Towards a bayesian seismotectonic zoning for use in Probabilistic Seismic Hazard Assessment (PSHA)

Boris Le Goff*, Delphine Fitzenz† and Céline Beauval**

*CGE, University of Évora, Portugal – email:boris@uevora.pt
†CGE, University of Évora, Portugal – email: delphine@uevora.pt
**LGIT-IRD , France – email: celine.beauval@obs.ujf-grenoble.fr

Abstract. The mathematical representation of seismic sources is an important part of probabilistic seismic hazard assessment. It reflects the association of the seismicity with the tectonically-active geological structures evidenced by seismotectonic studies. Given that most active faults are not characterized well enough, seismic sources are generally defined as areal zones, delimited with finite boundary polygons, within which the geological features of active tectonics and the seismicity are deemed homogeneous (e.g., focal depth, seismicity rate, and maximum magnitude). Besides the lack of data (e.g., narrow range of recorded magnitudes), the application of this representation generates different problems: 1) a large sensitivity of resulting hazard maps on the location of zone boundaries, while these boundaries are set by expert decision; 2) the zoning can not represent any variation in faulting mechanism; 3) the seismicity rates are distributed throughout the zones and we lose the location of the determinant information used for their calculation. We propose an exploratory study for an alternative procedure in area source modeling. First, different data (e.g., geomorphology, geology, fault orientations) will be combined by using automated spatial partitioning (investigation of both supervised and unsupervised methods) in order to obtain several information classes, which may be defined as areal source zones. Then, a given hypocenter belonging to a given "zone", from now on called seismicity model, will be expressed by a probability computed from the 2D (spatial) probability density function (pdf) for the active tectonic model used as an a priori and updated with specific data from seismicity catalogs (e.g., focal mechanism) or other new data sources (e.g., geomorphology, subsurface exploration). This hypocenter will thus be allowed to contribute to several models, with weights given by the value of the pdf for each model. The annual rate of occurrence, for a given model, will be calculated by the weighted average of the different hypocenter contributions contained in this model. Future applications will couple the seismicity models to Ground Motion Prediction Equations. In consequence, the results will provide the full spectrum of variability in the hazard and will highlight zones badly constrained and deserving to be more studied.

Keywords: Seismotectonic zoning, Probabilistic Seismic Hazard Assessment, propagation of uncertainties, Segmentation

1. INTRODUCTION

Several studies have shown the determinant impact of seismotectonic zoning in Probabilistic Seismic Hazard Assessment (PSHA) (Beauval, 2003; Beauval and Scotti, 2004; Bender, 1986; Woo, 1996). This technique allows to link the seismicity with the tectonically-active geological structures, in order to define sources for use in PSHA computation. Usually, and because faults are often not characterized well-enough, source zones are defined as surfaces and modeled as polygons. In that case, they are delimited with fixed, infinitely thin boundaries. In each zone, the geological expression of active tectonics and the seismicity are deemed homogeneous (e.g., focal depths and mechanisms, seismicity rate, and maximum magnitude), and each point of a zone is considered an equally likely source of earthquake.

Besides the lack of data (e.g., short period of instrumental observation of small events, short catalogue of large events, blind faults), the establishment of a traditional seismotectonic zoning generates different shortcomings. The finite boundaries of the different zones are set by expert decisions, leading to different problems : 1) the superposition of a resulting hazard map with areal source zoning model underlines the large sensitivity of the results to this method (Beauval, 2003); 2) it is not reproducible: different experts come up with different zonings using the same input data. 3) the final seismotectonic zoning is not provided with error maps reflecting the original density of information used for both the assessment of the common characteristics and the calculation of seismicity rates of each zone. 4) Moreover, the zoning does not account for any variation in faulting mechanisms with depth or for conjugate sets of faults.

This paper aims to propose an exploratory study for alternative procedures in area source modeling. We search to obtain a method which will be robust and reproducible. In this way, different data (e.g., geomorphology, geology, fault orientations) will be combined by using automated spatial partitioning in order to obtain several information classes, which may be defined as areal source zones. Supervised and unsupervised approaches are usually used in remote sensing and we will investigate how to adapt them to integrate data of different dimensions and to obtain soft boundaries. Then we will associate the hypocenters to the previously defined areal zones. Using Bayesian methods, a hypocenter will contribute in the calculation of annual rate occurrence of several zones.

2. INPUT DATA

Different data sets are used in order to establish a seismotectonic zoning. These data are of different kinds and are characterized by different dimensions.

The structural geology map contains much information about faults (e.g., fault orientations, fault geometry, fault type, tectonic ensembles) and has an important role in the development of seismotectonic zoning.

Two different seismicity catalogs are used in seismotectonic zoning computations : historical and instrumental. The idea is 1) to locate the hypocenters and 2) to compute the focal mechanisms and magnitudes to try and image active faults and characterize the type of faulting (strike-slip, or dip-slip faulting). The historical catalog is built by collecting stories/reports of subjective experiences (was the shaking enough to wake people up, did the building sway, questionnaires can be found in online forms from seismological surveys around the world, e.g., Did you feel it?, U.S. Geological Survey, or BCSF), observations of damages to buildings and perturbations to the landscape (liquefaction, stream offsets, etc) observed after an earthquake. These observations allow to assess an intensity of the seismic event in a predefined scale (e.g., MMI, EMS-98). Intensities are subjective measures by definition. The data points (density and location) depend namely on the distribution of the population at the time of the earthquake. They will be used to assess the size and depth of the events and to try and allocate the event to known active faults. Small or deep events are not likely to be represented in historical catalogs. Depending on the history of a region, it may cover 200 years (since the Gold Rush in California for example) or 2000 years (in the middle east). The instrumental catalog is obtained from seismic signals, recorded by seismic networks. The density of the network will affect directly the quality of the localization of the source and size of the events. Different magnitude scales exist to quantify the energy released by an earthquake. These catalogues contain uncertainties, notably with hypocenter locations and magnitude calculations. Because the evolution of seismic networks is recent (the first dense networks were dedicated to "listening" to nuclear tests by other countries, e.g., the LDG in France, since 1960), the period of instrumental observation of small events is short.

Other data are useful in the conception of seismotectonic zoning in order to have a better constrained model. Indirectly, geological maps, expressing the age of the formations, trenches, geophysical data and digital elevation model (DEM) allow to improve the location of faults.

An example of such composite datasets is shown in figure 1.

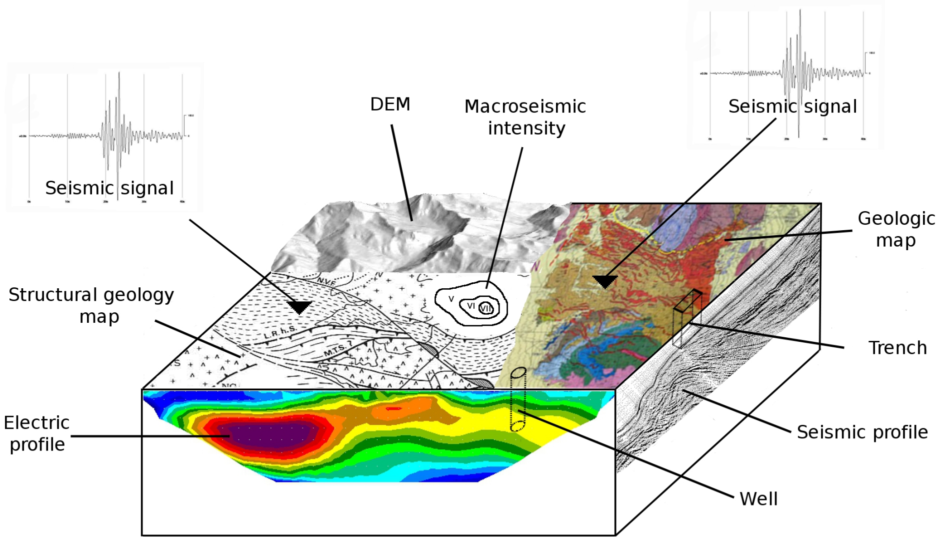

FIGURE 1. 3D diagram representing the useful data for using in PSHA. In this figure, the segmentation is done to show the kinds of data and do not express a seismotectonic zoning

In conclusion, we have 2D maps with features such as 1D fault lines, or dip angles; 3D points (hypocenters resulting from the inversion of seismograms), points on a surface with intensity data; borehole or seismic transects intersecting faults.

Bayesian data integration has been used in hydrogeology (Ezzedine, 1999), but not for seismic source modeling.

3. PSHA

There are several approaches to assess seismic hazard in a probabilistic sense. The most widely used is the Cornell-McGuire approach (Cornell, 1968; McGuire, 1976). It consists in estimating the probability of exceedance of a ground-motion target level at a site. Usually, this target level is described by the Peak Ground Acceleration (PGA) or the Peak Ground Velocity (PGV). Considering a Poissonian model ocurrence, it is common to refer to the return periods (inverse of the annual rate) instead of the annual rates. According to the application domain, these return periods range from 100 to 10^7 years.

A number of steps needs to be completed on the seismicity data before the actual PSHA can take place (Le Goff and al., 2009). First, the seismic catalogues have to be homogenized in a same magnitude scale. Indeed, the intensity scale used in historical catalogues has to be converted into a magnitude scale and instrumental catalogues can also use different magnitude scales. Usually, these conversions generate important uncertainties.

Second, the seismicity is usually modeled as a superposition of independent events (with a Poissonian distribution of inter-event times) and their related foreshocks and aftershocks (with a frequency that decreases with time before/after the main shock following well-known empirical relationships (Omori, 1894) and a frequency-size distribution following an empirical power-law, the Gutenberg-Richter law). In consequence, catalogues have to be filtered of their foreshocks and aftershocks, in order to model mainshock ocurrences only.

Third, from all available seismic events and seismotectonic data, a seismotectonic zoning is achieved in order to identify and characterize the source of seismicity. Inside each source zone,the Gutenberg-Richter model is assumed and the corresponding parameters are calculated (b-values, i.e., the slopes of the power law) and a maximum magnitude is estimated.

A ground motion prediction model is used to estimate exceedance probability of a target level and for a given magnitude/distance. Finally, the exceedance probability contributions of all couples magnitude/distance are summed.

Beyond random (aleatory) uncertainties, intrinsic of data (e.g., catalogue uncertainties), epistemic uncertainties are generated when the choice of different parameters or models is done. Sensitivity studies have shown parameters which have the most important impact in the estimation of probabilistic seismic hazard assessment. Beauval (2003, 2004) has demonstrated that both the truncation of the predicted ground-motion model and the choice of the magnitude-intensity correlation are dominant with respect to the level of hazard inferred. In terms of spatial distribution of hazard, the impact of the location of seismotectonic zoning boundaries is what matters most.

Usually, these epistemic uncertainties are accounted for using a logic tree approach where each branch is weighted arbitrarily by an expert or a panel of experts. The underlying assumptions are 1) the branches stemming from a given node in the logic tree represent completely distinct solutions (mutually exclusive), 2) all the branches stemming from a node represent all the possible solutions (collectively exhaustive), and that each level in the tree is independent from the level before (sequentially independent), meaning that the processes modelled by them are decoupled (Bommer and Scherbaum, 2008).

Several parts of this approach are disputable, such as the earthquake occurrence model or the predicted ground motion model. The choice to focus this study on the seismotectonic zoning is due to by the impact of its source zone boundaries into the hazard map. Moreover, the currentl seismotectonic zoning model does not allow to represent any variation in faulting mechanism. The assessment of seismotectonic zoning is not reproducible. Indeed, different experts, using same data, provide different zoning, based on their differing interpretation.

As reminded by an INGV open letter, the best approach to protect population and building from collapsing is not through earthquake prediction but through the application of appropriate safety measures. The development of seismic hazard maps provides the specifications required by building codes to avoid collapse of buildings and the resulting fatalities, and the information to convey to the population the basic concepts of earthquake hazard, awareness, preparedness and response.

4. METHODOLOGY

The exploratory approach described in this paper aims to obtain a reproducible method for the assessment of seismotectonic zoning. This method has as objective to avoid the abrupt changes in the seismicity rates at the source zones boundaries. It will allow to express the variability of seismic parameters (e.g., faulting mechanisms, faults orientations, maximum fault length). By propagating uncertainties, it will be possible to highlight the zones poorly constrained and deserving to be more studied.

The present approach will investigate how to adapt supervised and unsupervised methods, usually used in remote sensing (MacMillan, 2004) to seismotectonic zoning.

Foremost, it is important to differentiate three kinds of partitioning :

1) classification consists in labeling each data point independently, without using their spatial properties

2) regularized classification produces smoother and more consistent results by taking into account the spatial structure of the data.

3) region-based segmentation allows to partition the space into spatially consistent classes with one class per region.

Two approaches may be considered to apply these methods : supervised and unsupervised methods. In a supervised approach, statistical parameters (e.g., mean, variance, number of classes) are determined for predefined classes, using a training set made of expert-labeled data. To learn class parameters, a training step is executed. Then, the algorithm recognizes data which possess similar parameters to those introduced in the training step. It is possible to estimate the quality of the supervised classification with a confusion matrix. Each row of the matrix represents the instances in a predicted class, while each column represents the instances in an actual class.

In an unsupervised approach, statistical parameters are not known, and there is no training data. The number of classes is not known a priori. The parameters have to be learned from the data.

Automated spatial partitioning will allow to associate different data into several information classes, which may be considered as areal source zones.

Weatherill and Burton (2009) propose an approach of partitioning with K-means algorithms. It consists in partitioning regions from the location of observed seismicity. This method is reproducible, considering that the initial cluster centroids are the same, and underlines the problems of the seismotectonic zoning. However, the K-means method has problems determining the optimal number of clusters and defining of initial cluster centers and leads to a local optimum instead of a global optimum partition. Moreover, this approach puts away geologic and seismotectonic data. This assumption is not consistent with a seismotectonic zoning. Indeed, two hypocenters may be close to the mean cluster position and possess completely opposite geologic and seismotectonic features. The location of seismic event hypocenters is not an exclusive factor of seismotectonic zoning assessment.

The method, which will be investigated, will use a segmentation with disjoint classes. A supervised approach will be considered with the following a priori class parameters: faulting mechanisms, fault orientations, seismic event depth and maximum length of faults. A background class will be introduced for regions where the lack of data does not allow to choose one of the previous classes. Then, Bayesian inference will be used in order to identify regions which possess similar parameters those defined in the characterized classes. A generative model is used for randomly generating observable data. It is illustrated by a simplified graphical model (Figure 2), expressing the relationships between all the ingredients of the problem.

In this graphical model, each node represents a set of random variables. A black arrow, for example pointing from the seismotectonic zoning to the node structural map/seismicity catalogues, means that the structural map and seismicity catalogues are generated from the seismotectonic zoning. This relation brings into the model a conditional probabilistic density function (pdf). The graphical model expresses that the parameters of the zoning allow to predict surface faults traces, hypocenter locations, fault expressions in geophysical profile. This information, coupled with seismic signals and velocity model, allows to generate the structural map and the seismicity catalogues. The red arrow represents the Bayesian inference, which consists to estimate the

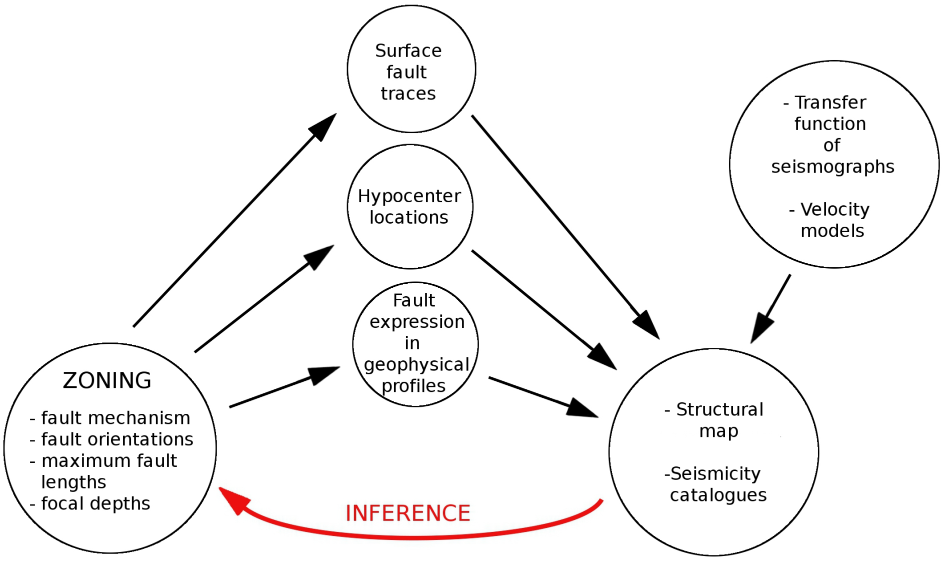

FIGURE 2. Graphical model for generating observable data. Black arrows represent interactions of the direct model. The red arrow represents the inference, which allows to estimate the number of geographic zones and their parameters, from the structural map and the seismicity parameters.

number of geographic zones and their parameters (e.g., mean, variance) knowing the structural map and the seismicity catalogues.

The resulting seismotectonic zoning will be determined from a supervised spatial partitioning. Boundaries of seismotectonic zoning will be fuzzy, because a point will be defined with a probability to belong to a zone and one other probability to belong to the next one. It will be possible to delimit, for example, zones with 95 % of confidence, allowing to isolate regions where the model is well-constrained. On the other hand, regions where studies are needed to improve the model, may be identified. In order to express the variability of parameters, defined in the classes (e.g., faulting mechanisms, faults orientations, maximum fault length), the seismotectonic zoning may be decoupled in layers, each one reflecting a particular parameter, used for the partitioning. For example, a zone containing two faulting mechanisms type will be decoupled in two layers, each one expressing one type of faulting mechanism. Moreover, this decoupling will allow to not lose the dominant information before the seismicity rate calculation.

In order to avoid the abrupt change in the seismicity rate at source boundaries, observable in the resulting hazard map, Bender (1986) proposes to use standard "hard" seismotectonic zoning but to provide smoothing using the epicenter location uncertainty, considered normally distributed. A point source will contribute to the seismicity rate calculation of several zones, in function of its epicenter location uncertainty. In that paper, the assumption on the location uncertainty characterization, arbitrary defined by 0, 10, 25 and 50 kilometers, is disputable. The actual location uncertainty depends on each seismic event. The value of 50 km is considered for historical earthquake and can not reflect the uncertainty of an instrumental seismic event. Instrumental seismic event uncertainty depends on its location (compared to the seismic network) and its date (improvement of seismic station accuracy and network density).

Wesson and al. (2003) propose a methodology, based on Bayesian inference, to associate earthquakes and faults. This method may be interesting to the seismicity rate calculation. A restricting factor is the requirement of exhaustive data (e.g., slip rates of recognized faults), limiting this application to regions where the velocity of plate tectonics is high.

Concerning the seismicity rate calculation, the method described in this paper will investigate two approaches. First, the achievement of the seismotectonic zoning, using partitioning with spatial information, allows to obtain fuzzy source zone boundaries. It is possible, with a supervised method, to define an additional class containing the seismicity. Thus, the seismicity rates will be smoothed into the whole zone and will not change abruptly at source zone boundaries.

Second, the method used by Bender may be adapted in order to calculate the seismicity rates. In our method, the uncertainty, liable to each seismic event instead of arbitrary defined for all events, will be used to define the contribution of this source to several zones. The fact that a source, close to a seismotectonic zoning boundary, may participate in the calculation of several seismicity rates allows to smooth these rates in source zone boundaries. Thus, in proximity to a source zone boundary, resulting acceleration levels for two close sites may not differ considerably.

These approaches will be achieved for the different decoupled layers to keep the dominant information before the hazard calculation. It will be possible, afterward, to establish an hazard uncertainty map. It is important to note that the method will be able to update the seismotectonic zoning and seismicity rates when new information will be available.

5. DISCUSSION

During about 20 years, the evolution of PSHA in current practice has been slowed down by using paying software within which code source is not known and shared. But recently, initiatives have been undertaken, first in California with openSHA and after in international-wide with GEM, to propose an open source code, making it easier to incorporate new methodologies. The methodology, presented in this paper, could therefore be inserted in the hazard calculation process.

The intervention of panels of experts, that was used in first attempts at assessing epistemic uncertainties, need to be reduced to go towards reproducible decisions based on traceable information.

With our contribution, we show how Bayesian Inference would be useful in Probabilistic Seismic Hazard Assessment. Other authors have pointed that out in the recent past, for renewal models (Biasi and Weldon, 2008; Fitzenz and al, 2007), or for a general view (Esmer, 2006). We are showing a very important and very concrete application, the questions of spatial partitioning. We hope to raise the interest of the MaxEnt community and gather suggestions or start collaborations. We advocate that the PSHA community would benefit from more frequent interaction with MaxEnt-related communities and that the latter would find interesting new applications to work on.

6. CONCLUSION

A key difficulty in PSHA is the representation of seismic sources. The method described in this paper is in the exploratory stage and propose an alternative proceeding to the seismic hazard calculation. It aims to overcome the problem of the large sensitivity of resulting hazard maps on the location of zone boundaries and to represent the variability of seismicity parameters, with bayesian method. This approach allows to conserve the dominant information used to characterize both seismotectonic zoning and seismicity rates, allowing to establish an hazard uncertainty map.

This approach integrates all data, however few data we have, with their different dimensions, from partitioning methods usually not used in PSHA.

This method is applicable in different regions, whether slowly or highly deformation. This method will be applied in the Lower Tagus Valley (Lisbon, Portugal), where seismotectonic data are poorly constrained.

ACKNOWLEDGEMENTS

This research and BLG are funded by the Fundação para a Ciência e a Tecnologia (PTDC/CTE-GIX/101852/2008 and FCOMP-01-0124- FEDER-009326). DF benefits from the FCT Ciência 2007 program. We thank André Jalobeanu for his guidance in data partitioning and Bayesian methodology.

REFERENCES

1. Beauval, C. , 2003. Analyse des incertitudes dans une estimation probabiliste de l alea sismique, exemple de la France, *Ph.D. thesis*
2. Beauval, C.and Scotti,O., 2004. Quantifying Sensitivities of PSHA for France to earthquake catalogue uncertainties, truncation of ground-motion variability, and magnitude limits, *Bull. seism. Soc. Am.*, *94* 1579-1594
3. Bender, B., 1986. Modelling source zone boundary uncertainty in seismic hazard analysis, *Bull. seism. Soc. Am.*, *76* 329-341
4. Biasi, G. P., and Weldon, R., 2008. San Andreas fault rupture scenarios from multiple paleoseismic records: Stringing pearls: in Working Group on California Earthquake Probabilities, Version 2 (UCERF2) *U.S.G.S., Appendix E: Open-File Report 2007-1437 and California Geological Survey Special Report 203*
5. Bommer J.J., Scherbaum F., 2008. The Use and Misuse of Logic Trees in Probabilistic Seismic Hazard Analysis *EARTHQ SPECTRA*, *24* 997-1009, ISSN:8755-2930
6. Cornell, C.A., 1968. Engineering seismic risk analysis *Bull. seism. Soc. Am.*, *58* 1583-1606
7. Esmer, O., 2006. Information Theoric Framework for the Earthquake Recurrence Models : Methodica Firma per Terra Non-Firma ; in A. Mohammad-Djafari *Bayesian Inference and Maximum Entropy Methods in Science and Engineering , 26th International Workshop, 8-13 July 2006, Paris American Institute of Physics (AIP) 872* 556-562
8. Ezzedine, S., Y. Rubin, and J. Chen (1999), Bayesian method for hydrogeological site characterization using borehole and geophysical survey data: Theory and application to the lawrence livermore national laboratory superfund site., *Water Resour. Res.*, *35*(9), 2671–2683.
9. Fitzenz, D. D., A. Jalobeanu, and S. H. Hickman, 2007. Integrating laboratory creep compaction data with numerical fault models: A Bayesian framework *J. Geophys. Res. 112* B08410, doi:10.1029/2006JB004792
10. Le Goff B., Bertil D., Lemoine A ., Terrier M., 2009. Systemes de failles de Serenne et de la Haute-Durance (Hautes-Alpes) : évaluation de l aléa sismique. *Rapport BRGM RP-57659-FR, 242 p, 85 ill., 15 Tab, 10Ann.*
11. McGuire, R. K., 1976. Fortran computer program for seismic risk analysis *U.S. Geol. Survey Open-File Report 76-67.*
12. MacMillan R. A. (Bob), 2004. Automated knowledge-based classification of landforms, soils and ecological spatial entities *open file report USDA*
13. Omori, F., 1894. *J. Coll. Sci. Imper. Univ. Toky* 7111
14. Weatherill, G., Burton, P.W., 2009. Delineation of shallow seismic source zones using K-means cluster analysis, with application to the Aegean region. *Geophysical Journal International 176* 565-588.v
15. Wesson, R. L., Bakun, W. H., and Perkins, D. M., 2003. Association of Earthquakes and Faults in the San Francisco Bay Area Using Bayesian Inference. *Bulletin of the Seismological Society of America 93(3)* 1306-1332. doi: 10.1785/01200200855
16. Woo, G., 1996. Kernal estimation methods for seismic Hazard area source modelling. *Bulletin of the Seismological Society of America 86* 353-362