

Seismic vulnerability assessment of urban environments in moderate-to-low seismic hazard regions using association rule learning and support vector machine methods

Ismaël Riedel · Philippe Guéguen · Mauro Dalla Mura ·
Erwan Pathier · Thomas Leduc · Jocelyn Chanusot

Received: 2 June 2014 / Accepted: 20 November 2014 / Published online: 30 November 2014
© Springer Science+Business Media Dordrecht 2014

Abstract The estimation of the seismic vulnerability of buildings at an urban scale, a crucial element in any risk assessment, is an expensive, time-consuming, and complicated task, especially in moderate-to-low seismic hazard regions, where the mobilization of resources for the seismic evaluation is reduced, even if the hazard is not negligible. In this paper, we propose a way to perform a quick estimation using convenient, reliable building data that are readily available regionally instead of the information usually required by traditional methods. Using a dataset of existing buildings in Grenoble (France) with an EMS98 vulnerability classification and by means of two different data mining techniques—association rule learning and support vector machine—we developed seismic vulnerability proxies. These were applied to the whole France using basic information from national databases (census information) and data derived from the processing of satellite images and aerial photographs to produce a nationwide vulnerability map. This macroscale method to assess vulnerability is easily applicable in case of a paucity of information regarding the structural characteristics and constructional details of the building stock. The approach was validated with data acquired for the city of Nice, by comparison with the RiskUE method. Finally, damage estimations were compared with historic earthquakes that caused moderate-to-strong damage in France. We show that due to the evolution of vulnerability in cities, the number of seriously damaged buildings can be expected to double or triple if these historic earthquakes were to occur today.

I. Riedel (✉) · P. Guéguen · E. Pathier
ISTerre, CNRS/IFSTTAR, Université Joseph Fourier Grenoble I, BP 53, 38041 Grenoble Cedex 9,
France
e-mail: ismael.riedel@ujf-grenoble.fr

M. Dalla Mura · J. Chanusot
Grenoble Images Parole Signal Automatique (GIPSA-LAB), Grenoble, France

T. Leduc
CERMA/CNRS, Université Nantes Angers Le Mans, Nantes, France

Keywords Seismic vulnerability · Moderate hazard · Existing building · Data mining · Support vector machine · Europe

1 Introduction

The extensive damage observed after the latest moderate-to-strong earthquakes together with population growth and the urbanization of megacities has considerably increased awareness regarding natural disasters over recent decades (Jackson 2006). There is also an increasing demand for detailed seismic risk analysis, to furnish adequate information for the insurance and reinsurance companies (Spence et al. 2008). A complete seismic risk assessment requires not only the estimation of the seismic hazard, but also the representation of the quality of existent buildings and their expected response based on the definition of their vulnerability. Even though some regions are considered to be of moderate hazard, they are not free of seismic risk, and particularly not if the vulnerability of their cities is high (Dunand and Gueguen 2012). Major earthquakes on the scale of France, for example, have caused genuine catastrophes during the last centuries. Reducing this risk has become a priority for local authorities in order to ensure the well-being and safety of local populations as well as for economic and social security. One of the areas contributing to the reduction in earthquake fatalities and losses, besides the improvement of technical norms and the reinforcement of existing buildings, is the anticipation and simulation of earthquake effects for crisis management. This simulation requires a representation of the structures' capacity to withstand the seismic ground motion: this is the objective of seismic vulnerability assessments. Coupled with real-time seismic ground motion estimates (e.g., Wald et al. 1999; Worden et al. 2010), macroscale vulnerability data are crucial for the early assessment of damage.

Even though earthquake codes can always be improved, the low rate of renovation of building stocks in cities makes existing buildings (mostly designed before the application of earthquake design rules) the center of physical vulnerability. Over the last two decades, many empirical methods have been published to assess the seismic vulnerability of buildings at a large scale, most of them calibrated using post-event damage information (e.g., GNDT 1993; Hazus 1997; Spence and Lebrun 2006) or directly derived from a macroseismic intensity scale (Lagomarsino and Giovinazzi 2006). Hybrid methods (e.g., Kappos et al. 2006) or experimental methods (Michel et al. 2012) have also been proposed as a complement of empirical methods. They estimate the probability of reaching a certain level of damage for a given class of buildings and a given seismic demand. An extensive description of these methods can be found in Gueguen (2013). Some challenges and difficulties these methods have to face are (1) the variability of the response of existing buildings to seismic loads, (2) misunderstanding of the seismic behavior of old buildings as well as inadequate information concerning the quality of construction materials, and (3) the lack of observation data to adjust empirical methods to the highest damage grade. These issues introduce significant epistemic uncertainty into seismic vulnerability assessment (Spence et al. 2008) and therefore into seismic risk analysis. These difficulties are even more critical in moderate-to-low seismic hazard regions, where the mobilization of resources for seismic evaluation is rather limited.

For example, France is considered as a moderate seismic hazard country. However, a major historic earthquake hit France in the twentieth century with an estimated magnitude

of more than 6 and major effects in the (at that time rural) region of Aix-en-Provence (southeastern France), causing 42 fatalities, many more injuries, and severe economic losses. Other important events make up the seismic history of metropolitan France (Bâle earthquake 1356, Chandeleur earthquake 1428, Ligure earthquake 1887). More recently, Ossau-Arudy 1980 ($M_L = 5.1$) and Annecy 1996 ($M_L = 4.8$) earthquakes caused estimated losses of 4 million Euro (Environment ministry—MEDD 1982) and 50 million Euro, respectively (AFPS, French Paraseismic Association 1996), even at these low magnitudes, with the same order of damage/magnitude ratio as in other moderate seismic regions (e.g., Pierre and Montagne 2004).

In this context, vulnerability assessment studies have been conducted in France, focused on large exposed cities and applying traditional empirical methods. However, the application of these methods requires so much information that the evaluation struggles to find sufficient political motivation and financial resources to complete the seismic inventory of buildings. For example, the RiskUE project (Spence and Lebrun 2006) aimed to propose a seismic vulnerability assessment method for Europe, but due to its complexity, no city in France has ever been studied using this method (except for the city of Nice, which was a test site for the RiskUE project). Consequently, the structural characteristics required for the seismic vulnerability assessment of existing buildings are not available for all exposed urban areas of the country. On the other hand, seismic exposure is higher than in the past, and a repetition of historic earthquakes may result in more casualties and economic losses (Jackson 2006).

To overcome the lack of building information at the macroscale, we propose in this paper to assess vulnerability not considering the information required for a conventional analysis, but the sole information already available in a region or country (Fig. 1). Two different data mining methods, association rule learning (ARL) (Agrawal et al. 1993) and support vector machine (SVM) (Boser et al. 1992; Cortes and Vapnik 1995), are applied to define vulnerability proxies between the elementary characteristics of buildings and the vulnerability classes of the European Macroseismic Scale EMS98 (Grunthal and Levret 2001). This is a two-step procedure: the first step (the learning phase) consists in defining the proxy using a sample of buildings for which elementary structural characteristics (or attributes) and vulnerability classes are available. The second step (the application phase) is to apply the proxy to a target region for which vulnerability classes are not available, but elementary attributes are.

In the first part of this paper, the dataset used in the first step is presented: the test bed of the city of Grenoble, one of the most exposed cities in France, for which an extensive vulnerability analysis has been performed (Gueguen et al. 2007). The SVM and ARL methods are then briefly presented and applied to the Grenoble target site, deriving two vulnerability proxies for a Grenoble city-like environment. In the third part of this study, the derived vulnerability proxies are applied to the entire country and validated by comparison with the RiskUE method applied in Nice, a test site for the RiskUE project. Finally, the probable damage produced by historic earthquakes was computed, considering (equivalent) earthquake-era and present-day urbanization to simulate the evolution of vulnerability over time.

2 Grenoble test-bed area

As described in Riedel et al. (2014), a simplified empirical method based on the Italian GNDT was proposed and tested in Grenoble as part of the VULNERALP project (Gueguen

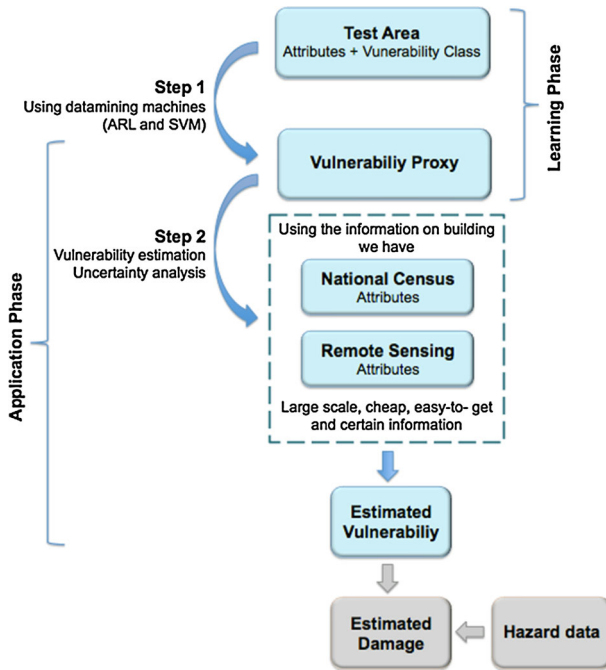


Fig. 1 Two-step process. During the learning phase, a vulnerability proxy is deduced from a test area for which a full seismic vulnerability evaluation is available. In the second step, this proxy is applied to a large region where only some attributes are available in order to estimate vulnerabilities. A final step combines the estimated vulnerability with hazard information to deduce probable damage

et al. 2007). Basic information was collected to assign elementary structural characteristics to existing buildings. The main pieces of information were date of construction ranked by period, number of floors ranked by category, roof shape (flat or slope), construction material, some qualitative description of plan and elevation irregularities, and building position in the block (corner, in-between, stand-alone, etc.). In addition to basic information, experts associated a type of building according to the EMS98 typology with the most likely vulnerability class (Grunthal and Levret 2001). The EMS98 scale was originally defined for macroseismic intensity assessment after an earthquake, but since buildings vulnerability is taken into account for defining intensity, vulnerability classification can be used to represent the seismic damage in a target region for a given intensity. Building vulnerability is established as belonging to a category of buildings (EMS98 typology) with six classes from A (most vulnerable) to F (least vulnerable). At the end of the process, the expert survey compiled the Grenoble building vulnerability database, in which 3,860 buildings were characterized according to their EMS98 vulnerability class and some essential attributes. These attributes are elementary since they are considered as reliable (no uncertainty in their definition) and can be obtained relatively easily on a large scale. For example, the information about the number of storeys and period of construction is available in the INSEE database (French national statistical institute, www.insee.fr), grouped by geo-localized cells called IRIS2000. Since their inception, the IRIS cells represent the national standard for geographical data distribution and must therefore meet geographic and demographic criteria. They also have contours that are stable over time and

easily identifiable. France has 50,100 IRIS units plus 700 in the overseas regions. Figure 2a shows the division of Grenoble and neighboring towns into IRIS units. Only residential dwellings are included in the INSEE database. Buildings per IRIS are described by attributes and grouped into categories.

Furthermore, during the NERA project (Network of European Research Infrastructures for Earthquake Risk Assessment and Mitigation—www.nera-eu.org), a building-by-building field survey was carried out in a small area of Grenoble (about 950×700 m) including all buildings within the surveyed area (Fig. 2b) (Spence et al. 2012). 560 residential buildings were characterized and classified according to EMS98. This subarea test was chosen because it shows a mix of building typologies representative of the Grenoble metropolitan area. Finally, remote sensing data are available in Grenoble, including a very high-resolution (VHR) orthorectified panchromatic image (airborne data, 25 cm resolution), a digital elevation model (DEM) (airborne acquisition, 1 m resolution in three dimensions), and building footprints from cadastral data. With this information, the Urbasis project (ANR-09-RISK-009) characterized the urban area based on building footprints and the surrounding open spaces within the NERA zone. Fifteen morphological indicators were computed according to Hamaina et al. (2012) for the characterization of urban fabric: length, width (W), elevation (H), area and volume of the building units, circularity according to Miller (ratio of footprint area to the area of circle having the same perimeter as the footprint), open space morphometry (proportion of the area occupied by open spaces), shared wall ratio (ratio between the length of perimeter walls shared with other buildings and the whole perimeter), average distance to nearest buildings (average distance between building footprints of neighboring cells), generalized ratio W/H , mean ratio of isovist area (area of space visible from a given point in space) divided by area of the enclosing circle, ground space index (ratio of a building's footprint area to the piece of land upon which it is built), floor space index (ratio between the building's volume and the area upon which it is built), among others. However, only a few were used for the vulnerability classification, as described in Sect. 3.1.

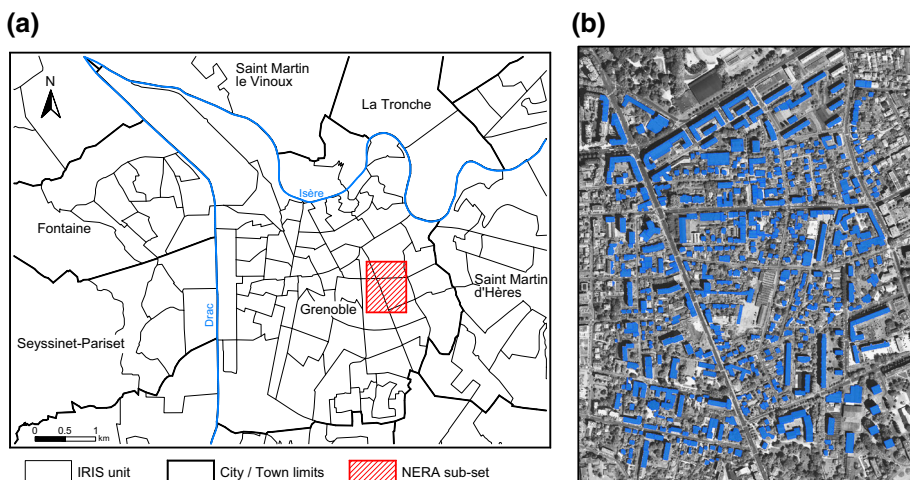


Fig. 2 **a** IRIS units in Grenoble. From INSEE database. **b** NERA study area in Grenoble, France. Building footprint layer superimposed on a VHR orthoimage. 560 buildings are characterized and classified according to EMS98 vulnerability classes

In this regard, remote sensing data and methods are increasingly used to assess seismic risk and particularly for the assessment of vulnerability (Geiss and Taubenböck 2012). Borsi et al. (2010) also illustrated how suitable processing of satellite images can contribute to the vulnerability evaluation of industrial areas, especially when no other sources of information are available.

3 Data mining methods

Data mining, a process at the intersection of computer science and statistics, attempts to identify patterns and establish relationships in large datasets. These techniques are used in many areas of research, including mathematics, cybernetics, genetics, and marketing. There are a number of different types of learning algorithms that can be used for the (exploratory) data analysis: decision trees, decision rules, association rules, neural networks, SVMs, Bayesian classifiers among others (Teukolsky et al. 2007).

3.1 Support vector machine (SVM)

The SVM is a state-of-the-art classification method (Boser et al. 1992; Cortes and Vapnik 1995). It is a supervised learning model with associated learning algorithms that analyze data and recognize patterns; it is used for classification and regression analysis (Teukolsky et al. 2007). A supervised classification task usually involves dividing data into training and testing sets. Each instance in the training set has one “target value” (i.e., the class label) and several “attributes” (i.e., the features or observed variables). The goal of SVM was to produce a model (based on the training data) that predicts target values for the test data (a set of patterns with a known label not considered in the training but used to evaluate the accuracy of the classification). A SVM model represents the samples as points in the space of the features. In an ideal case, after mapping, the separate categories are divided by a hyperplane. Unlabelled samples are then mapped into that same space and expected to belong to a category based on the side of the hyperplane into which they fall. SVMs are primarily designed for 2-class classification problems; therefore, in its most basic form, it is a binary and linear classifier, i.e., resulting in classification using a linear hyperplane function (see “Appendix”). It often happens that the sets to be classified cannot be separated linearly in that space. In such cases, the original finite-dimensional space can be mapped into a higher-dimensional space using the kernel trick, which is likely to make separation easier in that space (Cortes and Vapnik 1995). The multiclass problem (i.e., more than two classes) is often resolved by dividing the problem into smaller, simpler binary cases. The formal definition of the method and its principal aspects are presented in “Appendix.”

The effectiveness of SVM depends on the selection of the parameters controlling classification, i.e., the hyperplane parameters, the degree of misclassification, as well as the kernel parameters (see “Appendix”). The best parameter combination is selected by a grid search (Cortes and Vapnik 1995). The entire dataset is divided into smaller sets (n -folds). For each subset, one training set and one testing set are created, and the input variables are correlated in a grid search. The parameters with the best cross-validation accuracy in each n -fold are picked, and usually an average is then used for the classification. This work uses the PRTools toolbox for MATLAB (Duin et al. 2007).

Within a supervised classification framework, a SVM statistical learning algorithm is used on the Grenoble dataset to label the buildings according to the desired EMS98

standard for seismic vulnerability classes. Solving the optimization problem (“Appendix”) gives the parameters of the maximum-margin hyperplane needed for the classification. Having found the best hyperplane (using only the training set), accuracy is estimated automatically using the remaining data (the test set), i.e., by comparing the new estimated vulnerability class with the “real” one. Accuracy is thus measured by creating a confusion matrix and calculating the ratio between the sum of the diagonal values (correct classification) over the sum of all the elements in the matrix.

3.1.1 First phase: learning

In the first phase, the entire dataset is divided into two. The elements that form the training set are selected randomly each time the classifier is run, but respecting the distribution of classes. This introduces variability that has a slight effect on accuracy. To take this variability into account, 2,000 calculations were run (2,000 random training and testing divisions) and an accuracy histogram was created. The histogram shows a Gaussian-like distribution (Fig. 3). The median and the 16 and 84 percentiles can be estimated as a measurement of deviation. Figure 3a shows the histogram of accuracy for a training set of 30 % of Grenoble dataset and considering three attributes (i.e., construction period, number of floors, and shape of the roof).

Furthermore, accuracy will depend on the size of the training set (as a percentage of the total set). Figure 3b shows the evolution of median accuracy for growing sizes of training sets including dispersion (16 and 84 percentiles). The evolution shown in Fig. 3b is independent of the attributes included in the classification, and the same trend—regarding training set size—is found regardless of the dataset studied. Above 20 and 30 % of training set size, maximum attainable accuracy is reached, and the influence of increasing size is lessened. A training size of 30 % is therefore used for the calculations hereafter.

Finally, mean accuracy will depend on the building information (attributes) incorporated to train the machine. Keeping this idea in mind, the method was run on the Grenoble NERA subset, for which several building features are available including those obtained by

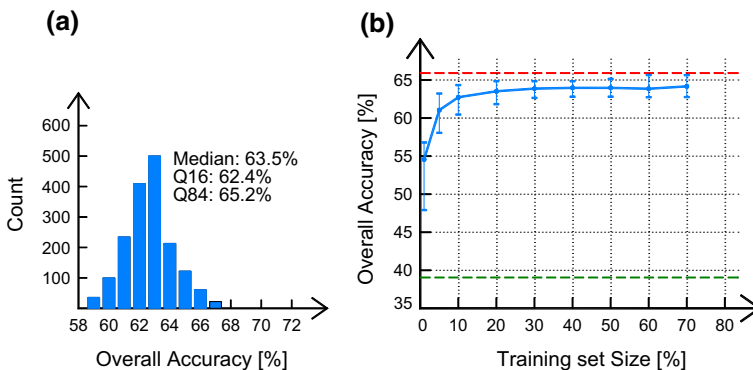


Fig. 3 **a** Histogram of accuracy for 30 % training set and 2,000 runs. **b** Evolution of accuracy and dispersion for different training set sizes. Accuracy increases and dispersion decreases as training size increases up to approximately 25 %; then, it stabilizes at the final value. The lower-limit is the accuracy obtained if all classes are simply assigned to the most probable class (*bottom green line*). The maximum possible accuracy is obtained using 100 % of data for training and testing on the same set (*upper-limit red line*)

processing remote sensing data. Each test involved different attributes, different numbers of attributes, and their combinations. In order to capture only the individual influence of these attributes on the accuracy of the estimation, exactly the same NERA building dataset and training set size (30 %) were used throughout.

The characteristics obtained by the NERA survey (i.e., construction period and number of floors) proved to be the basis of a relatively good classification and should always be included to achieve acceptable accuracy of 62.4 % in the estimation of EMS98 vulnerability class (buildings correctly classified) (Fig. 4a). By adding roof shape, a parameter obtained by processing aerial images, accuracy is improved slightly to 63.5 %. The shape of the roof is indirectly related to construction material. Accuracy is not enhanced drastically, since indirect construction material information might be also included in the other two attributes. In other words, the added information is not completely independent (Fig. 4b). Note that many features can be extracted from remotely sensed data, but not all are independent and therefore add no new information for the classifier to work with. Out

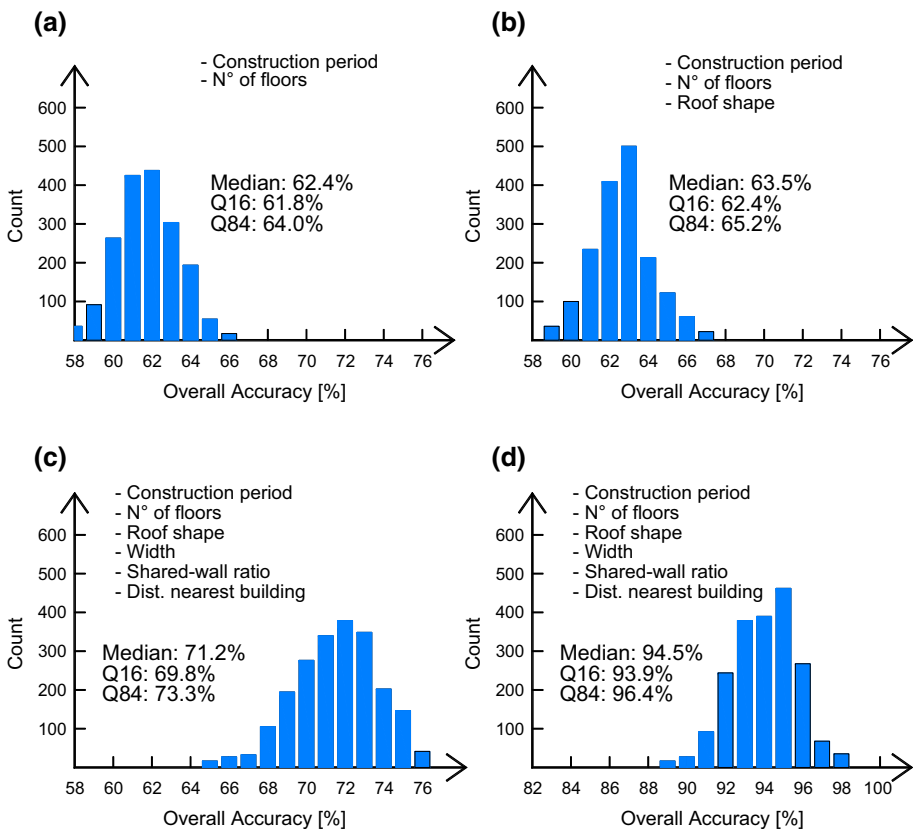


Fig. 4 Effects of different attributes on the accuracy. **a** Only two attributes: construction period and number of floors. **b** Three attributes, after adding shape of the roof. **c** Six attributes, after adding three parameters obtained from cadastral data processing: width of buildings, shared wall ratio (ratio between shared walls and the whole perimeter), and distance to nearest building (an indication of urban environment density). **d** Six attributes, but merging vulnerability classes into only three classes (A–B); (C–D); (E–F). Note change in x-axis range in Fig. 4d)

of the fifteen image-processing attributes available in NERA subset, only three produced a significant improvement of accuracy: width of the mean area-enclosing rectangle of the building footprint, shared wall ratio, and finally, average distance to nearest buildings. These three features represent the shape of urbanization. For example, average distance to nearest building is as sort of measurement of building density, a low-average distance indicates a cluster of buildings close to each other. By adding these pieces of information to the process, mean accuracy reaches 71.2 % of correctly classified buildings (Fig. 4c).

Figure 4 shows a general trend: the addition of more (independent) information on buildings improves the accuracy of the method. In all cases, the dispersion regarding the random selection of the training set elements is slight. Furthermore, 80 % of the misclassified buildings are labelled with a vulnerability class neighboring the correct one. The confusion matrix shows most values immediately bordering the diagonal and zero elsewhere (Table 1). Since the classifier struggles to “differentiate” nearby classes clearly, the effect of merging them was studied by reducing the multiclass problem from six to only three classes. Classes A and B were joined to make class 1, C and D form class 2, and E and F class 3. Classifier accuracy increased drastically, reaching 94 % of correctly assigned buildings (Fig. 4d). For this last example, it is worth noticing that even if accuracy in classification increases drastically, this does not mean that accuracy in vulnerability evaluation increases too, since we have a rougher vulnerability classification. For the rest of this study, a six-class classification is used.

3.1.2 Second phase: application to the Grenoble dataset

The second phase is then implemented to obtain the geo-localized distribution of vulnerability classes in each IRIS, knowing some attributes for the whole French territory. Since INSEE data only give information on two building features (per IRIS unit), the SVM was trained only with the “number of storeys” and “construction period” attributes for the Grenoble dataset, and then used at the scale of an IRIS unit.

As seen previously, the SVM assigns a class according to the side of the classification function (hyperplane) on which the point falls. However, classification is not always clear, even after the hyperplane has been defined in the first phase. For example, if a point falls

Table 1 Example of confusion matrix obtained by SVM on the NERA subset, considering a 30 % training set corresponding to the median case

	A	B	C	D	E	F	
A	52	24	2	0	0	0	
B	45	142	7	0	0	0	
C	0	3	45	22	0	0	
D	0	0	6	10	2	0	
E	0	0	0	0	28	0	
F	0	0	0	0	0	0	
	97	169	60	32	30	0	388
Accuracy	0.71						

Six classes (A–E) and six attributes (construction period, number of floors, roof shape, width, shared wall ratio, and average distance to nearest building). Columns correspond to the “real” vulnerability class and rows to the estimated vulnerability class (e.g., from the 169 class “B” buildings, 24 were assigned as “A,” 142 as “B,” and 3 as “C”). The values on the diagonal (in bold) are the correctly assigned building classes

Table 2 GVM proxy

INSEE attributes	(A)	(B)	(C)	(D)	(E)
<1945 and ≤3 fl.	0.221	0.636	0.103	0.031	0.009
(1945–1970) and ≤3 fl.	0.074	0.672	0.184	0.054	0.016
(1970–2000) and ≤3 fl.	0.019	0.205	0.043	0.041	0.691
>2000 and ≤3 fl.	0.013	0.207	0.021	0.014	0.746
<1945 and (4–5) fl.	0.119	0.660	0.175	0.037	0.009
(1945–1970) and (4–5) fl.	0.011	0.022	0.779	0.163	0.025
(1970–2000) and (4–5) fl.	0.010	0.055	0.069	0.075	0.793
>2000 and (4–5) fl.	0.009	0.065	0.026	0.030	0.871
<1945 and ≥6 fl.	0.043	0.058	0.802	0.084	0.013
(1945–1970) and ≥6 fl.	0.013	0.020	0.245	0.685	0.038
(1970–2000) and ≥6 fl.	0.010	0.026	0.096	0.606	0.261
>2000 and ≥6 fl.	0.025	0.068	0.101	0.281	0.526

“Confidence” values for the classification of each combination of attributes in EMS98 vulnerability classes. Obtained by SVM applied to the VULNERALP database with a 30 % training set
fl. floors

into a clearly divided region of the space, confidence in the classification will be near to one (or 100 %). But in some cases, confidence is lower for points falling near the hyperplane. In any case, SVM assigns the value with the highest confidence percentage. The method allows viewing of the “confidence” at each decision it makes.

Once the machine has been trained and to take this confidence into account, twelve points representing all the possible combinations of the two attributes (i.e., four categories of construction period and three ranges of number of floors) were added to the classification. A Grenoble Vulnerability Matrix (GVM) was created with the confidence distribution provided by the SVM considering each combination (Table 2). Considering the values of Table 2 as conditional probabilities to be in an EMS98 class knowing the building attributes, we calculate

$$N_j(X) = \sum_1^{12} N_{ji}P(X|Y_i) \tag{1}$$

where $N_j(X)$ is the number of buildings of vulnerability class $X_i = \{A, B, C, D, E\}$ in each j IRIS cell, N_{ji} the number of buildings with attributes Y_i in j , and $P(X|Y_i)$ the value of the probability given by the GVM proxy for the $X \rightarrow Y_i$ association (Table 2).

Since IRIS cells are geo-localized throughout France, a vulnerability map of the whole country can be produced, based on the GVM proxy. Figure 5 shows the vulnerability classes in Grenoble computed using the GVM proxy, considering (number of floors) and (construction period). The same main trends as those reported by Gueguen et al. (2007) and Michel et al. (2012) are also observed in Fig. 5: highest vulnerability in the historic downtown area, lowest vulnerability around the periphery and heterogeneous intermediate districts covering all periods of urbanization and mixing masonry and reinforced concrete buildings. Application of the proxy to the entire country assumes a Grenoble-like urbanization nationwide, and this assumption will be tested in Sects. 4 and 5.

3.2 Association rule learning (ARL)

Association rule learning is a popular and well-documented method for discovering relationships between variables in large databases. Agrawal et al. (1993) introduced

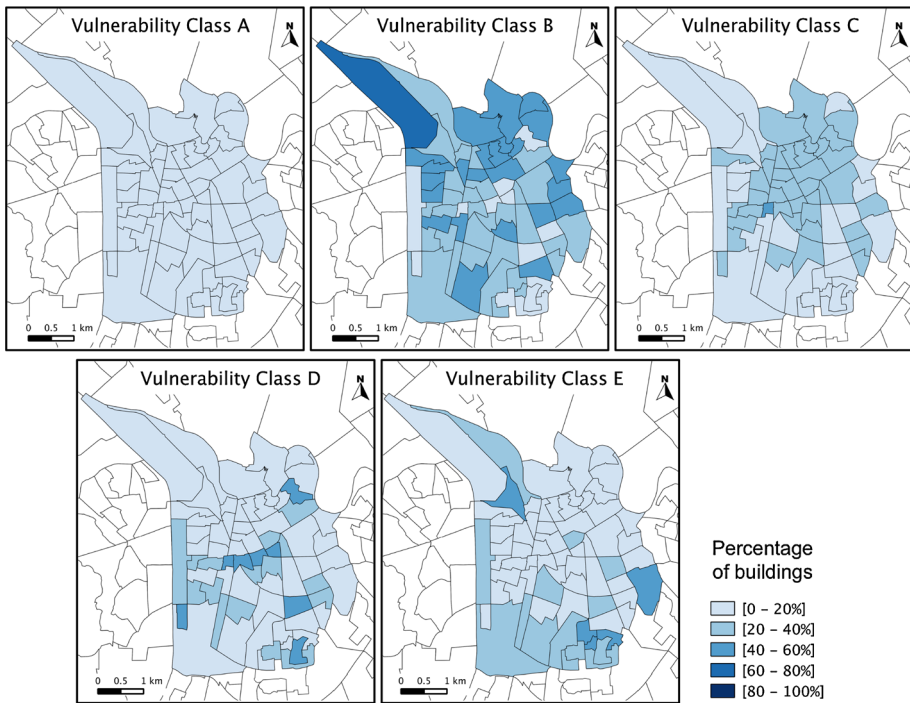


Fig. 5 Distribution of the EMS98 vulnerability class in Grenoble computed using the GVM proxy (SVM) considering INSEE attributes, i.e., construction period and number of floors

association rules as *if/then* statements to help reveal relationships between seemingly unrelated data in a relational database or other information repository. Riedel et al. (2014) presented the method and its application to France using a reduced Grenoble dataset. In this work, we develop a vulnerability proxy, using the simplified ARL method on this database of buildings in Grenoble. Structural information (attributes Y) and EMS98 vulnerability class (item X_i) allow definition of a conditional matrix between them (the learning phase). The conditional probability of having class $X_i = \{A, B, C, D, E\}$ knowing that an event Y has a nonzero probability (i.e., the probability of X_i , given Y) is the number denoted by $P(X|Y)$ and defined by

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} \tag{2}$$

Since the attributes available in INSEE are “number of floors” and “period of construction,” our study focuses on these two pieces of information, with the objective of extending the association to the whole French territory. Using Eq. (2), the vulnerability class X_i according to EMS98 is then associated with attributes and used as a vulnerability proxy. Table 3 summarizes the GVM for each conditional probability of being in EMS98 class X , knowing information related to Y . After the learning phase giving the GVM proxy, the second phase is implemented to obtain a geo-localized distribution of classes X_i in each IRIS, knowing Y for the whole French territory and using the formula:

Table 3 Conditional probabilities for each EMS98 vulnerability class according to building attributes

INSEE attributes	$P(A)$	$P(B)$	$P(C)$	$P(D)$	$P(E)$
<1945 and ≤ 3 fl.	0.390	0.483	0.086	0.039	0.002
(1945–1970) and ≤ 3 fl.	0.008	0.818	0.131	0.036	0.008
(1970–2000) and ≤ 3 fl.	0.000	0.345	0.005	0.010	0.641
>2000 and ≤ 3 fl.	0.000	0.200	0.000	0.000	0.800
<1945 and (4–5) fl.	0.113	0.556	0.289	0.042	0.000
(1945–1970) and (4–5) fl.	0.000	0.008	0.803	0.174	0.015
(1970–2000) and (4–5) fl.	0.000	0.000	0.016	0.000	0.984
>2000 and (4–5) fl.	0.000	0.000	0.000	0.000	1.000
<1945 and ≥ 6 fl.	0.000	0.029	0.912	0.059	0.000
(1945–1970) and ≥ 6 fl.	0.000	0.000	0.396	0.604	0.000
(1970–2000) and ≥ 6 fl.	0.000	0.000	0.017	0.521	0.462
>2000 and ≥ 6 fl.	0.000	0.000	0.100	0.250	0.650

Obtained by the learning phase applied to the VULNERALP database (GVM proxy)
fl. floors

$$P_j(X) = \frac{\sum_1^{12} N_{ji}P(X|Y_i)}{N} \tag{3}$$

where $P_j(X)$ is the probability of having vulnerability class $X_i = (A, B, C, D, E)$ in each j IRIS cell, N_{ji} the number of buildings with attribute Y_i in j , N the total number of buildings in IRIS j , and $P(X|Y_i)$ the value of the probability given by the GVM proxy for the $X \rightarrow Y_i$ association (Table 3). Figure 6 shows the computed vulnerability classes in Grenoble together with a comparison of the results of computed vulnerability using ARL and SVM. Similar results are found, and the general trends of urbanization can also be observed. For each IRIS unit in Grenoble, the ratio between the number of buildings in each vulnerability class obtained using the ARL proxy and the number obtained used the SVM proxy is also shown in Fig. 6. The average ratio for the city, while close to unity, is higher than 1 for the most vulnerable classes and lower for the least vulnerable classes. This suggests that, compared with SVM, ARL predicts more buildings of the more vulnerable classes and fewer of the less vulnerable classes. For a particular earthquake scenario, and on the broader scale, greater estimated damage would be expected if vulnerability was estimated using ARL rather than SVM, as will be shown in Sect. 5.

4 Validation in the city of Nice

The city of Nice, one of France’s cities most exposed to seismic risk, has undergone numerous vulnerability evaluations (e.g., Bard et al. 2005; Spence and Lebrun 2006). In order to validate the GVM proxies, seismic damage in Nice was predicted using both GVM proxies applied to INSEE data (obtained by SVM and ARL) and with the vulnerability indexes obtained by the RiskUE method. Validation was achieved by comparing the damage rate computed at the macroscale for different seismic scenarios. For the RiskUE analysis, vulnerability is measured in terms of a vulnerability index (I_v), which is defined taking into account the structural characteristics of buildings and adjusted according to damage observed during earthquakes in Italy. The hazard is described in terms of macroseismic intensity, according to the European Macroseismic Scale EMS98. The

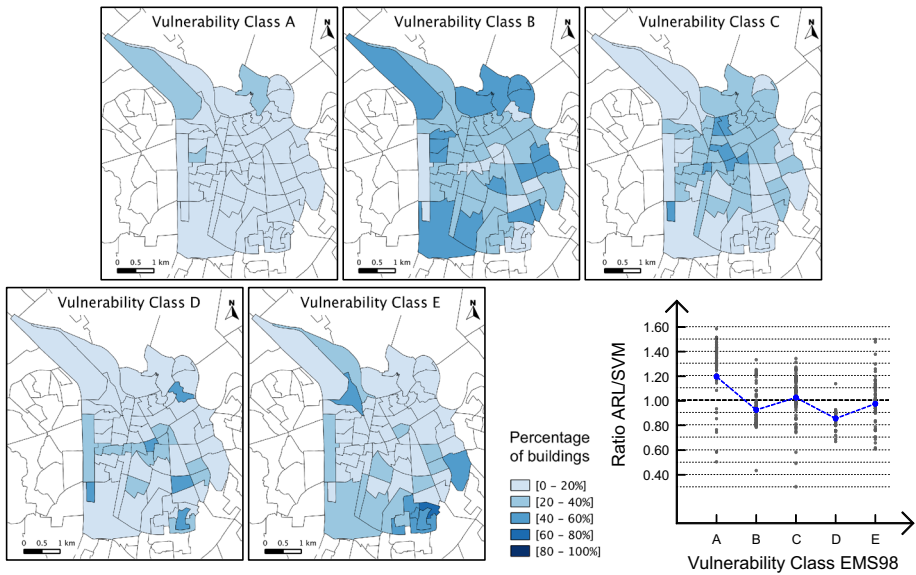


Fig. 6 Distribution of the EMS98 vulnerability class in Grenoble computed using the GVM proxy (ARL) considering INSEE attributes, i.e., construction period and number of floors. Comparison between estimated vulnerability classes using ARL and SVM proxies. For each IRIS unit and for each vulnerability class, the ratio between the number of buildings estimated by ARL and the number estimated by SVM is shown (gray dots). The average ratio for the city is shown in blue dots

correlation between seismic input and expected damage, as a function of the assessed vulnerability, is described by the analytical function

$$\mu_D = 2.5 \left(1 + \tanh \frac{I_{EMS98} + 6.25I_v - 13.1}{2.3} \right) \tag{4}$$

where μ_D is the average observed damage in buildings of the given vulnerability index I_v and subjected to a given macroseismic intensity. EMS98 characterizes damage according to six levels (D_k with $k = 0, 1, 2, 3, 4, 5$), ranging from D_0 (no damage) to D_5 (complete destruction). To take into account the variability of the damage level k in a set of buildings, Lagomarsino and Giovinazzi (2006) assume a binomial distribution.

Therefore, the probability $P(D_k)$ of observing each damage level D_k ($k = 0-5$) for a given mean damage μ_D is evaluated according to the probability function of the binomial distribution, namely

$$P(D_k) = \frac{5!}{k!(5-k)!} \left(\frac{\mu_D}{5} \right)^k \left(1 - \frac{\mu_D}{5} \right)^{5-k} \tag{5}$$

In Nice, the RiskUE project identified 27 zones considered homogeneous for vulnerability assessment (Fig. 7, top left). A random sample of buildings was selected to assess the vulnerability of each zone, with I_v between 0.365 and 0.849. Each zone was then geolocalized and characterized by an average vulnerability I_v and a probable range $I_v^{\max} - I_v^{\min}$. The spatial distribution of the EMS98 vulnerability classes deduced from the GVM proxy (SVM) is given in Fig. 7. Vulnerability is heterogeneous between IRIS cells, but the

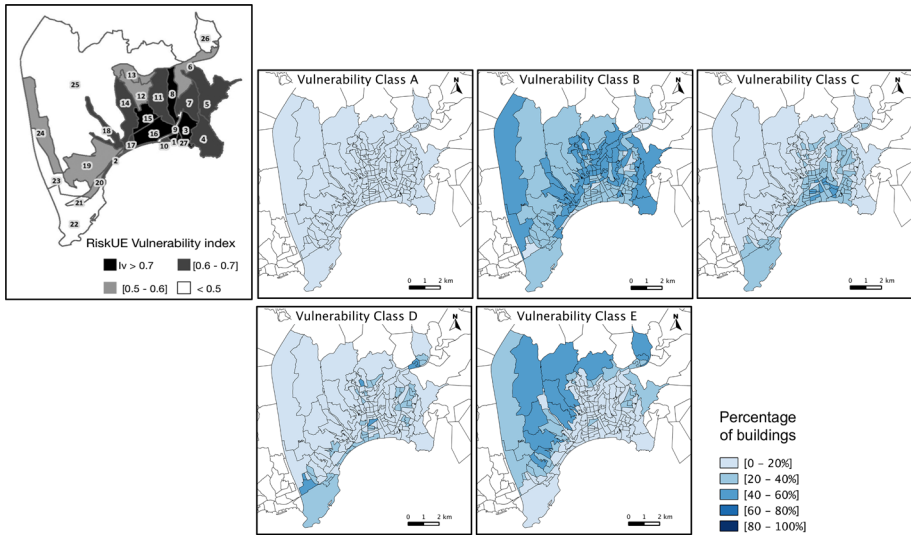


Fig. 7 Application of the GVM proxy to the city of Nice. Distribution of the seismic vulnerability index computed by the RiskUE method (*top left*). Distribution of the EMS98 vulnerability classes in Nice computed by the GVM SVM proxy

traditional trends observed in European urban centers are also observed in Nice, i.e., a more vulnerable, historic downtown area, and more modern, less vulnerable suburban areas. The methods used by INSEE and RiskUE to divide the city into zones are not the same; therefore, comparison is made at the scale of the city.

For the EMS98 scale, the frequency of expected damage is defined by linguistic terms (“few,” “many,” “most” buildings). The definitions provided by EMS98 can be regarded as damage matrices (Table 4, top). Lagomarsino and Giovinazzi (2006) proposed a numerical translation for these qualitative terms such as: “some” (5 %), “many” (35 %), and “most” (80 %). On this basis, damage matrices are established giving the occurrence probability distribution $P(D = D_k)$ for each intensity as a function of building vulnerability (Table 4, bottom). Damage to buildings occurs from intensity V , with D_1 damage grade affecting some buildings of classes A and B (Grunthal and Levret 2001). These matrices have to be completed for the damage range for which there is no definition, since the sum of the different damage grades must be equal to one for each intensity. According to EMS98, we assume (1) a monotonically decreasing function at a high damage level D_k ; (2) a normal distribution of probabilities around the mean damage grade for an intermediate level of damage; and (3) a monotonically increasing function at a low damage level D_k . For example, for buildings in class A and intensity VII, EMS98 says that “many (35 %) buildings in vulnerability class A suffer grade 3 damage and a few (5 %) suffer grade 4 damage.” The remaining 60 % are distributed over the lower levels of damage to propose a continuous, smoothed probability function of damage (Riedel et al. 2014).

The probability of occurrence of damage D_k for intensities V–XII, computed using RiskUE and the GVM proxies (i.e., ARL and SVM), averaged at the scale of the city, is shown in Fig. 8. The median I_v is used for RiskUE while the probabilities estimated using the range I_v^{\max} and I_v^{\min} are shown as dotted black lines (uncertainty range). For GVM methods, the proxy giving the median accuracy is used, while the estimations using the

Table 4 EMS98 macroseismic scale. Implicit damage probability matrix for vulnerability class A (top). Damage occurrence probability from EMS98 for vulnerability class A and macroseismic intensity from IV to XII (bottom)

I	D ₁	D ₂	D ₃	D ₄	D ₅
IV	–	–	–	–	–
V	Few	–	–	–	–
VI	Many	Few	–	–	–
VII	–	–	Many	Few	–
VIII	–	–	–	Many	Few
IX	–	–	–	–	Many
X	–	–	–	–	Most
XI	–	–	–	–	–
XII	–	–	–	–	–

I	D ₀	D ₁	D ₂	D ₃	D ₄	D ₅
IV	1.00	0.00	0.00	0.00	0.00	0.00
V	0.95	0.05	0.00	0.00	0.00	0.00
VI	0.60	0.35	0.05	0.00	0.00	0.00
VII	0.05	0.20	0.35	0.35	0.05	0.00
VIII	0.00	0.05	0.20	0.35	0.35	0.05
IX	0.00	0.00	0.05	0.25	0.35	0.35
X	0.00	0.00	0.00	0.00	0.20	0.80
XI	0.00	0.00	0.00	0.00	0.00	1.00
XII	0.00	0.00	0.00	0.00	0.00	1.00

proxy giving 16 and 84 percentile accuracy are plotted as dotted lines. Note that the values change so little that the curves overlap. Overall, slight differences are observed between the methods. Nevertheless, at the macroscale and for the intensities causing damage, the orders of magnitude of damage occurrence probability are quite similar. Although the GVM proxy was defined for a Grenoble-like environment, the damage prediction provides reliable information at the first order. Moreover, the simplified approach of computing the distribution of vulnerability class per IRIS based on just two very simple attributes allows generalization to the whole of the French territory, ultimately producing a geo-localized assessment of vulnerability.

5 Historic earthquake simulations

Once the seismic vulnerability of a region is estimated, probable damage to buildings can be assessed for any given seismic demand input. In this section, damage is modelled for a few historic earthquakes in France to enable (1) the estimation of damage if the same (or similar) earthquake were to strike today, using actual vulnerability; (2) validation of the model on the basis of damage estimations, using vulnerability at the time of the earthquake.

France is characterized by moderate seismicity, and destructive earthquakes are rare. Comparing modelled and real damage is not easy since the information concerning damage

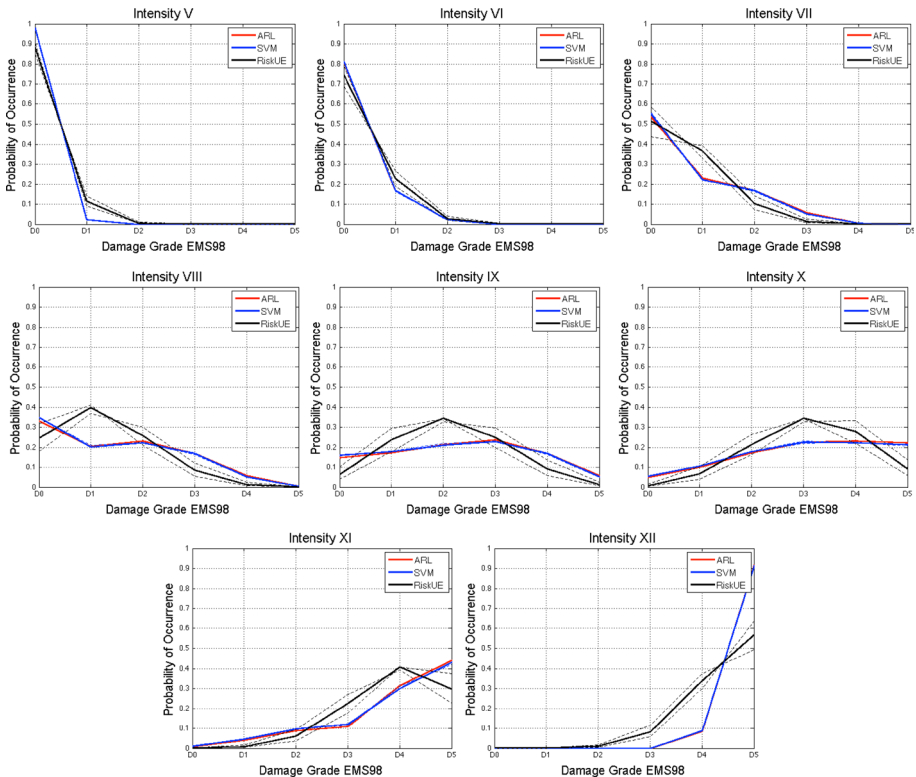


Fig. 8 Prediction of damage in Nice using RiskUE (*black continuous line*) with its uncertainty range (*black hidden line*) and using GVM proxy methods, i.e., ARL (*red continuous line*) and SVM (*blue continuous line*) for intensity scenarios ranging from V to XII

observation is old, sparse, and often imprecise. Nonetheless, for some historic earthquakes, quantitative information on observed damage can be retrieved (SisFrance, <http://www.sisfrance.fr>; Scotti et al. 2004), although the type of damage is not well detailed. Two of the best-documented historic French earthquakes are modelled in this section, using the macroseismic intensities observed as the seismic demand. For this evaluation, it is assumed that MSK reported intensities coincide with EMS98 scale intensities. This analysis, carried out as an example focused on the seismic vulnerability, eliminates the difficulties of simulating ground motion using prediction equations. Evaluation of the effects of a historic event occurring at the present time allows representation of the evolution of vulnerability over time.

5.1 Lambesc earthquake (1909)

The historic Lambesc earthquake, which hit southeastern France in June 1909, is probably the strongest earthquake in the recent history of France. This earthquake produced macroseismic intensities MSK between VIII and IX in the epicentral area (Fig. 9), 30 km from Marseille. Its magnitude was recently re-appraised and estimated at around 6.0 (Baroux et al. 2003). It was a shallow depth earthquake (less than 10 km), and it was felt

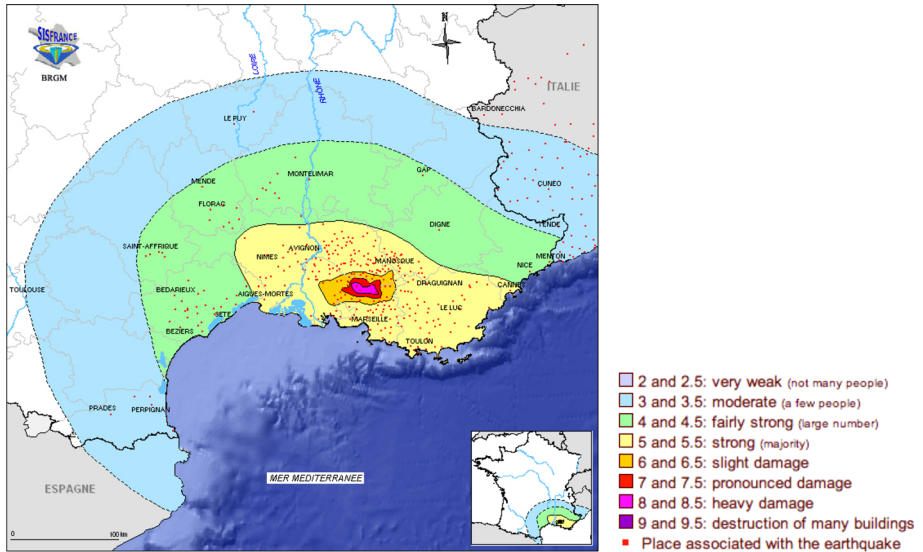


Fig. 9 Isoseists contour lines and intensity domains (on MSK scale) for the historic 1909 Lambesc earthquake (SisFrance catalogue, BRGM, EDF, IRSN)

more than 300 km from the epicenter. 46 casualties and about 250 injured were reported after the event and referenced in the SisFrance database (Scotti et al. 2004). In terms of losses, Lambert (1997) reported serious damage to buildings in different cities within the region affected. This earthquake is all the more important since it served as a scenario in 1982 to forecast seismic losses and casualties, taking into account urbanization evolutions between 1909 and 1982. The results provided information that increased the awareness of the authorities, an element (among others) that led to the establishment of the modern national earthquake rules for construction design, published about 15 years later.

In our analysis, we consider an area including all sectors with a macroseismic intensity above IV. In total, the studied area represents 4,162 IRIS cells, covering a large part of southeastern France. Using the GVM proxies calculated using the two data mining methods, ARL and SVM, the vulnerability class distribution was computed from INSEE data. Since the INSEE database gives the distribution of buildings present in 2008 according to the period of construction, and no information on the inventory of past—and now nonexistent—buildings, we assume that the number of buildings per IRIS corresponds to the buildings that were present in each period. We thus accept a slow rate of replacement and are able to provide an approximate simulation of the damage produced by the 1909 Lambesc earthquake, assuming that the buildings present in 1915 and still existing in 2008 were those present in 1909. We did not take into account the possible retrofitting or modifications of existing structures, as well as some special structural characteristics (e.g., short column, soft story, and irregularities). These characteristics certainly affect the seismic vulnerability of buildings, but for a macroscale evaluation, they are not available in national census and they cannot be obtained through the processing of aerial/satellite images.

The temporal evolution of seismic vulnerability can be assessed for different periods of construction in order to visualize the effects of the rate of urbanization on seismic

vulnerability. In general, probabilities for high vulnerability classes are reduced with time, and probabilities for the less vulnerable classes increase, reflecting the construction of new buildings that are less vulnerable. However, in terms of numbers, vulnerable buildings (classes A and B) still represent a large portion of the total buildings. Furthermore, the evolution of the number of buildings for the considered IRIS cells is significant, with more than 160,000 new constructions between 1945 and 2008, which is also coherent with the urbanization rate observed in Grenoble and reflecting the post-World War II needs for housing. New buildings are, in general, less vulnerable, thanks to the use of reinforced concrete rather than masonry and the application of new building codes, introduced after the 1970s. The number of buildings for each EMS98 damage grade D_0 – D_5 (or damage probability) is computed by crossing the GVM proxy applied to the INSEE attributes and using the 1909 macroseismic intensity curves as seismic demand, as follows:

$$N_{j,I_{EMS98}}^{D_k} = \sum_{i=A}^{i=F} N_{ji} P(D_k|i, I_{EMS98}) \quad (6)$$

where $N_{j,I_{EMS98}}^{D_k}$ is the number of buildings with damage grade D_k ($k = 0$ – 5) for each j IRIS cell and intensity I_{EMS98} . N_{ji} is the number of buildings of vulnerability class i ($i = A$ – F) for IRIS j , and $P(D_k|i, I_{EMS98})$, the probability of damage grade D_k of a vulnerability class i building for a given macroseismic intensity EMS98 (e.g., values of Table 4, bottom for class A). IRIS units not entirely within an iso-value (i.e., intersected by an isoseist line) are divided according to the surface ratio and thus have two intensity values. The number of buildings is distributed in proportion to the area of each subunit and respecting the vulnerability class distribution inside the original IRIS.

The number of buildings in each damage grade according to the ARL proxy is displayed on Fig. 10 and according to the SVM proxy on Fig. 11. They are grouped into three classes according to the EMS98 scale: slight damage (D_1), moderate damage ($D_2 + D_3$), and serious damage ($D_4 + D_5$). Figures 10a and 11a represent the number of buildings in each class of damage for the 1909 earthquake affecting dwellings built before 1915. The highest damage computed is localized close to the epicenter, which is where the highest intensities are found. Between 170 and 240 buildings suffer heavy damage, while between 2,600 and 2,700 are estimated as suffering moderate damage, the rest being distributed over the studied area. The historic information from 1909 concerning cities close to the epicenter enables a reliable estimate of the damage consequences, considering either the whole area or just the cities for which specific historic descriptions are available (Table 5). In this regard, our method allows the estimation of probable damage for each IRIS unit, therefore for each city or town.

Table 5 compares the number of buildings damaged according to historic information (from SisFrance archives) and the number simulated using GVM proxies from ARL and SVM methods using the 1915 catalogue of buildings. Slight differences exist, which may reflect the iso-intensity curves considered as seismic ground motion and especially the differences between the 2008 inventory of buildings built before 1915 and the actual state of urbanization in 1909. Nevertheless, we can assume that the damage obtained is appropriate in terms of damage estimation on the macroscale. The lack of more accurate descriptions of historic damage and information on urbanization at the time prevents better comparison. Estimations using the GVM proxy obtained with SVM seem to be closer to the damage observed, while estimations with the ARL proxy are more conservative, giving a larger number of damaged buildings.

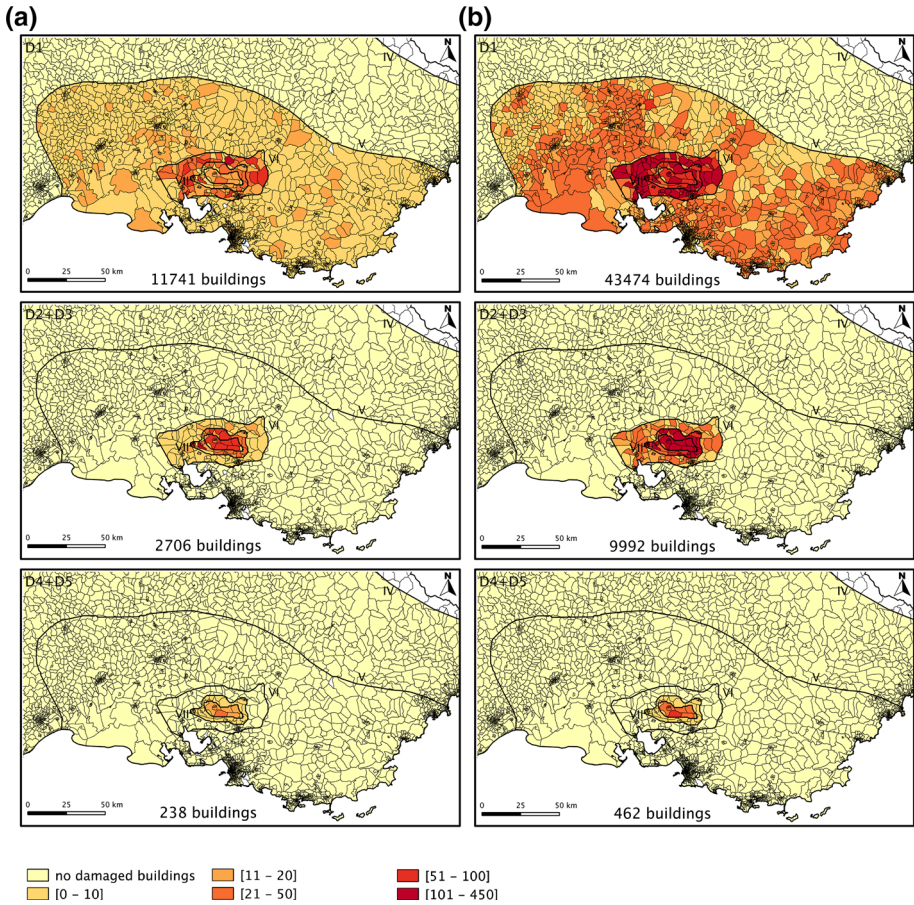


Fig. 10 Evaluation of the level of damage for the Lambesc earthquake scenario considering **a** 1909 urbanization (*left column*) and **b** 2008 urbanization (*right column*), using the GVM proxy obtained from ARL. Damage is grouped by slight D_1 (*top row*), moderate $D_2 + D_3$ (*middle row*), and severe $D_4 + D_5$ (*bottom row*) according to the EMS98 damage scale

Finally, the simulation can be continued by forecasting the impact of a future earthquake with the same characteristics as the 1909 Lambesc earthquake (i.e., same location and same macroseismic intensity) on the state of urbanization in 2008 (Figs. 10b, 11b). In 2008, the region suffering macroseismic intensity V or higher during the 1909 earthquake had more than 1.10 million buildings and a population of more than five million. 60 % of buildings were vulnerability class B or C, and classes D and E represent more than 31 %. If the 1909 earthquake re-occurred in 2008, about 50,000 buildings would be affected with different levels of severity, i.e., approximately 5 % of the total number of buildings. The small epicentral area (intensities VII and VIII) includes more than 14,000 damaged buildings, representing 44 % of the buildings present in this area. All the buildings suffering heavy damage and 81 % of those suffering moderate damage are within this area. Overall, if the same earthquake occurred again, it would cause more damage in terms of number of buildings for any damage type, closely linked to the urbanization growth

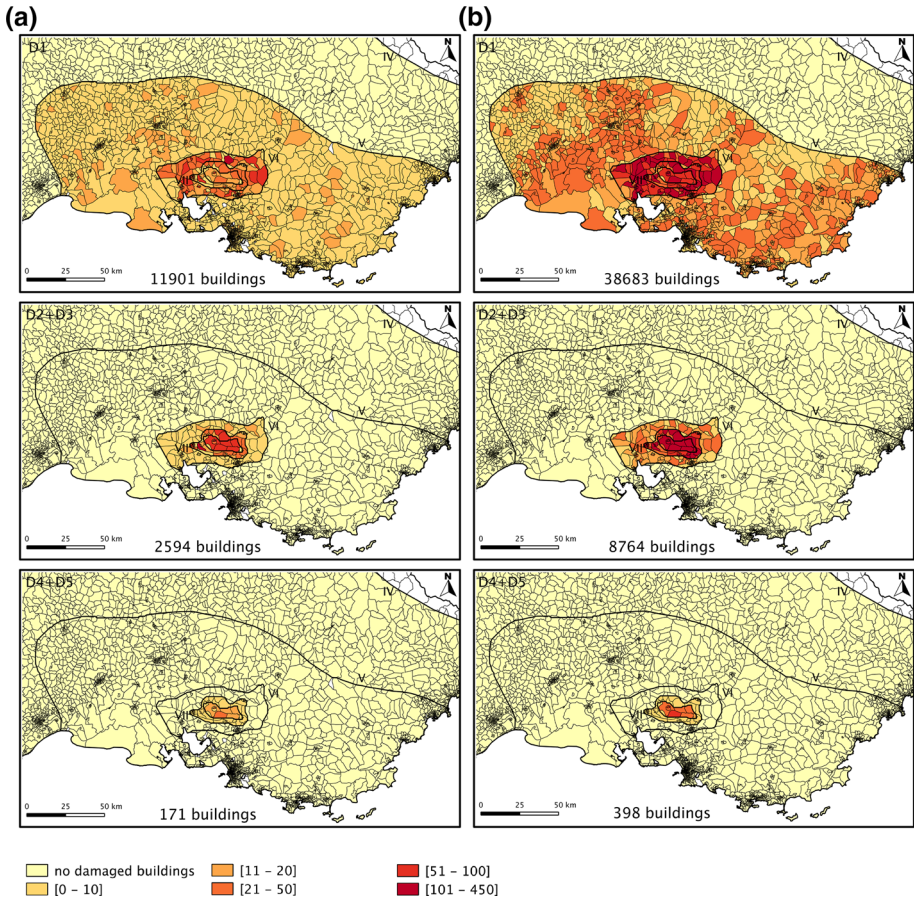


Fig. 11 Evaluation of the level of damage for the Lambesc earthquake scenario considering **a** 1909 urbanization (*left column*) and **b** 2008 urbanization (*right column*), using the GVM proxy obtained from SVM. Damage is grouped by slight D_1 (*top row*), moderate $D_2 + D_3$ (*middle row*), and severe $D_4 + D_5$ (*bottom row*) according to the EMS98 damage scale

Table 5 Comparison between damage observed during the 1909 Lambesc earthquake (historic records SisFrance) and simulated damaged using the GVM proxy obtained by ARL and SVM methods

City or town	Observed	Simulated (ARL)	Simulated (SVM)
Lambesc ($R_{epic} = 5$ km)	600 damaged 50 destroyed	361 D_{1+2+3} 77 D_{4+5}	376 D_{1+2+3} 58 D_{4+5}
Rognes ($R_{epic} = 3$ km)	250 damaged	172 D_{1+2+3} 18 D_{4+5}	173 D_{1+2+3} 14 D_{4+5}
Saint-Canat ($R_{epic} = 4$ km)	310 damaged 50 heavy damaged	148 D_{1+2+3} 28 D_{4+5}	152 D_{1+2+3} 21 D_{4+5}
La R. d'Anthéon ($R_{epic} = 7$ km)	110 heavy damaged	127 D_{1+2+3} 3 D_{4+5}	124 D_{1+2+3} 2 D_{4+5}
Aix-en-Provence ($R_{epic} = 20$ km)	1,500 damaged	1,409 D_{1+2+3} 25 D_{4+5}	1,433 D_{1+2+3} 18 D_{4+5}

D_{1+2+3} total number of buildings with damage grade D_1 , D_2 or D_3

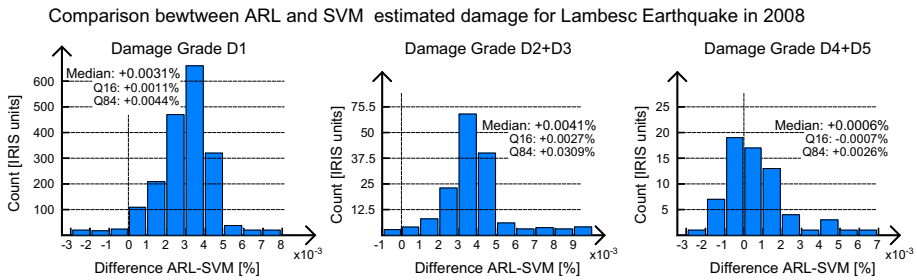


Fig. 12 Comparison between ARL and SVM estimated damage for Lambesc earthquake scenario in 2008 and for different damage levels (histogram). Difference between the percentage of buildings damaged at each IRIS unit estimated using ARL and the percentage estimated using SVM methods. Slight damage D_1 (left), moderate damage (middle), strong damage (right). Note change in axis ranges between figures

between 1909 and 2008 (increased number of buildings with a high percentage of vulnerable classes before 1970). As shown in Figs. 10b and 11b, the probable number of heavily damaged buildings doubles, reaching around 430 constructions [only a few buildings are completely destroyed (D_5)], and the number suffering moderate damage triples, with 9,400 buildings affected for the entire region. 40,000 buildings are expected to suffer slight damage, characterized by hairline cracks in very few walls, falling chimneys or small pieces of plaster, according to the EMS98 damage description.

A comparison between ARL and SVM damage estimations for the Lambesc earthquake scenario in 2008 is shown in Fig. 12. The difference between the percentage of buildings damaged in each IRIS unit estimated using ARL and the percentage estimated using SVM is represented in a histogram for each damage level. The ARL method gives slightly higher percentages (or number of damaged buildings) especially for the lower damages grades.

5.2 Arette earthquake (1967)

Another of most violent events experienced in France during the twentieth century occurred in August 1967 in Arette, in the western Pyrenees near the French–Spanish border. With a magnitude estimated at 5.8 M_L (Rothé and Vitart 1969), this earthquake produced a macroseismic intensity MSK of VIII in the epicentral area (Fig. 13). It was felt in an area with a radius of 220 km from the epicenter and caused 1 death, 15 injured, and major damage to buildings. This analysis considers the area including all sectors with a macroseismic intensity of more than IV on the French side of the border with Spain (1,092 IRIS units).

Vulnerability class distribution is computed following the same procedure as before (Eq. 6) and using the 1967 macroseismic intensity curves as seismic demand. We provide an approximate simulation of the damage caused by the Arette earthquake in 1967, considering buildings built before 1970 and existing in 2008 as those present in 1967. The number of buildings in each damage grade according to the ARL proxy is displayed on Fig. 14 and according to the SVM proxy on Fig. 15. Figures 14a and 15a show the number of buildings in each class of damage for the 1967 earthquake affecting dwellings built before 1970. As for the previously modelled earthquake, the information from 1967 concerning cities close to the epicenter enables a reliable estimate of the damage. Table 6 compares the number of buildings damaged according to historic observations (Rothé and Vitart 1969—SisFrance) and the number estimated by GVM proxy simulation. In spite of the differences, the damage caused is also comparable in this case in terms of damage

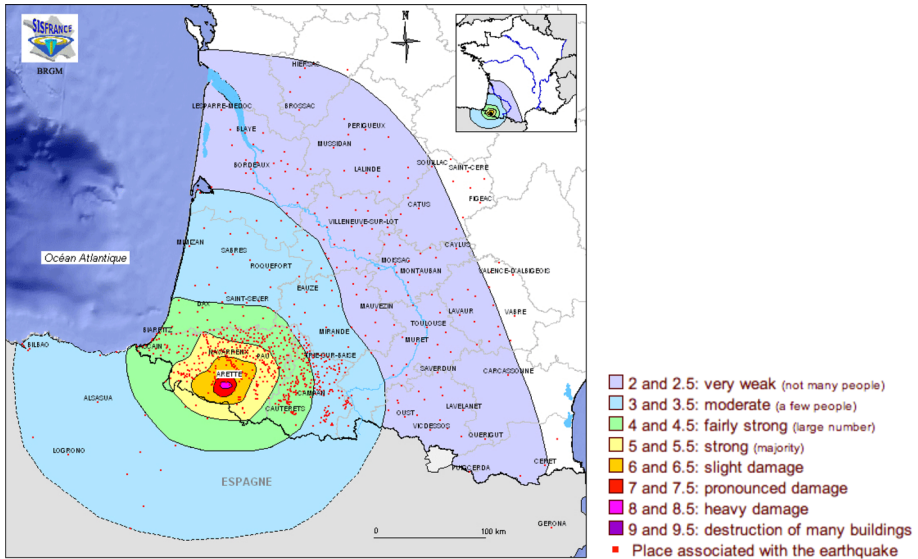


Fig. 13 Isoseists contour lines and intensity domains (on MSK scale) for the historic 1967 Arette earthquake (SisFrance catalogue, BRGM, EDF, IRSN)

estimated on the macroscale. Figures 14b and 15b predict the impact of an earthquake with the same characteristics as the 1967 Arette earthquake on the state of urbanization in 2008. The region that experienced macroseismic intensity V or higher during the earthquake (damage to constructions expected) had about 91,000 buildings and a population of more than 376,000 in 2008. If the 1967 earthquake had re-occurred in 2008, nearly 6,800 buildings would probably have been affected with different levels of severity, i.e., approximately 7 % of the total number of buildings. The epicentral area (intensities VII and VIII) includes more than 1,080 damaged buildings, representing 60 % of the buildings in the area. Every building with heavy damage and 58 % of those suffering moderate damage are inside this area. Even if, as for the previous earthquake, the same earthquake re-occurring in present times would cause more damage (in terms of number of buildings for any damage type), the increase in the number of buildings affected is smaller. As shown in Figs. 14b and 15b, the probable number of heavily damaged buildings remains almost the same, the number of buildings suffering moderate and slight damage increases by 10 and 15 %, respectively. Compared with the Lambesc simulation, the evolution of urbanization over this period of 41 years (1967–2008) is obviously less radical than over almost a century (1909–2008).

Figure 16 shows the comparison between ARL and SVM estimated damage for the Arette earthquake scenario in 2008. As in the previous case, the ARL method gives a slightly higher number of damaged buildings for any damage grade, with a median difference around 0.0035 %.

6 Conclusions

The aim of this paper was to validate a macroscale methodology for seismic vulnerability assessment, in a situation where only a poor description of construction characteristics

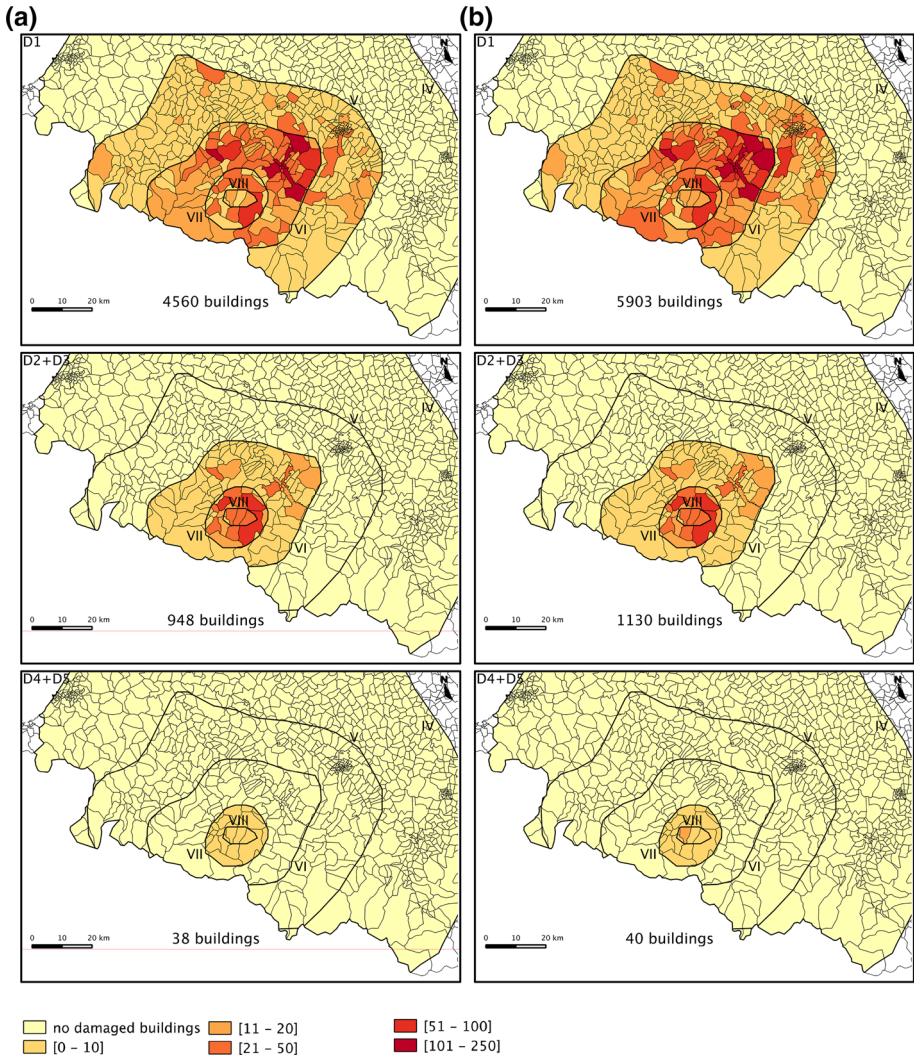


Fig. 14 Evaluation of the level of damage for the Arette earthquake scenario considering **a** 1967 urbanization (*left column*) and **b** 2008 urbanization (*right column*), using the GVM proxy obtained from ARL. Damage is grouped by slight D_1 (*top row*), moderate $D_2 + D_3$ (*middle row*), and severe $D_4 + D_5$ (*bottom row*) according to the EMS98 damage scale

(with respect to those necessary for an ad hoc analysis) is available for a large number of buildings. In a moderate seismic-prone region, where it is often difficult to mobilize resources for the reduction in seismic risk, the idea of using readily available data to expand the assessment to any given region is obviously of interest. Using the information available in Grenoble, we propose two vulnerability proxies (GVM proxy) defined using the ARL and SVM methods. These proxies create a relationship between two building characteristics (present in the French national census database) and their most probable EMS98 vulnerability class. Since INSEE data are available for the whole of the French territory, it is possible to apply the GVM proxy to simulate the impact of historic

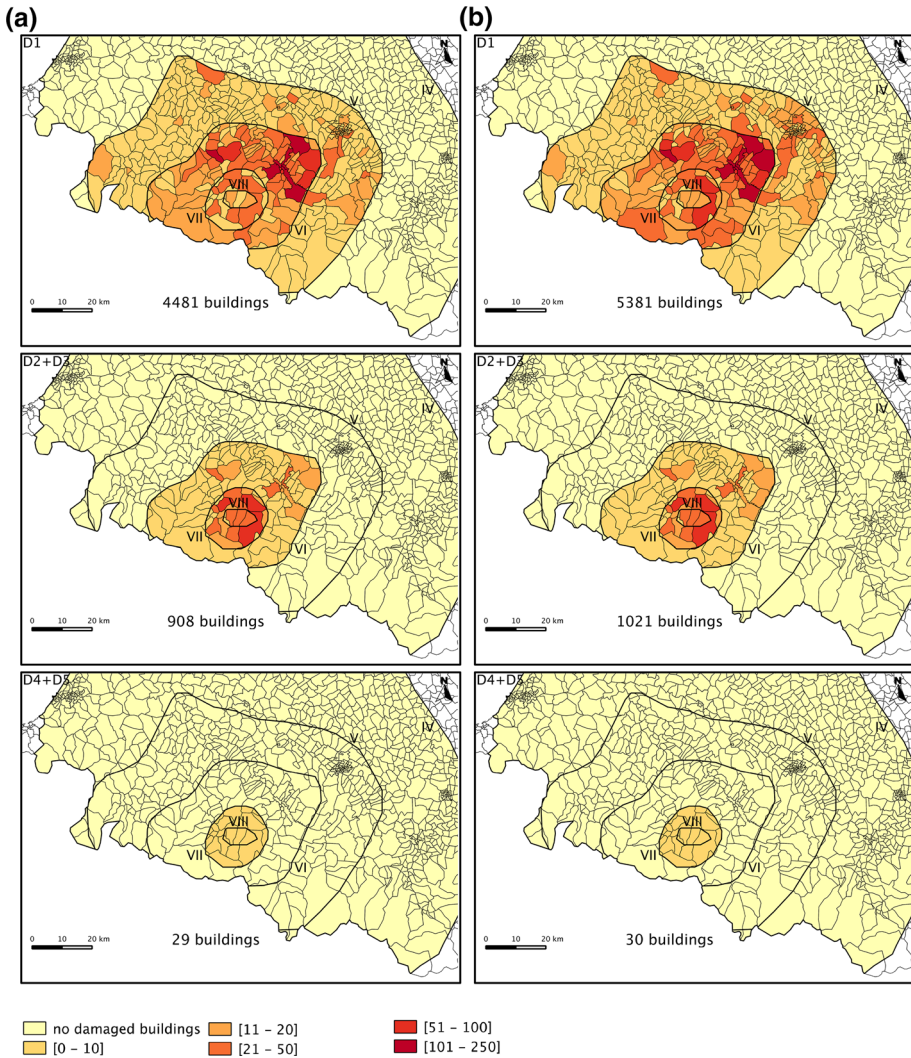


Fig. 15 Evaluation of the level of damage for the Arette earthquake scenario considering **a** 1967 urbanization (*left column*) and **b** 2008 urbanization (*right column*), using the GVM proxy obtained from SVM. Damage is grouped by slight D_1 (*top row*), moderate $D_2 + D_3$ (*middle row*), and severe $D_4 + D_5$ (*bottom row*) according to the EMS98 damage scale

earthquakes on present-day urbanization and/or to forecast damage levels in the impacted zone a few seconds after the occurrence of an earthquake. Even though the proxies were created for France-like environments, their application to other European cities should be tested. Furthermore, the INSEE dataset provides information on residential buildings only. No commercial buildings were included in our damage estimations.

The flexibility and adaptability of the method is one of its main advantages. If the information available is on the scale of a building, the estimation of vulnerability and damage can obviously be carried out on this same scale. This method can easily be applied

Table 6 Comparison between damage observed during the 1967 Arette earthquake (SisFrance) and damaged simulated using the GVM proxy obtained by ARL and SVM methods

City or Town	Observed	Simulated (ARL)	Simulated (SVM)
Basses_Pyrénées (global)	2,283 damaged	5,319 D ₁₊₂	5,229 D ₁₊₂
	340 heavy damaged or destroyed	189 D ₃ 38 D ₄₊₅	160 D ₃ 29 D ₄₊₅
Arette, Lanne and Montory (epicentral area)	40 % heavy damaged or destroyed	270 D ₁₊₂ 104 D ₃₊₄₊₅ (22 %)	269 D ₁₊₂ 95 D ₃₊₄₊₅ (20 %)
Arette ($R_{epic} = 6$ km)	Many heavy damaged some destroyed	150 D ₁₊₂ 30 D ₃ (11 %) 6 D ₄₊₅ (2 %)	146 D ₁₊₂ 28 D ₃ (10 %) 6 D ₄₊₅ (2 %)
Montory ($R_{epic} = 2$ km)	All slight damaged 40 heavy damaged	51 D ₁₊₂ 35 D ₃₊₄₊₅ (total 87 %)	53 D ₁₊₂ 32 D ₃₊₄₊₅ (total 86 %)
Lanne ($R_{epic} = 0.5$ km)	Many damaged some destroyed	94 D ₁₊₂₊₃ (75 %) 8 D ₄₊₅ (6 %)	93 D ₁₊₂₊₃ (74 %) 6 D ₄₊₅ (5 %)

anywhere provided basic information on the buildings is available. We show the adaptability of the method regarding the information available. Having more (or more detailed) independent attributes during the training phase increases the accuracy of the vulnerability class estimation. For example, national census information, satellite or airborne photographs, and cadastral data are cheap sources of information available over a large scale, and further exploration of the impact of urban parameters on vulnerability could be tested in more detail in the future. According to our analysis, SVM provides a better estimate of damage classification compared with historic data. Unfortunately, historic descriptions of damage are sparse and imprecise, and the effectiveness of SVM compared with the ARL method must be confirmed.

The technique was validated in Nice and finally tested for two historic earthquakes that caused damage in France. Although the attributes describing the buildings are very basic, the analyses provide results that confirm the suitability of our solution, providing reliable estimates of damage for earthquake scenarios. EMS98 intensity scale can be closely considered as a definition of the ground motion since it includes in its definition the seismic vulnerability of buildings. In this study, validations of the method were performed with historic earthquakes and based on reported macroseismic intensities considered as ground motion. Forecasted intensities as produced by ShakemapTM might be available minutes after an earthquake.

In Nice, a more sophisticated method (RiskUE), based on a relatively detailed description of structural features and using macroseismic intensity as the ground motion parameter, produced similar levels of damage across the city.

Because of the lack of elements of comparison and the shortage of details about historic damage, it is difficult to quantify the assessment errors that might be obtained for a given earthquake. However, the data mining method, which consists in defining the best relationship between attributes and vulnerability class during the learning phase, appears to be well suited to the large-scale assessment of seismic vulnerability and thus to the simulation of seismic damage. We were able to highlight certain obvious trends, such as the reduction in the proportion of vulnerable buildings with the development of urbanization. We also confirmed and quantified the increasing effects of earthquakes in terms of damage, mainly due to the explosion of urbanization and urban concentrations in certain areas prone to

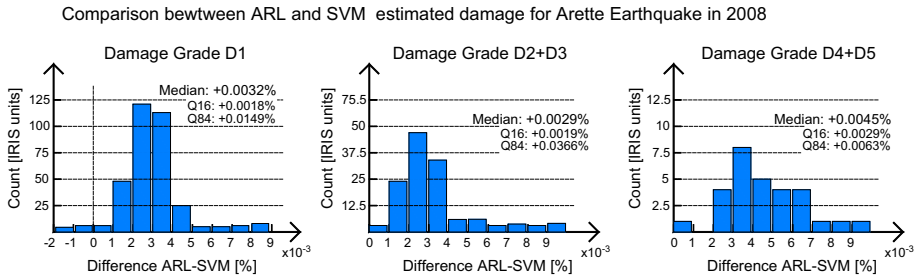


Fig. 16 Comparison between ARL and SVM estimated damage for the Arette earthquake scenario in 2008 and for different damage levels (histogram). Difference between the percentage of buildings damaged in each IRIS unit estimated using ARL and the percentage estimated using the SVM method. Slight damage D_1 (left), moderate damage (middle), severe damage (right). Note the change in axis ranges between figures

seismic hazard. For example, in the Lambesc region, if the 1909 earthquake had occurred in 2013, there would have been serious consequences in terms of casualties and economic losses: 430 buildings would have suffered severe levels of damage (D_4 and D_5), a dozen buildings would have been completely destroyed (D_5), and more than 9,400 buildings would have been affected by moderate damage (D_2 and D_3). Even over a period of 40 years, urbanization development increases the seismic risk of a region (Arette earthquake simulation). We observed a strong increase in damage, even for an earthquake of moderate magnitude, with levels comparable to those observed during earthquakes of similar magnitude in L'Aquila in Italy or Christchurch, New Zealand. It is clear that with a smaller information sample (attributes/vulnerability classes), a particular machine or proxy may be developed for any location to estimate regional damage. These elements are essential to enable the evaluation of economic and human losses. Once the distribution of vulnerability classes is known, the consequences in terms of damage can be simulated rapidly after an earthquake, providing an additional element to the simulation of ground motion via ShakemapTM for a seismic warning system.

Acknowledgments This work was supported by the French Research Agency (ANR). Ismaël Riedel is funded by the MAIF Foundation. INSEE data were prepared and provided by the Centre Maurice Halbwachs (CMH).

Appendix

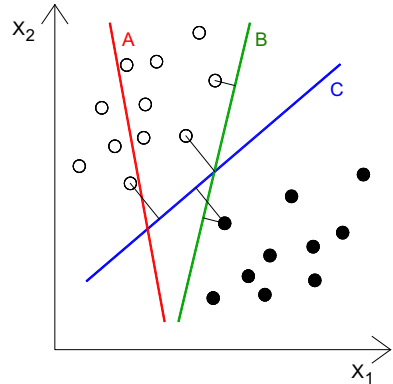
Support vector machine definitions (SVM)

For the sake of simplicity, a formal definition of the linear binary case is first presented. The nonlinear case (still binary) is then studied. At last, the multiclass case is considered (n -class classification problem). Definitions are built following Teukolsky et al. (2007) and Cortes and Vapnik (1995).

Linear classification

Before entering into the mathematical definitions, a qualitative graphical description will help understanding the basic foundation of the method. Given some data points belonging to one of two classes (binary problem), viewed as p -dimensional vectors (a list of

Fig. 17 Different separating hyperplanes. A does not separate the classes. B does, but only with a small margin. C separates them with the maximum margin



p numbers) for SVM, many planes might exist that classify the data (Fig. 17). Intuitively, a good separation is achieved by the plane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin is, the lower the generalization error of the classifier. Therefore, the basic idea is to choose the plane so that the distance from it to the nearest data point on each side is maximized.

Given some training data D , a set of points of the form

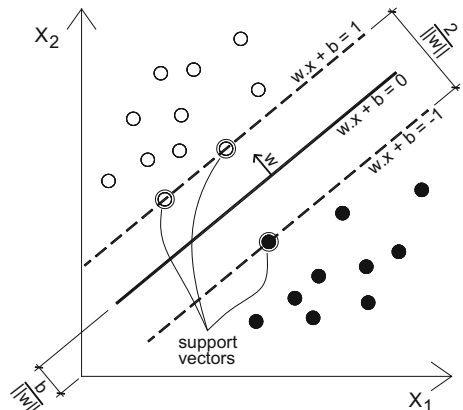
$$D = \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in \mathbb{R}^p, \quad y_i \in \{-1, 1\}\}_{i=1}^n$$

where the y_i is either 1 or -1 , indicating the class to which the point \mathbf{x}_i belongs. Each \mathbf{x}_i is a p -dimensional real vector. We want to find the maximum-margin hyperplane that divides the points having $y_i = 1$ from those having $y_i = -1$. Any hyperplane can be written as the set of points \mathbf{x} satisfying

$$\mathbf{w} \cdot \mathbf{x} + b = 0$$

where \cdot denotes the dot product and \mathbf{w} the normal vector to the hyperplane. The parameter $\frac{b}{\|\mathbf{w}\|}$ determines the offset of the hyperplane from the origin along the normal vector \mathbf{w} (Fig. 18). If the training data are linearly separable, we can select two hyperplanes in a way that they separate the data and there are no points between them, and then try to maximize

Fig. 18 Maximum-margin hyperplane and margins for an SVM after training with samples from two classes. Samples on the margin are called the support vectors



their distance. The region bounded by them is called “the margin.” These hyperplanes can be described by the equations (see Fig. 18)

$$\mathbf{w} \cdot \mathbf{x} + b = 1 \text{ and } \mathbf{w} \cdot \mathbf{x} + b = -1$$

By using geometry, we find the distance between these two hyperplanes is $\frac{2}{\|\mathbf{w}\|}$, so we need to minimize $\|\mathbf{w}\|$. As we also have to prevent data points from falling into the margin, we add the following constraint: for each i either

$$\begin{aligned} \mathbf{w} \cdot \mathbf{x}_i + b &\geq 1 && \text{for } \mathbf{x}_i \text{ of the first class, or} \\ \mathbf{w} \cdot \mathbf{x}_i + b &\leq -1 && \text{for } \mathbf{x}_i \text{ of the second class} \end{aligned}$$

This can be rewritten as

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \quad \text{for all } 1 \leq i \leq n$$

The optimization problem is then posed as:

$$\text{Minimize}(\mathbf{w}, b) \|\mathbf{w}\|; \quad \text{subjected to (for any } i = 1, \dots, n) y_i(\mathbf{w} \cdot \mathbf{x}_i) \geq 1$$

To simplify the problem, it is possible to alter the equation by substituting $\|\mathbf{w}\|$, the norm of \mathbf{w} , with $\frac{1}{2} \|\mathbf{w}\|^2$ without changing the solution (the minimum of the original and the modified equation has the same \mathbf{w} and b). This is a quadratic programming optimization problem.

$$\text{Minimize}(\mathbf{w}, b) \frac{1}{2} \|\mathbf{w}\|^2; \quad \text{subjected to (for any } i = 1, \dots, n) y_i(\mathbf{w} \cdot \mathbf{x}_i) \geq 1$$

In mathematical optimization, the method of Lagrange multipliers is a strategy for finding the local maxima and minima of a function subject to equality constraints.

By introducing Lagrange multipliers α , the previous constrained problem can be expressed as

$$\min_{\mathbf{w}, b} \max_{\alpha \geq 0} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1] \right\}$$

This problem can now be solved by standard quadratic programming techniques and programs. The “stationary” *Karush–Kuhn–Tucker* condition implies that the solution can be expressed as a linear combination of the training vectors

$$\mathbf{w} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i$$

Only a few α_i will be greater than zero. The corresponding \mathbf{x}_i is exactly the *support vector* that lies on the margin and satisfies

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1$$

From this, we can derive that the support vectors also satisfy

$$\mathbf{w} \cdot \mathbf{x}_i + b = \frac{1}{y_i} = y_i \Leftrightarrow b = \mathbf{w} \cdot \mathbf{x}_i - y_i$$

which allows defining the offset b . In practice, it is more robust to average over all support vectors N_{sv}

$$b = \frac{1}{N_{sv}} \sum_{i=1}^{N_{sv}} (\mathbf{w} \cdot \mathbf{x}_i - y_i)$$

A modified maximum-margin idea was proposed, allowing for mislabelled examples. If there exists no hyperplane that can split the examples (some points may fall within the margins), the *Soft Margin* method will choose a hyperplane that splits the examples as cleanly as possible, while still maximizing the distance to the nearest cleanly split examples. The method introduces slack variables ζ_i , which measure the degree of misclassification of the data x_i .

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \zeta_i \quad 1 \leq i \leq n$$

The optimization becomes a trade-off between a large margin and a small error penalty. The final equation leads to a quadratic programming solution. The membership decision rule is based on the sign function, and the classification is done by $y_{new} = \text{sgn}(\mathbf{w} \cdot \mathbf{x}_{new} + b)$ where (\mathbf{w}, b) are the hyperplane parameters found during the training process, and x_{new} is an unseen sample.

Nonlinear classification

In addition to performing linear classification, SVMs can efficiently perform nonlinear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. For machine learning algorithms, the kernel trick is a way of mapping observations from a general set S into an inner product space V , in the hope that the observations will gain meaningful linear structure in V . Linear classifications in V are equivalent to generic classifications in S . The trick to avoid the explicit mapping is to use learning algorithms that only require dot products between the vectors in V , and choose the mapping such that these high-dimensional dot products can be computed within the original space, by means of a kernel function. The resulting algorithm is formally similar, and the maximum-margin hyperplane can be fitted in the transformed feature space. The transformation may be nonlinear, and the transformed space was high dimensional; therefore, even if the classifier is a hyperplane in the high-dimensional feature space, it may be nonlinear in the original input space (Fig. 19). There exist several choices of kernel

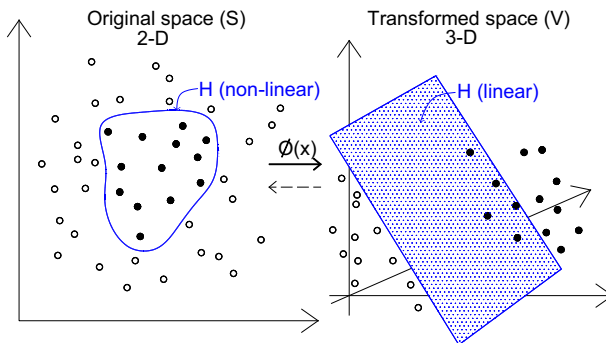


Fig. 19 Kernel machine. The separation surface can become linear when feature vectors are mapped in a high-dimensional space (here 3D—right) while it may be nonlinear in the original input space (here 2D—left)

function k . The Kernel is related to the transform $\phi(x_i)$ by the equation $k(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$.

Generally, the Gaussian kernel is a common good choice $k(x_i, x_j) = \exp\left(-\frac{1}{2}|\mathbf{x}_i - \mathbf{x}_j|^2/\sigma^2\right)$, and it proved to give the best results in our study. Therefore, the classifications in this work are done using this kernel.

Multiclass SVM

Even if SVM are intrinsically binary classifiers, in practice several-classes classifications are usually of interest. Different multiclass classification strategies can be adopted, based on the binary analysis or the less used “all-together” method. The former is the dominant approach, which reduces the single multiclass problem into multiple binary classification problems and can be of the form (among others):

One versus all

Involves training N different binary classifiers, each one trained to distinguish the data in a single class from the data in all remaining classes. Classification of new instances is done by a winner-takes-all strategy, in which the classifier with the highest output function assigns the class.

One versus one

Builds binary classifiers that distinguish between every pair of classes. Classification is done by a max-wins voting strategy, in which every classifier assigns the instance to one of the two classes, then the vote for the assigned class is increased by one vote, and finally, the class with the most votes determines the instance classification. The one-versus-one classification proved to be more robust in the majority of cases, and showing the best results is the one selected in our study.

References

- Agrawal R, Imieliński T, Swami A (1993) Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD international conference on Management of data—SIGMOD '93, p 207. doi:[10.1145/170035.17007](https://doi.org/10.1145/170035.17007)
- Bard PY, Duval AM, Bertrand E, Vassiliadès JF, Vidal S, Thibault C, Guyet B, Mèneroud JP, Gueguen P, Foin P, Dunand F, Bonnefoy-Claudet S, Vettori G (2005) Le risque Sismique à Nice: apport méthodologique. résultats et perspectives opérationnelles. Final technical report of the GEMGEP project. CETE-Méditerranée
- Baroux E, Pino NA, Valensise G, Sotci O, Cushing ME (2003) Source parameters of the 11 June 1909. Lambesc (Provence. southeastern France) earthquake: a reappraisal based on macroseismic, seismological, and geodetic observations. *J Geophys Res* 108(B9):2454. doi:[10.1029/2002JB002348](https://doi.org/10.1029/2002JB002348)
- Borsi B, Dell'Acqua F, Faravelli M, Gamba P, Lisini G, Onida M, Polli D (2010) Vulnerability study on a large industrial area using satellite remotely sensed images. *Bull Earthq Eng* 9:675–690. doi:[10.1007/s10518-010-9211_9](https://doi.org/10.1007/s10518-010-9211_9)
- Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: Haussler D (ed) 5th annual ACM workshop on COLT. ACM Press, Pittsburgh, pp 144–152
- Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20:273–297

- Duin RPW, Juszczak P, Paclik P, Pekalska E, de Ridder D, Tax DMJ, Verzakov S (2007) PRTTools 4.1, a Matlab toolbox for pattern recognition. Delft University of Technology, Delft
- Dunand F, Gueguen P (2012) Comparison between seismic and domestic risk in moderate seismic hazard prone region: the Grenoble City (France) test site. *Nat Hazards Earth Syst Sci* 12:511–526. doi:10.5194/nhess-12-511-2012
- Geiss C, Taubenböck H (2012) Remote sensing contributing to assess earthquake risk/from a literature review towards a roadmap. *Nat Hazards* 68:7–48. doi:10.1007/s11069-012-0322-2
- Grunthal G, Levret A (2001) L'échelle macrosismique européenne. European macroseismic scale 1998(EMS-98) Conseil de l'Europe—Cahiers du Centre Européen de Géodynamique et de Séismologie, vol 19
- Gruppo Nazionale per la Difesa dai Terremoti, Roma (GNDT) (1993) *Rischio Sismico di Edifici Pubblici*. Consiglio Nazionale delle Ricerche
- Gueguen P (2013) *Seismic vulnerability of structures*. Civil engineering and geomechanics series. ISTE Ltd and Wiley, London. ISBN 978-1-84821-524-5. Edited by Philippe Gueguen
- Gueguen P, Michel C, LeCorre L (2007) A simplified approach for vulnerability assessment in moderate-to-low seismic hazard regions: application to Grenoble (France). *Bull Earthq Eng* 2007(5):467–490. doi:10.1007/s10518-007-9036-3
- Hamaina R, Leduc T, Moreau G (2012) Towards urban fabrics characterization based on building footprints. In: Gensel J et al (eds) *Bridging the geographic information sciences, lecture notes in geoinformation and cartography*. doi:10.1007/978-3-642-29063_18
- HAZUS (1997) *Earthquake loss estimation methodology*, Hazus technical manuals. National Institute of Building Science, Federal Emergency Management Agency (FEMA), Washington
- Jackson J (2006) Fatal attraction: living with earthquakes. The growth of villages into megacities and earthquake vulnerability in the modern world. *Philos Trans R Soc* 364(1845):1911–1925
- Kappos AJ, Panagopoulos G, Panagiotopoulos C, Penelis G (2006) A hybrid method for the vulnerability assessment of R/C and URM buildings. *Bull Earthq Eng* 4(4):391–413
- Lagomarsino S, Giovinazzi S (2006) Macroseismic and mechanical models for the vulnerability and damage assessment of current buildings. *Bull Earthq Eng* 2006(4):415–443. doi:10.1007/s10518-006-9024-z
- Lambert J (1997) *Les tremblements de terre en France: hier, aujourd'hui, demain*. BRGM Eds. Orléans, France
- Michel C, Guéguen P, Causse M (2012) Seismic vulnerability assessment to slight damage based on experimental modal parameters. *Earthq Eng Struct Dyn* 41(1):81–98. doi:10.1002/eqe.1119
- Pierre J-P, Montagne M (2004) The 20 April 2002, Mw 5.0 Au Sable Forks, New York, earthquake: a supplementary source of knowledge on earthquake damage to lifelines and buildings in Eastern North America. *Seismol Res Lett* 75(5):626–635
- Riedel I, Gueguen P, Dunand F, Cottaz S (2014) Macro-scale vulnerability assessment of cities using association rule learning. *Seismol Res Lett* 85(2):295–305. doi:10.1785/0220130148
- Rothé J-P, Vitart M (1969) Le séisme d'Arette et la séismicité des Pyrénées. 94e Congrès national des sociétés savantes, Pau, sciences, t. II, pp 305–319
- Scotti O, Baumont D, Quenet G, Levret A (2004) The French macroseismic database SISFRANCE: objectives, results and perspectives. *Ann Geophys* 47(2/3):571–581. doi:10.4401/ag-3323
- Spence R, Lebrun B (2006) Earthquake scenarios for European cities: the risk-UE project. *Bull Earthq Eng* 4 special issue
- Spence R, So E, Jenny S, Castella H, Ewald M, Booth E (2008) The global earthquake vulnerability estimation system (GEVES): an approach for earthquake risk assessment for insurance applications. *Bull Earthq Eng* 6:463–483. doi:10.1007/s10518-008-9072-7
- Spence R, Foulser-Piggott R, Pomonis A, Crowley H, Guéguen P, Masi A, Chiauzzi L, Zuccaro G, Cacace F, Zulifikar C, Markus M, Schaefer D, Sousa ML, Kappos A (2012) The European building stock inventory: creating and validating a uniform database for earthquake risk modeling and validating a uniform database for earthquake risk modeling risk modeling. In: *The 15th world conference on earthquake engineering*, Sept 2012, Lisbon, Portugal
- Teukolsky SA, Vetterling W, Flannery BP (2007) *Section 16.5. Support vector machines*. Numerical recipes: the art of scientific computing, 3rd edn. Cambridge University Press, New York. ISBN 978-0-521-88068-8
- Wald DJ, Quitoriano V, Heaton TH, Kanamori H, Scrivner CW, Worden CB (1999) Trinet “shakes maps”: rapid generation of peak ground motion and intensity maps for earthquakes in Southern California. *Earthq Spectra* 15(3):537–555
- Worden CB, Wald DJ, Allen TI, Lin K, Garcia D, Cua G (2010) A revised ground motion and intensity interpolation scheme for shakemap. *Bull Seismol Soc Am* 100(6):3083–3096