different classes in terms of attenuation, despite the great variety of situations.

As exemplified in section 4.2, in each case study the choice of a suitable learning set is carried out by comparing the attenuation trends of the earthquakes under study with those of the four classes, and by choosing the most similar one.

Figure 5 shows, conditioned on $I_0 = VII$, the prior and posterior estimates of the parameters p_j of the binomial distribution of the intensity at site I_s obtained for each of the classes A, B, C, and D as explained in section 2.2.2. The estimates were obtained under the assumption of isotropic decay and the width of the distance bins was assumed equal to 10 km. The red curves depicted in Figure 5 represent the trends of the estimates obtained by applying the smoothing inverse power function $g(d) = [c_1/(c_1 + d)]^{c_2}$ with the values given in Table 3. We emphasize that these trends should not be confused with deterministic attenuation laws; in fact, they are substantially linked, by Eq. 3, to the variation of the probability of null decay with respect to the distance from the epicentre. For I_0 = V the estimates are shown just for classes A, B, and C since class D does not include earthquakes with I_0 = V. Similarly, for I_0 = X and I_0 = XI the estimates are shown just for classes A and B since classes C and D do not include earthquakes with I_0 = X and I_0 = XI.

The 298 macroseismic fields taken into account for building the learning sets have also been used to costruct the classification tree [3] depicted in Figure 6. Very briefly, the tree has to be interpreted as follows. In the root node we have all the data

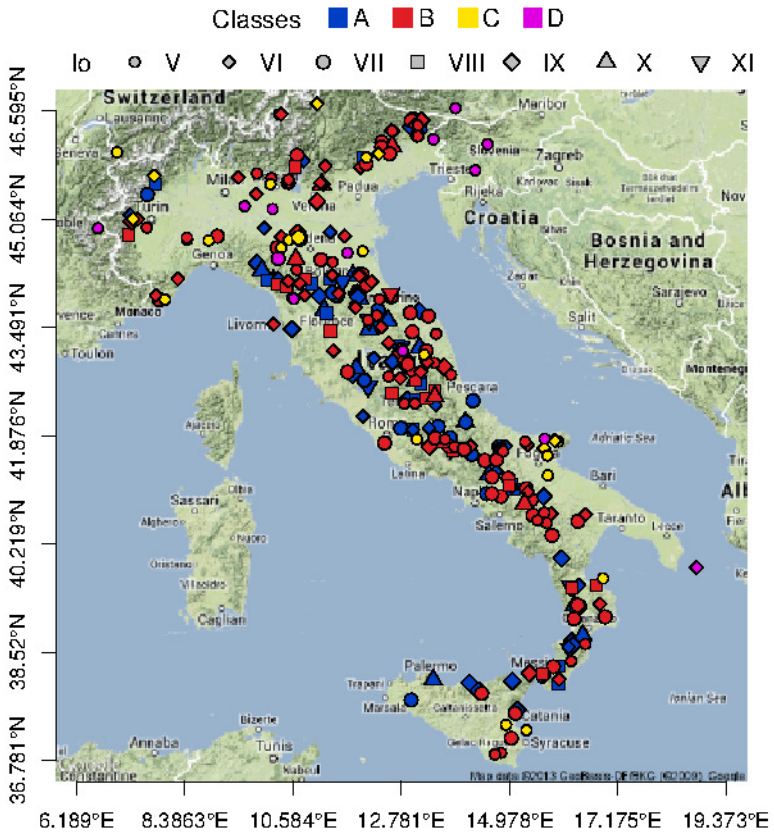


Fig. 3 Spatial location of the 298 earthquakes subdivided into the four attenuation classes A, B, C, and D.

set, that is the 298 macroseismic fields, each characterized by the attenuation class it belongs to. Each split is driven by a condition concerning one of the summaries of the spatial distribution of the intensity decay used to describe the macroseismic fields in the clustering procedure (see section 2.1). For example, the condition $mean[2] < 29.56$ ruling the very first split concerns the mean of the distances of the sites with $\Delta I = 2$, while the condition $mean[4] < 82.72$ ruling the split at the very left of the tree concerns the mean of the distances of the sites with $\Delta I = 4$, and so on.

The macroseismic fields for which the condition is satisfied are directed to the left, those for which the condition is not satisfied are directed to the right. As it is seen, each node of the tree, including the terminal nodes (leaves), is characterized by its composition in terms of macroseismic fields (in order, classes D, A, B, C, as in Figure 1) and by the class assigned to it according to the composition itself. To exemplify, let us take again the leaf at the very left of the tree: it contains 80 macroseismic fields of class A, 4 of class B and none of the two other classes, and class A is the class associated with the leaf. At the opposite side, on the extreme right of the tree,

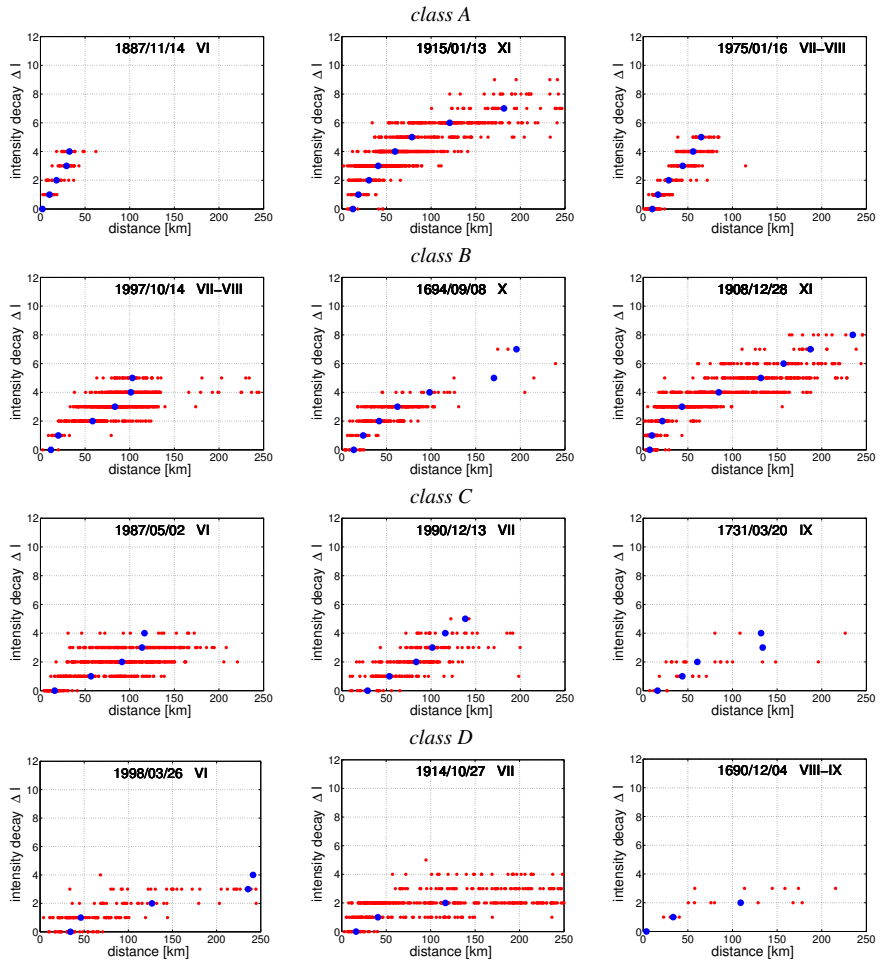


Fig. 4 Intensity decay versus epicentral distance for some macroseismic fields of class A, B, C, and D. Each blue dot signifies the median of the corresponding set of epicentral distances (red dots) for sites with the same ΔI. The date and the epicentral intensity of the relative earthquakes are shown too.

we have a leaf containing 116 macroseimic fields of class B, 1 of class A and C and none of class D. The class associated with the leaf in this case is class B. The idea underlying the construction of a classification tree is that of having leaves as uniform as possible in terms of the class they belong to. In our case the result is very satisfactory since the percentage of macroseismic fields directed to leaves associated to classes different from the ones they belong to is very low, as it can be easily verified by looking at the leaves themselves.

Generally speaking, given a set of known classes, a classifier is used to assign one of these classes to an object whose class is unknown. A classification tree achieves this aim by entering the object into the tree and directing it to one of the leaves according to the conditions on the splits. In this study the classification tree we constructed has

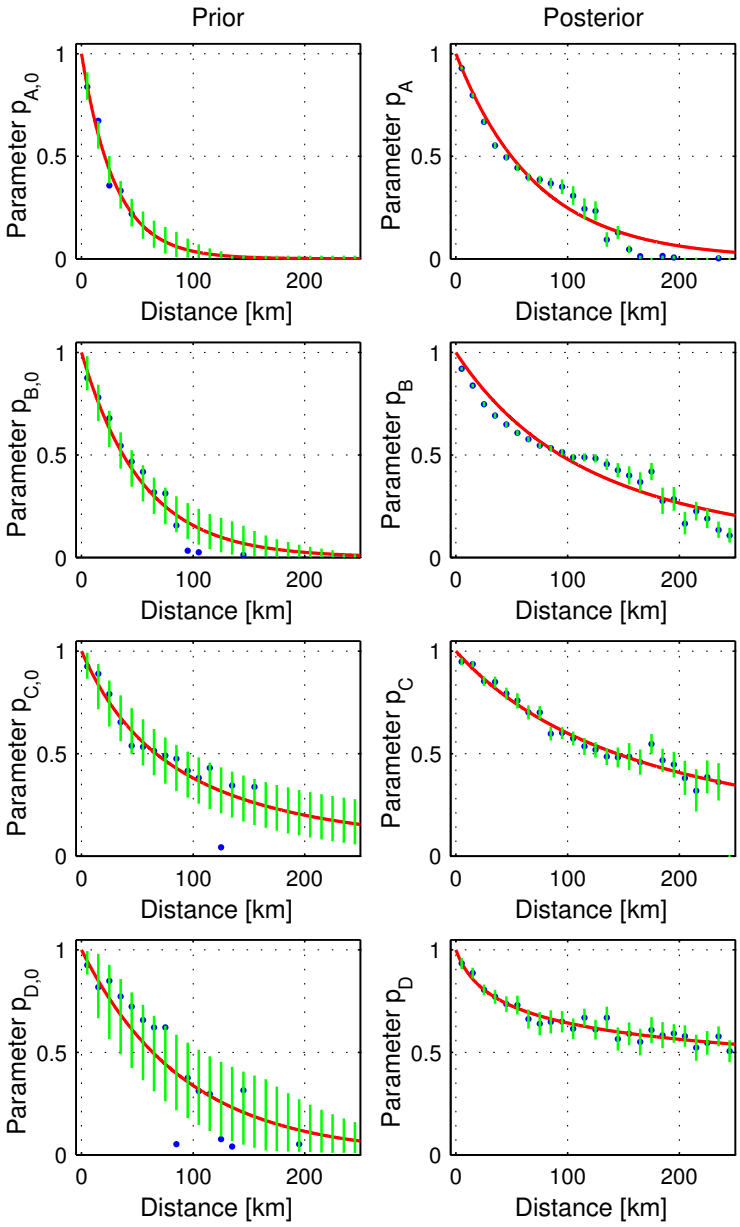


Fig. 5 Prior and posterior estimates of parameters p_j (blue dots). The red curves indicate the smoothing inverse power functions, and the green bars the 80% confidence intervals of the p_j. Classes A, B, C and D; I_0 = VII. We recall that parameters p_j and probability of null decay are linked by Eq. 3

Table 3 Parameters of the inverse power functions smoothing the values of the $p_{j,0}$ prior and p_j posterior parameters, respectively, for the attenuation classes A, B, C and D.

I_0	Class	$c_{1,0}$	$c_{2,0}$	c_1	c_2
V	A	8391.50	290.52	13587.91	244.38
	B	10169.59	186.29	433.66	3.66
	C	177.84	2.18	65.93	0.72
	D	-	-	-	-
VI	A	7026.81	232.13	12776.89	216.50
	B	15909.20	316.81	36125.39	261.90
	C	15703.24	192.29	165.87	1.08
	D	22464.64	353.85	208.48	0.87
VII	A	4786.27	160.30	13679.45	191.53
	B	18650.52	346.79	361.63	3.01
	C	159.88	1.99	261.29	1.58
	D	28026.65	304.56	13.99	0.21
VIII	A	8425.44	273.20	13515.28	189.91
	B	17759.64	331.79	209.08	1.70
	C	1187.72	11.84	56.81	0.59
	D	25159.44	300.90	26.57	0.30
IX	A	8908.18	285.60	13973.95	162.15
	B	16013.49	291.74	123.10	1.10
	C	2905.35	29.91	201.73	1.72
	D	20684.05	219.65	83.61	0.75
X	A	6354.20	226.24	31832.60	329.98
	B	18600.38	331.09	93.92	0.84
	C	-	-	-	-
	D	-	-	-	-
XI	A	10650.37	357.73	33942.32	307.74
	B	17819.76	333.78	20376.75	115.65
	C	-	-	-	-
	D	-	-	-	-

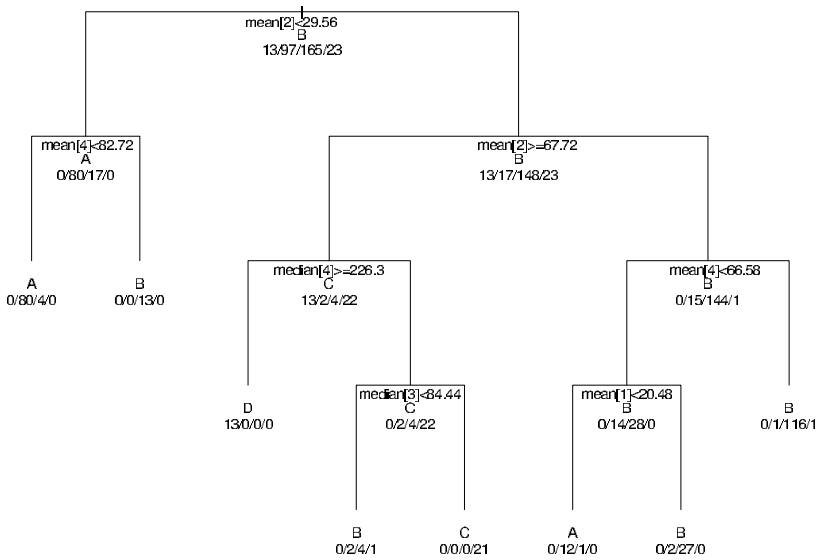


Fig. 6 Classification tree built with the 298 macroseismic fields.

been used to help choosing the attenuation class of the macroseismic fields considered in the different case studies. In studies on seismic hazard assessment in Italy, the classification tree has been used for assigning the attenuation class to those macroseismic fields of the DBMI11 database that did not enter into the clustering procedure since they provide little information.

The software used to obtain the classification tree is the free software R [10].

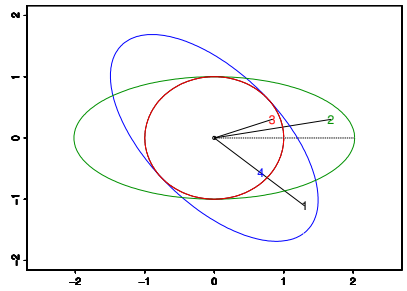
4 Solutions to some specific problems

4.1 How to exploit information on anisotropy?

In Agostinelli and Rotondi [1], data depth functions have been applied to identify attenuation patterns in sets of macroseismic fields. In addition to the general circular decay trend corresponding to the assumption of isotropic attenuation, it has turned out that it can be appropriate to use an elliptical shape when we have information on the fault rupture that caused an earthquake, in particular on the direction and length of the rupture. Indeed, it can happen that more rapid decay is visibly recognizable along the direction perpendicular to that of the fault in the macroseismic field generated by a strong earthquake with long fault rupture. It is however clear that it is not possible to collect a large number of fields with the same fault characteristics on which to base parameter estimation. The solution we have found consists in a plane transformation that turns the ellipse of major axis equal to the fault rupture into the circle of radius equal to the width of the bins. This allows to exploit in assigning prior distributions the information collected in the isotropic case. As for the assumption of anisotropic decay, we just consider the hypothesis of elliptic isoseismal lines, but, in principle, other curves could be considered, as long as it is possible to find a plane transformation that brings back to the circular pattern characterizing the isotropic decay.

In the ellipse case we proceed in the following way [2]: let us consider an ellipse (blue line in Figure 7) rotated by an angle ψ with respect to the positive semi-axis, with center at origin and semimajor a and semiminor b axes, respectively (in Figure 7 $\psi = -0.785$ rad, azimuth = 2.356 rad); first, we rotate it counterclockwise by the angle ψ, overlapping it to the (green) ellipse in canonical position, then we scale this ellipse to the (red) circle with radius b, and finally rotate this circle clockwise to the original position. In Figure 7 the sequence of movements is indicated by the points 1, 2, 3, 4; at the side of the figure there are the equations linking the coordinates of the moving points.

Once the plane transformation is performed, the probability model can be applied, and the parameters estimated, as in the isotropic case (section 2.2). The estimated probability distribution of the intensity I_s that will be felt at a site is then associated to the original position of that site. Since the anisotropic effect decreases with the distance from the epicentre, we draw the subsequent elliptical bins by increasing both the axes of a fixed equal segment Δr, so that the eccentricity of the increasing ellipses tend to 0. An example of the results obtained is given in Figure 8 which refers to the 1865/07/19 earthquake that hit Fondo Macchia, a place on the eastern flank of Mt Etna volcano. The top panels show the observed (left) and estimated (right)



$(1) \longrightarrow (2)$

$x_2 = \cos(-\psi)\ x_1 - \sin(-\psi)\ y_1$

$y_2 = \sin(-\psi)\ x_1 + \cos(-\psi)\ y_1$

$(2) \longrightarrow (3)$

$x_3 = x_2 \times b/a \qquad y_3 = y_2$

given $\theta = \arctan(y_2/x_2)$ and $\theta' = \arctan(y_3/x_3)$, we set $\phi = -\theta' + \theta + \psi$

$(3) \longrightarrow (4)$

$x_4 = \cos(\phi)\ x_3 - \sin(\phi)\ y_3$

$y_4 = \sin(\phi)\ x_3 + \cos(\phi)\ y_3$

Fig. 7 Graphical representation and equations of the transformation of the (blue) ellipse into the (red) circle.

Table 4 Number of macroseismic fields analysed for the European test areas involved in UPStrat-MAFA project, subdivided by epicentral intensity.

		Epicentral intensity								
	type of MFs	V	VI	VII	VIII	IX	X	XI	XII	$\geq$ V
Italy	isotropic	70	72	65	33	28	21	9	-	298
Mainland Portugal	an/isotropic	3	4	1	-	1	1	-	-	10
Offshore Portugal	an/isotropic	1	-	-	1	4	1	1	1	9
Azores Islands	isotropic	-	4	6	3	1	-	1	-	15
SE Spain	isotropic	5	14	8	2	1	-	-	-	30
Mt Etna	isotropic	-	30	14	8	2	-	-	-	54
	anisotropic	-	-	7	9	1	-	-	-	17
Iceland	isotropic	-	3	1	1	2	1	-	-	8
	anisotropic	-	-	-	-	2	1	-	-	3

macroseismic field, whereas the botton panels depict the values of intensity I_s that may be exceeded with 25% (left) and 75% (right) probability respectively, according to the estimated smoothed binomial distribution (5).

4.2 How to tune learning sets with data sets?

In the project UPStrat-MAFA we have analysed the macroseismic attenuation of European countries: Portugal, Spain, Iceland, which do not have so large database of macroseismic fields as to allow a reliable estimation of the probability distribution of I_s. Table 4 shows the number N of macroseismic fields examined for the European test areas involved in the UPStrat-MAFA project, subdivided by epicentral intensity.

A way to proceed has been to compare the attenuation trend in these countries with that of the four classes A, B, C, and D of macroseismic fields identified in the Italian database DBMI11 [5] and used as learning sets (section 3). The comparison

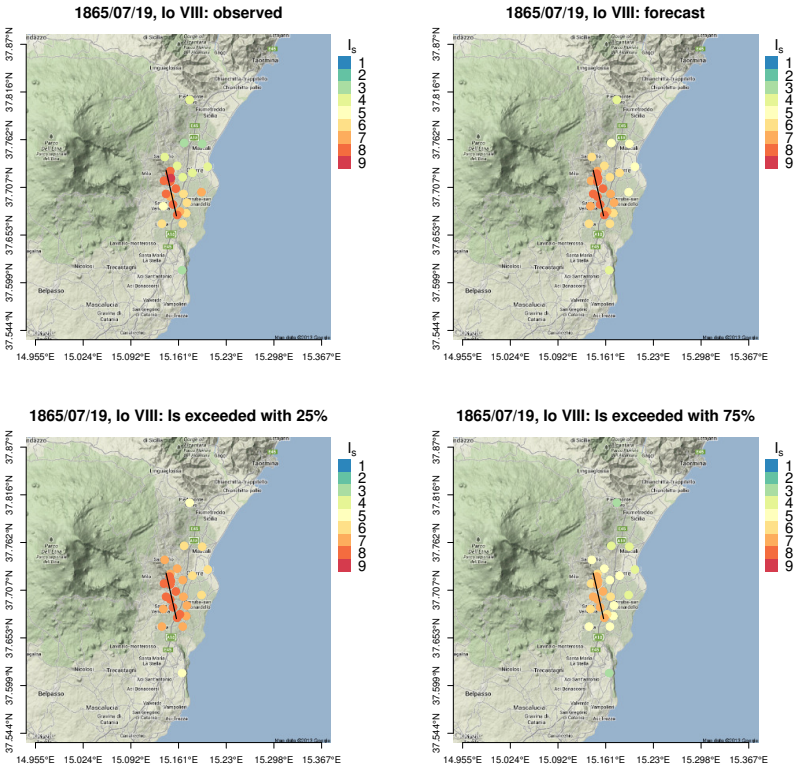


Fig. 8 *Mt Etna, Italy*: Anisotropic case. (*top*): Observed (left panel) and estimated (right panel) macroseismic field of the 1865/07/19 earthquake of I_0 = VIII; (*bottom*): Intensities at site that may be exceeded with probability 25% (left panel) and 75% (right panel).

was mainly carried out graphically by overlapping summaries (mean, median, 3^{rd} quartile) of the spatial distribution of the seismic decay in the classes A, B, C and D with those of the country under examination. If necessary, we applied a scaling factor k to the data from the European countries to improve matching; so, e.g., as for Icelandic earthquakes of $I_0 \geq IX$, we noticed that what is felt at a distance d in the case of Italian earthquakes in class A is similar to what is felt at distance $d/2$ in Iceland. This means that the scaling factor $k = 2$ has to be adopted to make usable the results obtained for class A as prior information on the decay in Iceland; therefore, the inverse power function $g(d) = [(c_{1A}/k)/(c_{1A}/k+d)]^{c_{2A}}$ is used to obtain the values $p_{j,0}$ from which to deduce the hyperparameters of the prior distribution of the parameter p_j (section 2.2.2), where c_{1A} and c_{2A} are the parameters estimated through the method of least squares from the macroseismic fields of the earthquakes in class A (see Table 3). The correspondence between scaled ($k = 2$) Icelandic earthquakes of $I_0 \geq IX$ and class A was also supported by the result obtained by examining the Icelandic macroseismic fields with the classification tree built on the basis of the 298 Italian fields (see section 3). Figure 9 compares summaries of the spatial distribution

of the seismic decay in class A with those of the Icelandic earthquakes of $I_0 \geq IX$ without and with the scaling.

Table 5 Input parameters adapting the model to the observations from European countries: attenuation class used as learning set, scaling factor, bin size, width of the first isoseismal line, length of the minor axis of the first ellipse in the anisotropic case.

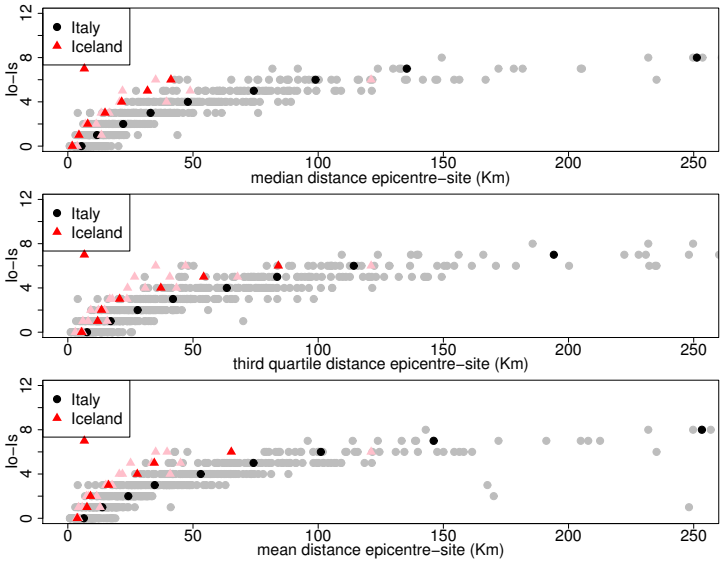
			Isotropic input		Anisotropic input
	attenuation class	scaling factor	bin size	first isoseismal	minor axis of the ellipse
Italy	A, B, C, D	1	10 Km	10 Km	-
Mainland Portugal	B	1/2	20 Km	20 Km	fault rupture/3
Offshore Portugal	B	1/3	30 Km	30 Km	fault rupture/3
Azores Islands	A	2	5 Km	5 Km	-
SE Spain	A	1	10 Km	10 Km	-
Mt Etna	B	10	1 Km	1 Km	fault rupture/5
Iceland	A, Io$\geq$ IX	2	5 Km	5 Km	fault rupture/3
	D, Io $\leq$ VIII	1	10 Km	10 Km	-

Other critical points concern how to assign the size of the first circle, the width of the bins, and the length of the minor axis in the anisotropic case. These values have been typically determined through an exploratory analysis of the epicentral distances of the sites at which $\Delta I < 1$ in the isotropic case, and through an exploratory analysis of the distances of the same sites from the fault rupture in the anisotropic case. As for Iceland, in the light of these values we have set the radius of the first circle and the width of the subsequent bins equal to 20 km in the isotropic case, and the minor axis in the anisotropic case equal to the third of the faul rupture (major axis). Table 5 summaries all of the input parameters which adapt the model to the specific conditions of the countries considered in the project.

4.3 How to integrate the fields of offshore earthquakes?

Offshore earthquakes are characterized by the lack of intensity data points at short distances from the epicentre or from the fault rupture when it is known. In our approach this means that we were not able to compute some summaries (mean, median, 3^{rd} quartile) of the spatial distribution of the seismic decay that are the starting point of our modelling. For instance, the use of the classification tree in order to select the most suitable learning set among the four classes A, B, C, and D, requires that some specific summaries are known; indeed, on the basis of the learning set we used to build the classification tree, the summaries required for using it as a classifier are 5,

Class A - Original distances for $I_0 \geq IX$.



Class A - Distances $\times$ scaling factor $I_0 \geq IX$

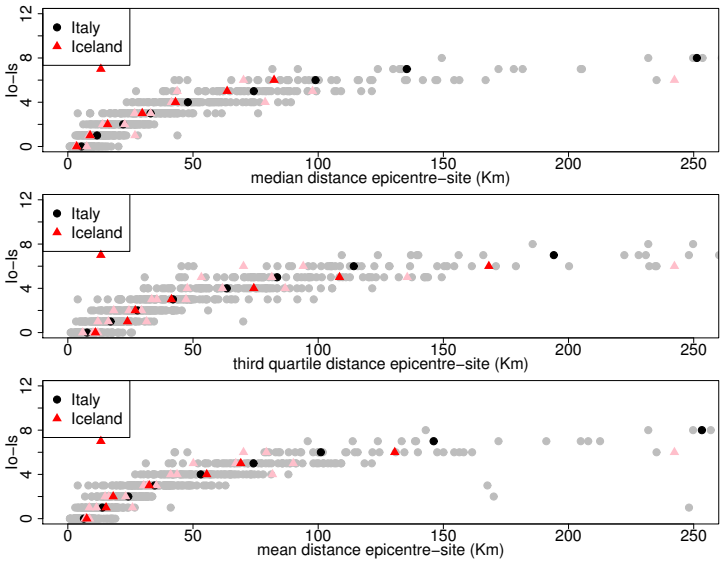


Fig. 9 Plots of summaries (median, 3^{rd} quartile and mean) of the spatial distribution of the intensity decay in class A (grey dots) and for Icelandic earthquakes having $I_0 \geq IX$ (pink triangles), without (above) and with (below) the scaling factor $k = 2$. Black dots and red triangles denote median values.

and precisely the mean of the distances at which $\Delta I = 1$ (mean[1]), $\Delta I = 2$ (mean[2]) and $\Delta I = 4$ (mean[4]), and the median of the distances at which $\Delta I = 3$ (median[3]) and $\Delta I = 4$ (median[4]); we refer to section 3, in particular to Figure 6. When missing, they were derived by fitting the regression model $\Delta I = 3\log(d/h) + b(d-h)$, where d denotes distance and b and h are parameters of the model, estimated by the least squares method. Figure 10 shows the values of the 5 summaries that were provided in input to the tree for some offshore earthquakes of Portugal.

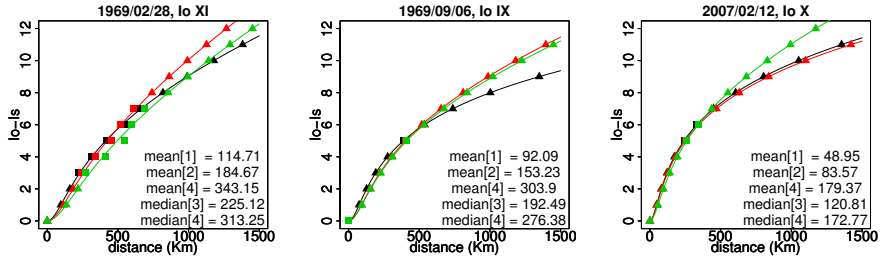


Fig. 10 *Offshore Portugal*: Examples of plots of the mean (red), median (black) and 3^{rd} quartile (green) of the epicentral distances where the same ΔI is recorded; d in the place of the missing summaries is estimated according to the model $\Delta I = 3\log(d/h) + b(d-h)$. Triangles denote observed values, squares estimated values. The values of the summaries needed by the classification tree are shown too.

5 Conclusions

The UPStrat-MAFA project has offered the opportunity to make, for some European seismic regions, the whole path from the collection of data to the evaluation of the seismic risk impact on buildings and network systems. All of the steps have been done in consistent way with the decisions of adopting the macroseismic intensity as proxy measure of the size of an earthquake and of treating the various sources of uncertainty through probabilistic methods. In this picture we have dealt with the problem of attenuation modelling. A key point of our approach has been the construction of learning sets from which to draw information to enhance insufficient data sets and to make reliable the parameter estimates. Then we have matched the data set of each test area with the most similar learning set from the decay point of view so that, according to the Bayesian paradigm, the prior distributions of the model parameters are borrowed from the suitably modified posterior distributions obtained through the learning set. Moreover, we have shown how to include in the same modelling also the macroseismic fields of offshore earthquakes by completing the missing summaries of the sets of distances from the epicentre to the sites where the same decay is recorded. Finally, when information on the finite source is available, we have illustrated how to return to a similar environment in which the decay is isotropic by a transformation of the plane, to exploit the information collected in estimation of other data sets.

We refer the reader to the final report of the project [12] for more detailed information and all the figures.

Acknowledgements The authors thank the two reviewers for very useful and constructive suggestions and comments, which appreciably improved the paper. This study was co-financed by the EU - Civil Protection Financial Instrument, in the framework the European project Urban disaster Prevention Strategies using MAcroseismic fields and FAult sources (Acronym: UPStrat-MAFA, Grant Agreement N. 23031/2011/613486/SUB/A5)

References

1. C Agostinelli and R Rotondi. Analysis of macroseismic fields using statistical data depth functions: Considerations leading to attenuation probabilistic modelling. *Bulletin of Earthquake Engineering*, this volume, 2015.
2. R Azzaro, S D'Amico, R Rotondi, T Tuvè, and G Zonno. Forecasting seismic scenarios on Etna volcano (Italy) through probabilistic intensity attenuation models: A Bayesian approach. *Journal of Volcanology and Geothermal Research*, 251:149–157, 2013.
3. L Breiman, J H Friedman, R A Olshen, and C J Stone. *Classification and regression trees*. Chapman Hall, New York, 1993.
4. L Kaufman and P J Rousseeuw. *Finding groups in data*. Wiley, New York, 1990.
5. M Locati, R Camassi, and M (eds.) Stucchi. DBMI11, the 2011 version of the Italian Macroseismic Database. `http://emidius.mi.ingv.it/DBMI11/`. Milano, Bologna, 2011.
6. L Magri, M Mucciarelli, and D Albarello. Estimates of site seismicity rates using ill-defined macroseismic data. *Pure and Applied Geophysics*, 143(4):617–632, 1994.
7. C Meletti, E Patacca, and P Scandone. Construction of a seismotectonic model: the case of Italy. *Pageoph.*, 157:11–35, 2000. Zonation ZS.4 available from `http://emidius.mi.ingv.it/GNDT/P511/home.html`.
8. F Meroni, V Petrini, R Rotondi, and G Zonno. Expected damage for alternative seismic hazard evaluations. In *Proceedings of 4th ICSZ*, volume 2, pages 801–818, Standford, California, 1991.
9. B Pizzo and V Fabietti. Environmental risk prevention, post-seismic interventions and the reconstruction of the public space as a planning challenge. *Italian Journal of Planning Practice*, 3(1):1–8, 2013.
10. R Development Core Team. R: A language and environment for statistical computing. `http://www.r-project.org`. Vienna, Austria, 2008.
11. T R C Read and N A C Cressie. *Goodness-of-fit statistics for discrete multivariate data*. Springer-Verlag, New York, 1988.
12. R Rotondi, C Brambilla, E Varini, and G Zonno. Task B - Probabilistic analysis of macroseismic data for forecast damage scenarios. *OA Earth-prints Repository*, 2014. available from `http://hdl.handle.net/2122/9143`.
13. R Rotondi and G Zonno. Bayesian analysis of a probability distribution for local intensity attenuation. *Annals of Geophysics*, 47(5):1521–1540, 2004.
14. P J Rousseuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.
15. M Stucchi, R Camassi, A Rovida, M Locati, E Ercolani, C Meletti, P Migliavacca, F Bernardini, and R Azzaro. DBMI04, il database delle osservazioni macrosismiche dei terremoti italiani utilizzate per la compilazione del catalogo parametrico CPTI04. *Quaderni di Geofisica*, 49:1–38, 2007. available from `http://emidius.mi.ingv.it/DBMI04/`.
16. T M Tsapanos, O Galanis, G Koravos, and R M W Musson. A method for Bayesian estimation of the probability of local intensity for some cities in Japan. *Annals of Geophysics*, 45(5):657–671, 2002.
17. Urban Disaster Prevention Strategies using MAcroseismic Fields and FAult Sources (UPStrat-MAFA) - EU Project. Num. 23031/2011/613486/SUB/A5), DG ECHO Unit A5. `http://upstrat-mafa.ov.ingv.it/UPStrat/`, 2012.
18. R L Winkler. Scoring rules and the evaluation of probabilities. *Test*, 5(1):1–60, 1996.
19. G Zonno, R Rotondi, and C Brambilla. Mining macroseismic fields to estimate the probability distribution of the intensity at site. *Bulletin of Seismological Society of America*, 99(5):2876–2892, 2009.