Bayesian estimation and selection of seismic source models

Merlin Keller[1*], Sanaa Zannane[1], Clara Duverger[2,5], Gloria Senfaute[3] and Jessie Mayor[4]

[1]EDF R&D, Street, Chatou, 100190, France.
[2]CEA, DAM, DIF, Arpajon, F-91297, France.
[3]EDF R&D, Street, Saclay, 100190, France.
[4]EDF TEGG, Street, Aix-en-Provence, 100190, France.
[5]now at Lithium de France, 16 rue des Couturières, Bischwiller, 67500, France.

*Corresponding author(s). E-mail(s): merlin.keller@edf.fr;
Contributing authors: sanaa.zannane@edf.fr;
clara.duverger@cea.fr; gloria.senfaute@edf.fr; jessie.mayor@edf.fr;

Abstract

In the context of probabilistic seismic hazard analysis (PSHA), we propose a full Bayesian methodology to estimate seismic recurrence parameters of area source models, which are subdivisions of a region of interest into zones assumed homogeneous in terms of seismic activity rate, and to discriminate between several models based on the data provided by an earthquake catalogue. For each candidate source model, our method outputs fast and accurate estimations of both the joint posterior distribution of the model's uncertain parameters, and the model's marginal likelihood. We then extend the approach to zone clustering, by allowing zones to be merged in all possible ways. In this new setting, we propose a simple Gibbs sampling algorithm to efficiently explore and optimize all possible clusterings. We first apply the proposed methodology to a toy example, and then to the Metropolitan French context, where at least four national competitive and published area source models are used by engineers and researchers for seismic hazard assessment. Our results show that current seismotectonic models are over-parameterized, in that they contain too many zones with respect to the limited amount of data available in low seismic regions, such as the Continental French territory.

Clustering zones using the Gibbs sampler allows to obtain more accurate recurrence parameter estimates, and points to several promising research avenues to improve the predictive power of PSHA models.

Keywords: bayesian updating, PSHA, seismic source model, Gutenberg-Richter law, Gibbs sampler, mixture modeling

1 Introduction

1.1 Context of PSHA

To construct and justify the seismic resistance of specific buildings and sensitive facilities, engineers must assess the seismic risk, partly quantified by the seismic hazard curve, *i.e.* the probability of exceedance of a certain seismic ground motion threshold a (according to an intensity measure, such as the peak-ground acceleration, or PGA) during a certain reference period and at a given site of interest. In the following, we will note $\tau(PGA > a)$ the exceeding rate.

This evaluation is based on statistical models known as *PSHA* (Probabilistic Seismic Hazard Analysis) models, whose seismic activity parameters are estimated using catalogues of earthquakes, according to the methodology developed by the PSHA community, and whose general framework can be found in Cornell (1968), among others.

A central element of the PSHA analysis is provided by seismic source models (SMM). In low-to-moderate seismic regions, these models usually take the form of area-source models (or seismotectonic models), which consist of a partition of the territory under study into a number I of neighbouring regions, assumed to be seismotectonically homogeneous. More precisely, the number n_i of earthquakes observed over a period of t years within each region i, for $i = 1, \ldots, I$, is modelled by a Poisson law, of intensity $\lambda_i \times t$, where λ_i is the annual rate of earthquakes. Earthquake magnitudes $(m_{i,1}, \ldots, m_{i,n_i})$ are modelled by an exponential law of parameter β, truncated between two terminals $M_{\min}$ and $M_{\max}$, following the reference approach developed by Gutenberg and Richter (1944) (see Appendix A for more details).

The estimation of such recurrence models from a catalogue of earthquakes is often made difficult by the incompleteness of the latter. First of all, the magnitudes of observed earthquakes are never precisely known, which generally gives rise to censorship at contiguous intervals, in the form

$$\left[M_j \pm \frac{\delta}{2}\right[,$$

for $j = 1, \ldots, J$ with $M_1 - \frac{\delta}{2} = M_{\min}$, $M_J + \frac{\delta}{2} = M_{\max}$ and $M_j - M_{j-1} \equiv \delta$.

On the other hand, earthquake measurements have only been available for a limited period of time, called the complete observation period, whose

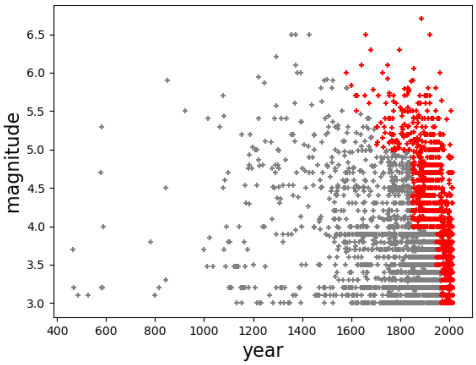


Fig. 1: French earthquake catalog (FCAT-17 from Manchuel et al (2018)): moment magnitude (Mw) larger than 3.0 versus time in years. Points highlighted in red belong to the complete observation periods and are the only one considered in our approach.

duration t_j (in years) varies according to the magnitude M_j, as illustrated in Figure 1. Thus, the weakest earthquakes have been recorded only since the installation of seismic sensor networks a few decades ago, while the traces left by high magnitude earthquakes remain visible for several centuries, whether in the natural environment (paleoseismology), in buildings (archeoseismology), or through written testimonies (historical seismicity). Moreover, the robustness of recurrence models depends on the number and the magnitudes of seismic contributions that have been observed. In low to moderate deformation contexts such as in France, this number is obviously low especially for the large magnitudes, of interest for seismic hazard, and therefore models are usually less well constrained than elsewhere.

As a result of this double censorship in magnitude and time, we actually observe a number n_{ij} of earthquakes in each zone i, for each class of magnitude M_j, and during a period of completeness t_j . The assumptions of the above model imply that the n_{ij} are independently distributed according to:

$$n_{ij} \overset{ind}{\sim} \mathcal{P}(\lambda_i t_j p_{ij}), \tag{1}$$

with

$$p_{ij}(\beta_i) = \frac{e^{-\beta_i m_j}}{\sum_\ell e^{-\beta_i m_\ell}}. \tag{2}$$

The estimation of this model therefore requires particular attention, and is generally done by maximum likelihood, using the seminal Weichert method Weichert (1980), or one of its more recent adaptations. However the Bayesian

approach, which quantifies more precisely the uncertainty on the seismic recurrence parameters within each zone, is an attractive alternative, as described by Keller et al (2014).

However, neither the number I nor the shape or the physical characteristics of the seismotectonic zones are precisely known, so that many seismotectonic models have been proposed in recent decades, without it being possible to clearly separate them. We study here the problem of selecting an optimal seismotectonic model from a given seismic catalogue CAT, which is decomposed according to each considered zoning model $\mathcal{M}$, as follows

$$CAT(\mathcal{M}) = \{M_j, t_j, n_{ij}^{\mathcal{M}}\}_{1 \leq i \leq I_{\mathcal{M}} 1 \leq j \leq J}$$

where:

- $I_{\mathcal{M}}$ is the total number of distinct zones within model $\mathcal{M}$
- $n_{ij}^{\mathcal{M}}$ is the number of counts for zone i and magnitude class j of model $\mathcal{M}$.

The magnitude classes and completeness periods are usually chosen independently from the seismotectonic model. However, it is also a common practice to re-define completeness periods for each zone of each model, which doesn't fundamentally change the methodology, but can surely have an influence on the results.

For a possible choice $\mathcal{M}$ of seismic source model, the rate of exceeding a threshold a by an intensity measure type (IMT) such as the peak-ground acceleration (PGA) is calculated as the sum of:

$$\tau\left(PGA > a|\mathcal{M}\right) = \sum_{i=1}^{I_{\mathcal{M}}} \int P\left(PGA > a, m_i, r_i\right)) \lambda_i(m_i|\mathcal{M}) f_R(r_i|\mathcal{M}) dr_i dm_i$$

where:

- $P(PGA > a|m_i, r_i)$ is the probability of exceeding the peak ground acceleration threshold a for an earthquake with magnitude m_i at a distance r_i of the considered site, in the zone i of the SSM. This probability is thanks to a ground motion model (GMM) ;
- $\lambda_i(m_i|\mathcal{M})$ is the occurrence rate of earthquake with magnitude m_i in zone i;
- $f_R(r_i|\mathcal{M})$ is the probability density associated to a distance r_i between the earthquake source and the interest site.

Many software packages provide Monte-Carlo estimators from $\tau(PGA > a)$ for each possible choice of SSM and GMM. This raises the problem of selecting the 'best' model (in a sense to be specified), each choice leading to different values of the exceeding rate, and therefore to different seismic hazard assessments.

We address this in the framework of Bayesian inference, which is to our knowledge the only coherent way to account for multiple sources of uncertainty Robert (2007). These include model parameters and the choice of the models

themselves, through the use of posterior model probabilities, which are conveniently interpreted as the probability for each model in competition to be the one that has generated the observed dataset. We insist here on the fact that this interpretation only holds under the assumption that the data have indeed been generated by one of the models under study. Thus, selecting a model according to a given criterion does not guarantee that the selected model is the model that has generated the data, but rather that it is the model most adapted to the available data, *ie* that best fits the data while also maintaining a low number of uncertain parameters.

Previous works on Bayesian model selection applied to PSHA models include for instance Viallet et al (2019), which use the same Bayesian framework, but in a more integral approach, wherein the complete PSHA models are compared according to how well ther predict the final interest quantity, that is, the hazard curve itself. In contrast, we focus here on validating the area source model only, based on how well they fit to the observed earthquake catalog. Similar Bayesian updating and selection of ground motion models has been studied in Bodda et al (2021a), and in Bodda et al (2021b), and could be combined to the results of the present work, with the potential to provide more global PSHA testing methods, which could be directly compared to that in Viallet et al (2019). Meanwhile, Mosca et al (2019) proposes a similar Bayesian approach, tailored to the British context.

This paper is organised as follows. In Section 2.1, we show how Bayesian model selection can easily be applied to PSHA modeling, and furthermore can be broken down in independent inferences for each region of a given seismotectonical model. Such region-specific analyses can be conducted using the Importance Sampling (IS) approach introduced in Keller et al (2014), a summary of which is provided in Section 2.2.

In Section 3, we demonstrate the efficiency of our methodology and illustrate its workings on a toy example, involving a source model with two rectangular zones. In Section 5, we then apply it on a real-life application concerned with the quantification of French seismicity. We then conclude in Section 6 with a discussion on the perspectives for the further development of our method.

2 Method

2.1 Selection of seismic source models based on Bayesian approach

The current approach to address uncertainty in the choice of source and ground motion models, which are the two main building blocks of PSHA modelling, is to associate prior probabilities to each competing model and then generate predictions from them. We propose to update these weights conditional on the available seismic data, through a rigorous application of Bayes' theorem. To do this, let us note $\mathcal{M}_k$ for $k = 1, \ldots, K$ the seismotectonic models, The prior

weights $P(\mathcal{M}_k)$ of these models are then updated according to Bayes' theorem:

$$P(\mathcal{M}_k|CAT) = \frac{P(CAT|\mathcal{M}_k)P(\mathcal{M}_k)}{\sum_{k'\ell'} P(CAT|\mathcal{M}_{k'})P(\mathcal{M}_k)} \quad (3)$$

where CAT represents the available data.

Two approaches exist to account for uncertainty on model choice during inference and prediction:

- select the *a posteriori* most probable model (maximizing $P(\mathcal{M}_k|CAT)$)
- aggregate models, weighted by their *a posteriori* probabilities (Bayesian model averaging, or BMA)

Computing posterior model weights requires the marginal likelihood $P(CAT|\mathcal{M}_k)$.Under the previously discussed assumptions, this is conveniently factored into:

$$P(CAT|\mathcal{M}_k) = \prod_i \int P(\lambda_i, \beta_i|\mathcal{M}_k)P(CAT_i(\mathcal{M})|\lambda_i, \beta_i)d\lambda_i d\beta_i. \quad (4)$$

Each factor can be calculated using importance sampling, as proposed in Keller et al (2014) (see Appendix B for technical details).

however, an important remark is that the classical PSHA model formulation in Weichert (1980), based on the Poisson distribution of the earthquake counts $n_{ij}^{(k)}$ in each seismotectonic zone i and magnitude class j, is not compatible with the model selection framework summarized by Equation (3). Indeed, the latter implies that every model is defined in terms of the *same* set of observations, noted *obs*, whereas the former is intrinsically based on a spatial clustering of the original dataset (the raw earthquake catalog) according to a specific seismotectonic model, defined as a partition $R = R_1 \sqcup \ldots \sqcup R_I$ of the region of interest R.

Hence, it is necessary to re-write the PSHA model in terms of the raw earthquake catalog that we can summarize as:

$$CAT = \{x_k, y_k, m_k\}_{1 \leq k \leq n}$$

The mathematical details of this re-formulation can be found in Appendix C. This leads to an 'imbalance' factor equal to:

$$\frac{n! \prod_i |R_i|^{n_i}}{\prod_{i,j} n_{ij}!}$$

To correct the likelihood values in Equation 4, as well as the posterior expression in Equation (5), used to compute model weights, one only needs to divide them by the above factor, to recover valid values, *ie*, that allow model comparison.

In conclusion, the Bayesian method we propose, allows an efficient computation, due to the factorization of the resulting marginal likelihoods, and the fact that it only requires integration in relation to the recurrence parameters (λ_i, β_i), independently by source.

2.2 Updating recurrence parameters

In this section, we recall some of the main elements of the methodology developed by Keller et al (2014) to compute the joint posterior distribution of the recurrence parameters (λ_i, β_i) in region R_i of a given seismotectonical model, given the observed counts n_{ij} for each magnitude class m_j over the complete observation period t_j (in years).

Following Keller et al (2014), independent gamma prior densities are chosen for each parameter:

$$\pi(\lambda_i) = \mathcal{G}(\lambda_i|n_0, t_0) = \frac{t_0^{n_0}}{\Gamma(n_0)}\lambda_i^{n_0-1}e^{-t_0\lambda_i}$$

$$\pi(\beta_i) = \mathcal{G}(\lambda_i|r_0, s_0) = \frac{s_0^{r_0}}{\Gamma(r_0)}\beta_i^{r_0-1}e^{-s_0\beta_i}$$

where, loosely speaking, n_0, t_0 can be interpreted by saying that the resulting prior on λ_i provides the same information as n_0 virtual earthquakes, observed during a virtual time period of t_0 years. r_0, s_0 have no particular interpretation.

An important result is that the conditional posterior of λ_i given β_i is still Gamma-shaped:

$$\pi(\lambda_i|\beta_i, CAT_i(\mathcal{M}), \mathcal{M}) = \mathcal{G}\left(\lambda_i|n_0 + n, t_0 + \sum_j t_j p_{ij}(\beta_i)\right).$$

Hence, it is easily integrated out, yielding the marginal posterior of β:

$$\pi(\beta_i|CAT_i(\mathcal{M}), \mathcal{M}) \propto \pi(\beta_i)\prod_j p_{ij}(\beta_i)^{n_{ij}}\left(t_0 + \sum_j t_j p_{ij}(\beta_i)\right)^{-(n_0+n)} \quad (5)$$

$$\times \frac{t_0^{n_0}}{\Gamma(n_0)}\Gamma(n_0 + n)\prod_j \frac{t_j^{n_j}}{n_j!}. \quad (6)$$

Details about the importance sampling approach to generate a sample $(\beta_i^{(s)})_{1\leq s\leq S}$ from this posterior density, as well as an estimate of the marginal likelihood $P(CAT_i(\mathcal{M})|\mathcal{M}) = \int P(CAT_i(\mathcal{M})|\lambda_i, \beta_i)P(\lambda_i, \beta_i|\mathcal{M})d\lambda_i\beta_i$ can be found in Keller et al (2014).

We want to highlight here a critical aspect of β_i's posterior density. Indeed, leaving the prior $\pi(\beta_i)$ aside, notice that it depends on β_i only through the

probabilities $p_{ij}(\beta_i)$ defined in Equation (2). Furthermore, it is easy to see that:

$$\lim_{\beta_i \to 0^+} p_{ij}(\beta_i) = \frac{1}{n}$$
$$\lim_{\beta_i \to +\infty} p_{ij}(\beta_i) = \mathbf{1}_{\{j=1\}}$$

In other terms, when $\beta_i \to 0^+$, all classes become equally probable, whereas, when $\beta_i \to +\infty$, the first class (corresponding to the smallest magnitude) has a probability of one, and all the others, zero.

As a consequence, the posterior density of β (if we skip $\pi(\beta)$) does not decay to zero when $\beta_i \to 0^+$, and decays to zero when $\beta_i \to +\infty$ only if the number of counts n_{ij} is nonzero for at least one $j > 1$. This shows that the choice of prior hyper-parameters for β, must be done carefully. Indeed, choosing $r_0 < 1$ yields a posterior density for β_i which is unbounded in any neighborhood of 0, making the estimation difficult. Hence in the following we advocate the choice $r_0 = 1$, meaning that we use in practice an exponential prior on β_i, with prior mean and standard deviation both equal to $1/s_0$.

This also means that we cannot use so-called non-informative improper priors, obtained by having hyper-parameters r_0, s_0 go to 0, since they would result in an unbounded posterior density in the neighborhood of 0^+. In practice, we avoid this issue by constraining β_i to vary between $[\beta_{\min};\ \beta_{\max}]$. This means that we adopt a doubly truncated exponential prior on β, which converges to the uniform density on $[\beta_{\min};\ \beta_{\max}]$, for the minimally informative choice $r_0 = s_0 = 0$.

3 Illustration on a toy example

This first experiment is easily reproduced using the `toy_SourceModelTree.py` script from the `demos/ToyDataset` sub-directory. To better understand how our methodology works, we devised a simplistic toy example, in which we considered two candidate source models defined over a certain rectangular search domain R :

1. $\mathcal{M}_1$, containing as its single zone the whole search domain R, and defined by a couple of recurrence parameters $\theta = (\lambda, \beta)$;
2. $\mathcal{M}_2$, containing two zones R_1 and R_2, with equal surfaces ($|R_1| = |R_2|$) and defining a partition of the search domain $R = R_1 \sqcup R_2$,, and defined by two couples of recurrence parameters $\theta_1 = (\lambda_1, \beta_1)$ for R_1 and $\theta_2 = (\lambda_2, \beta_2)$ for R_2.

Note that, $\mathcal{M}_1$ can be seen as the special case of $\mathcal{M}_2$ in which $(\lambda_1, \beta_1) = (\lambda_2, \beta_2)$. In other terms, the models are *nested*, with $\mathcal{M}_1 \subsetneq \mathcal{M}_2$.

3.1 Synthetic data simulation and first results

In order to simulate from model $\mathcal{M}_2$ (hence also $\mathcal{M}_1$), we first had to choose specific values for the recurrence parameters. For simplicity's sake, we first

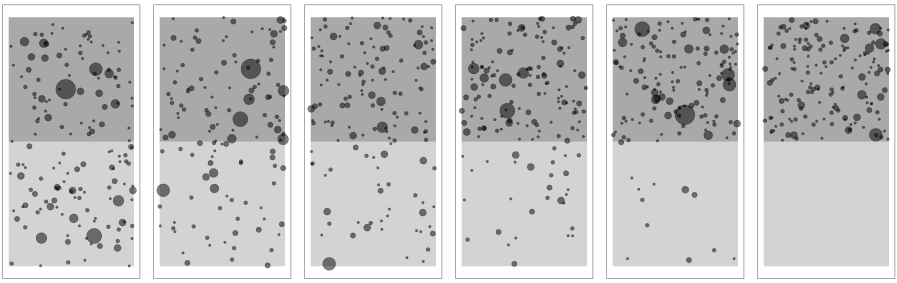
Dataset	(a)	(b)	(c)	(d)	(e)	(f)
α	0.5	0.6	0.7	0.8	0.9	1.0
$P(\mathcal{M}_2\|obs)$	0.0	0.02	1.0	1.0	1.0	1.0

Table 1: Posterior probabilities for the two-zones model $\mathcal{M}_2$, corresponding to different α (proportion of earthquakes in zone R_1 compared to R) and associated subplots of Figure 2. For $\alpha = 0.5$, we remind that the zones R_1 and R_2 have similar recurrence parameters, so the one-zone model $\mathcal{M}_2$ should be favored, that is the case with $P(\mathcal{M}_2|obs)$ equal to zero. On the other hand, for $\alpha \neq 0.5$, the two-zones model $\mathcal{M}_2$ should be favored.

used a fixed Gutenberg-Richter slope $\beta_1 = \beta_2 = \log(10)$ (i.e. a b-value equal to 1) throughout the study, focusing on the Poisson intensities λ_1 and λ_2. To this end, we introduce the alternative parameterization $(p, \alpha) \in \mathbb{R}^+ \times [0, 1]$, such that:

$$\begin{aligned} \lambda_1 &= 10^p \alpha; \\ \lambda_2 &= 10^p (1 - \alpha). \end{aligned}$$

In other words, p controls the expected yearly earthquake rate throughout the search domain, equal to 10^p, while α is the expected proportion of earthquakes happening in zone R_1. As explained earlier, $\mathcal{M}_1$ corresponds to the special case $\alpha = 50\%$.



(a) $\alpha = 0.5$ (b) $\alpha = 0.6$ (c) $\alpha = 0.7$ (d) $\alpha = 0.8$ (e) $\alpha = 0.9$ (f) $\alpha = 1.0$

Fig. 2: Synthetic catalogs simulated from the two-zones toy source model, for $p = 1$ and increasing values of α.

Figure 2 shows different catalogs simulated according to $\mathcal{M}_2$ for $p = 1$ and different values of α ranging from 0.5 to 1.0, together with the posterior probabilities (Table 1) that each catalog has been simulated according to $\mathcal{M}_2$ rather than $\mathcal{M}_1$, as defined in Equation (3), assuming equal prior probabilities $P(\mathcal{M}_j) \equiv 0.5$. As can be seen, posterior model probabilities are extremely contrasted, taking almost systematically $0 - 1$ values, meaning that model selection was done with almost absolute certainty. Furthermore, for $\alpha = 0.5$,

our algorithm correctly selected the single-zone $\mathcal{M}_1$ model, the posterior probability $\mathcal{M}_2$ being equal to zero. Likewise, for all tested $\alpha > 0.5$, $\mathcal{M}_2$ is correctly selected, with posterior probability equal to 1, except for $\alpha = 0.6$, which yields $P(\mathcal{M}_2|obs) = 2\%$. This can seem counter-intuitive at first glance, because for $\alpha = 0.6$ it is extremely difficult to visually discriminate, based on the observed catalog, between the two-zones $\mathcal{M}_2$ and the single-zone $\mathcal{M}_1$ models. Hence, we would expect the posterior probabilities to be close to $1/2$.

To better understand this phenomenon, it is worth recalling that Bayesian model selection tends to favor models with fewer parameters. In particular, it is shown in the seminal work of Schwarz (1978) that asymptotically (for large datasets) the posterior probability of model $\mathcal{M}_j$ is approximately proportional to:

$$\pi(\mathcal{M}_j|CAT) \propto \frac{\max_{\theta_j} P(CAT|\theta_j, \mathcal{M}_j)}{n^{d_j/2}}, \tag{7}$$

where:

- $\max_{\theta_j} P(obs|\theta_j, \mathcal{M}_j)$ denotes the maximum achievable likelihood for model $\mathcal{M}_j$ when observing *obs*, *ie*, the standard way to measure the "goodness-of -it" of a certain model to a given dataset;
- $n^{d_j/2} \geq 1$ is a penalty factor, increasing polynomially with the total number n of observations (here, earthquakes) and exponentially with the number d_j of components of the parameter vector θ_j.

This explains why, in general, Bayesian model selection favors models that combine goodness-of-fit and parsimony, *ie*, that explain the data as well as possible, while depending on as few free parameters as possible. For instance, between two models that fit the data equally well (according to the maximum likelihood criterion), the one with fewest parameters systematically receives the largest posterior probability; this is typically the case when the available data is too scarce to discriminate between both models, as seen here with $\alpha = 0.6$.

%endfigure

3.2 Reproducibility study

To assess the reproducibility of the above results, we conducted a second numerical experiment, wherein we simulated S *iid* independent datasets $(obs_{k,s})_{\substack{1\leq k\leq K\\ 1\leq s\leq S}}$, for several couples of values (p_k, α_k). Thus, we were able to evaluate the capacity of our Bayesian model selection procedure to correctly identify whether each dataset was simulated according to model $\mathcal{M}_1$, meaning that $\alpha_k = 0.5$, or to model $\mathcal{M}_2$, meaning that $\alpha_k \neq 0.5$.

More specifically, we tested all possible combinations of:

- six different values for p, linearly spaced between 0 and 3;
- two distinct values of α : 0.5, for which the data are simulated according to the $\mathcal{M}_1$ model, and 0.6, for which the data are simulated according to the two-zones $\mathcal{M}_2$ model.

Finally, we simulated $S = 10$ replicates for each couple (p_k, α_k).

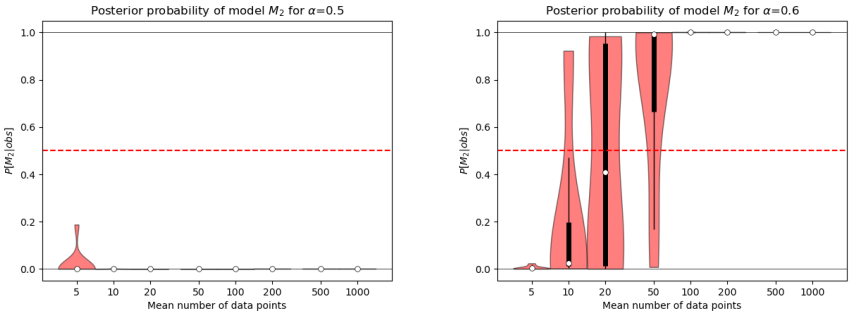


Fig. 3: Violin plots of the empirical distribution of posterior probabilities for model $\mathcal{M}_2$ for increasing sample sizes, for $\alpha = 0.5$ (left) and $\alpha = 0.6$ (right).

Figure 3 shows the results of this reproducible study for λ analysis, which confirms our first results. Indeed, for $\alpha = 0.5$, our algorithm correctly selected the $\mathcal{M}_1$ model for all 60 simulated datasets, with a posterior probability for $\mathcal{M}_2$ that was never greater than 0.2. For $\alpha = 0.6$ on the other hand, the results depended crucially on the expected sample size. For small datasets (less than 15 observations per time unit), $\mathcal{M}_1$ is almost systematically chosen. As explained earlier, this is due to the fact that Bayesian model selection penalizes model complexity, especially when the available data is scarce, so $\mathcal{M}_1$ is preferred over the more complex model $\mathcal{M}_2$ in the absence of significant evidence towards the latter.

For slightly larger datasets, corresponding to 15 observations per unit of time, a large variability of the posterior model probabilities can be witnessed, as more data are available and, in some cases, provide enough evidence to correctly identify the $\mathcal{M}_2$ model. For even larger datasets, corresponding to 63 observations per unit of time and above, this variability vanishes as enough data are observed to systematically select the $\mathcal{M}_2$ model.

In conclusion, this simulation study illustrates the standard behavior of Bayesian model selection, which favors the simplest models (with fewer parameters), except when the data provides overwhelming evidence in favor of more complex ones. This is why, in our reproducible study, for low sample sizes, $\mathcal{M}_2$ systematically received near-zero posterior probability, whether it had been used or not to generate the data. This shouldn't be interpreted as a strong confidence in $\mathcal{M}_1$ being "the true data-generating model", but rather as a strong confidence that $\mathcal{M}_1$ is the most realistic model that can be learned given the available data.

4 Optimal zone clustering

As illustrated in the above simulation studies, Bayesian model selection allows to adapt the complexity of seismic source models (measured by the number of zones) to the amount of available data. This means that the list of candidate

models should ideally contain as many different models as possible, ranging from the simplest (single-zone) to the most complex.

Now, given a seismic source model $\mathcal{M}$ comprising I disjoint regions $R_1, \ldots, R_I$, an obvious way to generate alternative models consists in merging the I regions R_i into (at most) K clusters, with $K \leq I$. Formally, this amounts to defining additional parameters $\gamma_i \in \{1, \ldots, K\}$, for $i = 1, \ldots, I$, such that $\gamma_i = k$ iff R_i is in the k-th cluster. Note $\mathcal{M}_{\boldsymbol{\gamma}}$ the source model obtained from $\mathcal{M}$ by merging zones according to $\boldsymbol{\gamma}$. The total number of such models is given by the Bell number,which increases very quickly with I. So computing $P\left(\mathcal{M}_{\boldsymbol{\gamma}}|obs\right)$ for all possible choices of $\boldsymbol{\gamma}$ becomes intractable.

A more efficient strategy is to sample from the posterior distribution of $\boldsymbol{\gamma}$, as given by Bayes' theorem, assuming an uniform prior $\pi(\boldsymbol{\gamma}) \propto 1$:

$$\pi\left(\boldsymbol{\gamma}|CAT\right) \propto P\left(CAT|\mathcal{M}_{\boldsymbol{\gamma}}\right), \tag{8}$$

where the marginal likelihood $P\left(CAT|\mathcal{M}_{\boldsymbol{\gamma}}\right)$ can be computed as explained in Section 2.2. A simple and widespread algorithm to do so is the Gibbs sampler Robert and Casella (2004), which consists in sequentially updating each component of the cluster-allocating parameter $\boldsymbol{\gamma}$, given an arbitrary starting point $\boldsymbol{\gamma}^{(0)}$. At step $t = 1, 2, \ldots$, the current value $\boldsymbol{\gamma}^{(t-1)}$ is updated according to the following conditional simulation steps for $i = 1, \ldots, n$:

$$\begin{aligned}
\gamma_1^{(t)} &\sim \pi\left(\gamma_1^{(t)}|\gamma_2^{(t-1)}, \ldots, \gamma_I^{(t-1)}, \mathcal{M}, CAT\right) \\
\gamma_2^{(t)} &\sim \pi\left(\gamma_2^{(t)}|\gamma_1^{(t)}, \gamma_3^{(t-1)}, \ldots, \gamma_I^{(t-1)}, \mathcal{M}, CAT\right) \\
\gamma_3^{(t)} &\sim \pi\left(\gamma_3^{(t)}|\gamma_1^{(t)}, \gamma_2^{(t)}, \gamma_4^{(t-1)}, \ldots, \gamma_I^{(t-1)}, \mathcal{M}, CAT\right) \\
&\vdots \\
\gamma_I^{(t)} &\sim \pi\left(\gamma_I^{(t)}|\gamma_1^{(t)}, \ldots, \gamma_{I-1}^{(t)}, \mathcal{M}, CAT\right).
\end{aligned}$$

Noting $\boldsymbol{\gamma}_{-i} = (\gamma_1, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots, \gamma_I)$ the vector $\boldsymbol{\gamma}$ minus its $i - th$ component, γ_i's conditional posterior density is given by:

$$\pi\left(\gamma_i = k|\boldsymbol{\gamma}_{-i}, \mathcal{M}, CAT\right) \propto P(CAT|\mathcal{M}_{(\gamma_1, \ldots, \gamma_{i-1}, k, \gamma_{i+1}, \ldots, \gamma_I)}).$$

Hence, updating the i-th component of $\boldsymbol{\gamma}$ can be done by:

1. Computing $\pi\left(\gamma_i = k|\boldsymbol{\gamma}_{1:i-1}^{(t)}, \boldsymbol{\gamma}_{i+1:I}^{(t-1)}, \mathcal{M}, CAT\right)$ for every candidate value $k = 1, \ldots, K$;
2. Drawing γ_i from the discrete distribution defined by these conditional distributions.

This simple solution is made possible thanks to the fact that we can efficiently evaluate the marginal likelihood $P\left(CAT|\mathcal{M}_{\boldsymbol{\gamma}}\right)$ for any value $\boldsymbol{\gamma}$.

Furthermore, because this quantity can be factorized across clusters, when updating a single component γ_i, say from $\gamma_i = k$ to $\gamma_i = k'$, it suffices to recompute the marginal likelihood of clusters k and k', since all other clusters are left unchanged.

Likewise, for each posterior draw $\boldsymbol{\gamma}^{(t)}$, *ie* for each possible clustering, it is possible to simulate a realization $(\boldsymbol{\lambda}^{(t)}, \boldsymbol{\beta}^{(t)})$ from the joint conditional posterior distribution of recurrence parameters in each individual zone of the considered model:

$$\boldsymbol{\lambda}^{(t)}, \boldsymbol{\beta}^{(t)} \sim \pi\left(\boldsymbol{\lambda}^{(t)}, \boldsymbol{\beta}^{(t)} | \mathcal{M}_{\boldsymbol{\gamma}^{(t)}}, CAT\right).$$

By construction, this yields a sample $(\boldsymbol{\lambda}^{(t)}, \boldsymbol{\beta}^{(t)})_{t=1,\ldots,T}$ from the Bayesian-model-averaged (BMA) joint posterior distribution of recurrence parameters for each zone, that is, having integrated out the uncertain model-choice parameters $\boldsymbol{\gamma}$:

$$\pi\left(\boldsymbol{\lambda}, \boldsymbol{\beta} | \mathcal{M}, CAT\right) = \int_{\boldsymbol{\gamma}} \pi\left(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\gamma} | \mathcal{M}, CAT\right) d\boldsymbol{\gamma}.$$

4.1 Results on the toy example

We now apply the above Gibbs sampling algorithm to the toy example introduced in Section 3. In this simplified, two-zones, setting, the clustering parameters are reduced to $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$, with $\gamma_1, \gamma_2 \in \{1, 2\}$ since there are at most two clusters, corresponding to the two initial zones. Note that the four possible values of $\boldsymbol{\gamma}$ correspond in fact to only two possible models:

- the single-cluster model, termed $\mathcal{M}_1$ in Section 3, corresponding to $\alpha = 0.5$, and $\gamma_1 = \gamma_2$, *ie* $(\gamma_1, \gamma_2) = (1, 1)$ or $(\gamma_1, \gamma_2) = (2, 2)$;
- the two-clusters model, termed $\mathcal{M}_2$ in Section 3, corresponding to $\alpha \neq 0.5$, and $\gamma_1 \neq \gamma_2$, *ie* $(\gamma_1, \gamma_2) = (1, 2)$ or $(\gamma_1, \gamma_2) = (2, 1)$.

Hence, based on the sample $(\boldsymbol{\gamma}^{(t)})_{1 \leq t \leq T}$ from the posterior distribution (8) given by the Gibbs sampler, the probability of the two-clusters model $\mathcal{M}_2$ can be estimated by the empirical ratio:

$$\widehat{\pi}(\mathcal{M}_2 | CAT) = \frac{\sum_{t=1}^{T} \mathbf{1}_{\{\gamma_1^{(t)} \neq \gamma_2^{(t)}\}}}{T}.$$

We chose to simulate the data for $\alpha = 0.6$ and $p = 1.3$, leading to an average number of simulated earthquakes equal to $10^p \approx 20$, since, according to Figure 3, discriminating between the one and two-cluster models seemed most difficult in this case.

We ran 3 Markov chains of 5 000 iterations each, according to the above Gibbs algorithm. Convergence to the stationary distribution was checked by the Gelman-Rubin statistic, which was equal to 1.00 for both γ_1 and γ_2. Next,

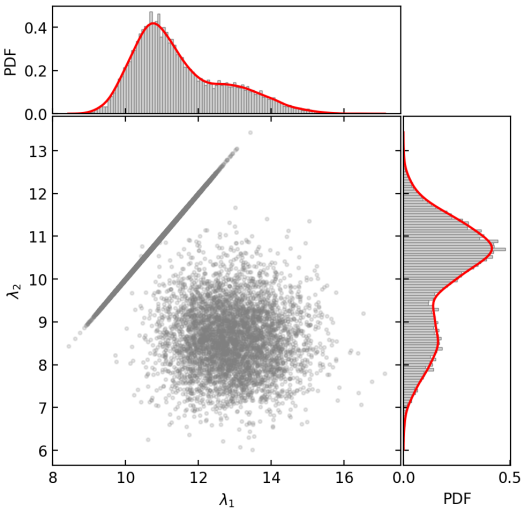


Fig. 4: Model-averaged joint posterior distribution of (λ_1, λ_2)

IS	0.329
Gibbs	0.327
Gibbs LCL	0.319
Gibbs UCL	0.335

Table 2: Posterior probability for model $\mathcal{M}_2$ estimated by different methods

we estimated the effective sample size for the concatenation of the three chains, which was equal to: 14 617, out of the 15 000 realizations.

Figure 4 illustrates the resulting scatter plot and marginal histograms of the joint Bayesian model-averaged posterior distribution for (λ_1, λ_2). It is the mixture of two following distributions, which explains the bimodality of the histograms:

- the posterior, conditional on model $\mathcal{M}_1$. In this case, $\lambda_1 = \lambda_2$ by construction since both zones are merged, hence the points are contained in the $y = x$ line;
- the posterior, conditional on model $\mathcal{M}_2$. In this case, λ_1 and λ_2 are independent by construction, since they are estimated separately within each zone.

Consequently, the proportion of points in the scatterplot that are not contained in the $y = x$ line is an empirical estimate of the posterior probability of model $\mathcal{M}_2$. Table 2 compares this estimate, termed "Gibbs" to the Importance sampling ("IS") estimate. As expected, both estimates are coherent.

5 Real case study application

We selected the French territory to carry out a large scale application which is based on the homogenized and declusterized French earthquake catalog (noted FCat-D20 in the following) built and used in Drouet et al (2020) and illustrated in Figure 5. This catalogue integrates the FCAT-17 catalogue from Manchuel et al (2018) with more recent seismicity from the LDG bulletins (https://www-dase.cea.fr/) and further large earthquakes from European catalogue SHEEC (Stucchi et al, 2013)

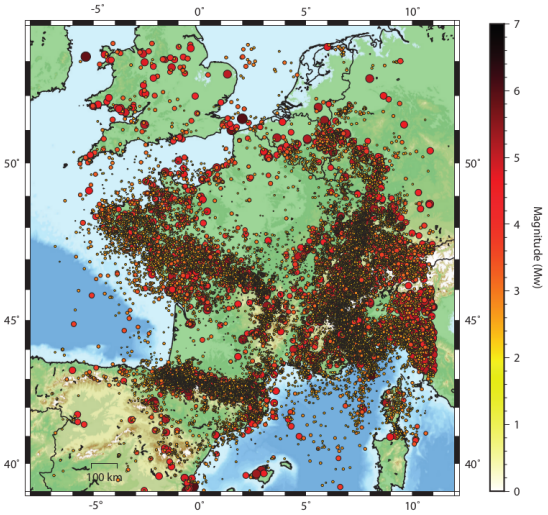


Fig. 5: The French earthquake catalog for PSHA .

We considered the seismotectonic models proposed by four different organizations: EDF (34 zones) (Drouet et al, 2020), Geoter (noted GTR in the following, 58 zones), Drouet et al (2020), IRSN (32 zones) (Baize et al, 2013), and CEA (33 zones) (Marin et al, 2004), as illustrated in Figure 6. As complementary information, the first three models have also been used in the PSHA study of Drouet et al (2020). An important remark is that these models do not recover the same total geographical footprint, meaning that each one applies to a different sub-dataset of the catalog, obtained by removing from the FCat-D20 catalog earthquakes outside their external boundaries. This makes it impossible to compare them directly from a statistical viewpoint, since a minimal condition is that all models share the same dataset, noted *obs* in Equation (3). We avoided this issue by limiting ourselves to the intersection of the model's footprint, which involved modifying zones that weren't strictly enclosed in that intersection On the downside, this reduced the total number of considered earthquakes for the study.

Choice of prior laws

To simplify the analysis, we adopted the same prior law for Bayesian inference on the recurrence parameters (λ_i, β_i) within each zone i of each of the seismotectonic models considered, parameterized by a single hyperparameter σ_0 representing the same prior standard deviation for all parameters, according to:

$$\begin{aligned}\lambda_i|\sigma_0 &\sim \mathcal{G}(1/\sigma_0^2, 1/\sigma_0^2)\\ \beta_i|\sigma_0 &\sim \mathcal{G}(1, 1/\sigma_0)\mathbf{1}_{\{\beta_{\min}\leq\beta_i\leq\beta_{\max}\}}.\end{aligned}$$

Notice that λ_i has prior mean 1 and prior variance σ_0^2, while β_i has prior mean σ_0 and prior variance σ_0^2, before truncation between $\beta_{\min}$ and $\beta_{\max}$. We set $\sigma = 10^{19}$ as a weakly informative choice; a sensitivity analysis to this hyperparameter, which we do not develop here for lack of space, showed that the results of the analysis were robust to this choice.

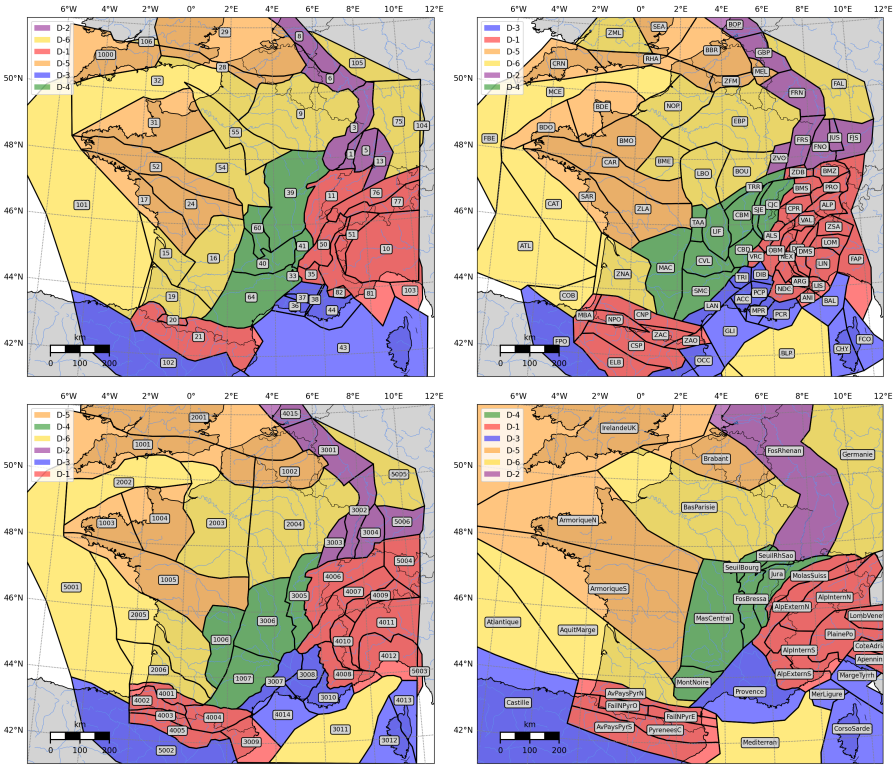


Fig. 6: From left to right, then top to bottom: seismotectonic models of the national territory proposed by EDF, GTR, IRSN, and CEA respectively. Colors indicate the large domains as defined in Drouet et al (2020)

Model (nb. zones)	IRSN (32)	CEA (33)	EDF (34)	GTR (58)
log-marginal likelihood	**-15 687**	-16 054	-15 911	-17 790
posterior weight	**1.0**	0.0	0.0	0.0

Table 3: Bayesian model selection for the four French zoning models.

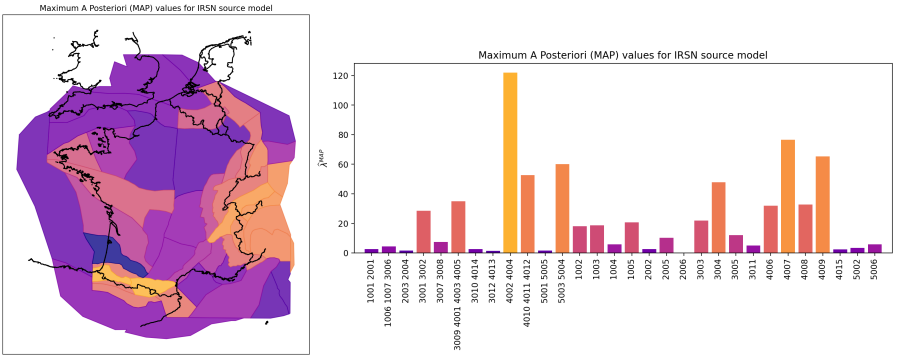


Fig. 7: Normalized λ_i MAP values estimated using the FCat-D20 catalog, for a reference area of 10^6 km^2, and for magnitudes over $M_{\min} = 3.0$ for the IRSN model, for a prior standard deviation of $\sigma_0 = 10^{19}$

Meanwhile, uniform prior weights where assigned to each of the four candidate models, though we will see how these had in fact very little influence on the results of the inference.

First results

We performed Bayesian inference in the four candidate models, using the above prior choice, and the importance sampling procedure described in Section 2.2. The posterior probabilities, given in Table 3, were extremely contrasted, since the IRSN model received a posterior probability of one. Based on this result, one is tempted to abandon all remaining models, who received a zero posterior probability, and simply update the IRSN source model; the result is illustrated in Figure 7, which shows the map of annual earthquake rates per surface unit.

However, this solution is hardly satisfying, for at least two reasons. First, the IRSN model deemed optimal is also the one with the lowest number of parameters. Hence, as in the toy example in Section 3, a possible explanation is that the considered catalog didn't provide sufficient information to discriminate between the candidate models, in which case Bayesian model selection is known to favour simpler models. This explanation is supported by the fact that the log marginal likelihoods of the different models are seen to decrease with the number of zones (with an exception between the LDG and EDF models).

Secondly, the source models are seen to fit the dataset more or less well, according the amount of available data in each seismotectonical zone. In fact, after removing earthquakes with magnitudes below 3, GTR and IRSN ended

Domains-based model	IRSN	CEA	EDF	GTR
log-marginal likelihood	-13 613	-13 691	**-13 552**	-13 684
posterior weight	0.0	0.0	**1.0**	0.0

Table 4: Bayesian model selection for the large domains of the four French zoning models.

up with an empty zone each (in which no earthquake had been recorded), while EDF had one zone with a single recorded earthquake, and GTR one zone with only three earthquakes. This is a clear indication that existing source models contain too many zones with respect to the available data, whose recurrence parameters cannot be properly estimated.

This strongly suggests to use source models with fewer, larger zones, better adapted to the limited information available in the earthquake catalog. We explore several options to do so in the following sections.

5.1 Domains-based inference

As shown in Figure 6, the zones within each model can be aggregated into 6 large domains spanning the whole French territory. These domains can be considered as alternative, simplified, source models. Thus, we applied the same Bayesian updating procedure as for the full zoning models. Results of the Bayesian model selection are given in Table 4. It is worth noting that the log marginal likelihood values based on these domains are well above those associated to the original zoning models. this confirms that the latter contained too many zones with respect to the available catalog. However, the weights of the four domain-based models, which now all share the same number of zones (6), are still very constrasted, with the EDF model now receiving a posterior probability of 1. Figure 8 shows the resulting map of annual earthquake rates per surface unit.

Again, this result is not entirely satisfactory from a statistical point of view, since it depends strongly on the specification of the large domains. Indeed, nothing prevents the existence of yet another source model, or another way of aggregating the existing ones, with potentially even fewer large domains, that could outperform our current best solution.

5.2 Clustering-based inference

The Gibbs-sampling algorithm presented in Section 4, allows to explore all possible clusters of the model zones, and identify which are in best agreement with the available data, as measured by the log-maginal likelihood. We ran this algorithm, for each of the four candidate source models, for 500 iterations, including 100 so-called "burn-in" iterations to allow the Markov Chain to reach its equilibirum state. In order to check the convergence, we ran the algorithm from different, randomly chosen, initial values for the clustering parameter γ. We found empirically that the initial number of nonempty classes had a major influence on the resulting posterior sample, and that choosing as little

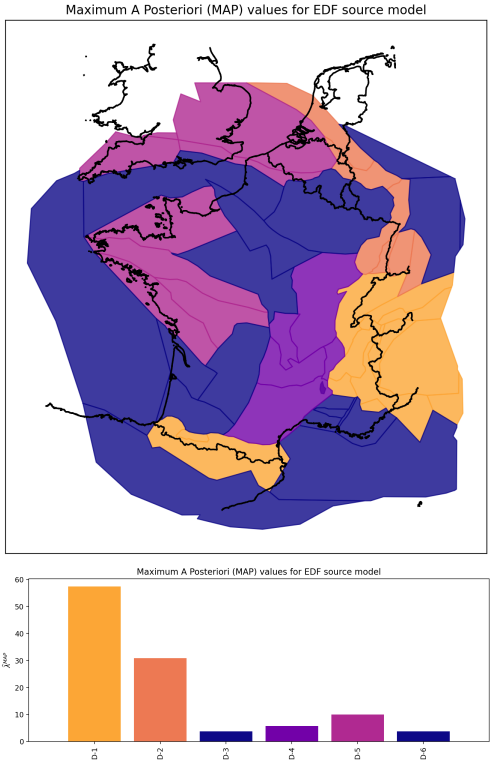


Fig. 8: Normalized λ_i MAP values estimated within each large domain of the EDF model, using the FCat-D20 catalog, for a reference area of $10^6 km^2$, and for magnitudes over $M_{\min} = 3.0$.

Clustered model (zones)	IRSN (2)	CEA (3)	**EDF (4)**	GTR (7)
log-marginal likelihood	-14 338	-13 247	**-12 661**	-13 008
posterior weight	0.0	0.0	**1.0**	0.0

Table 5: Bayesian model selection of French seismic source models, considering the most probable clustering *a posteriori*.

as many active clusters consistently led the chain to areas of highest posterior densities. Such results suggest both a multimodal posterior, and the difficulty of our Gibbs algorithm to "jump" between modes.

Table 5 show the final scores obtained by the different source models, with optimal clustering via Gibbs sampling. As could be expected, the optimal number of clusters was found to be much less than the initial number of zones, and close to the number of domains (6). The best score was achieved by optimal clustering of the EDF source model, which received followed by optimal clustering of the GTR source model. However the gap in terms of log-marginal likelihood is of several hundreds points, meaning that the EDF model receives

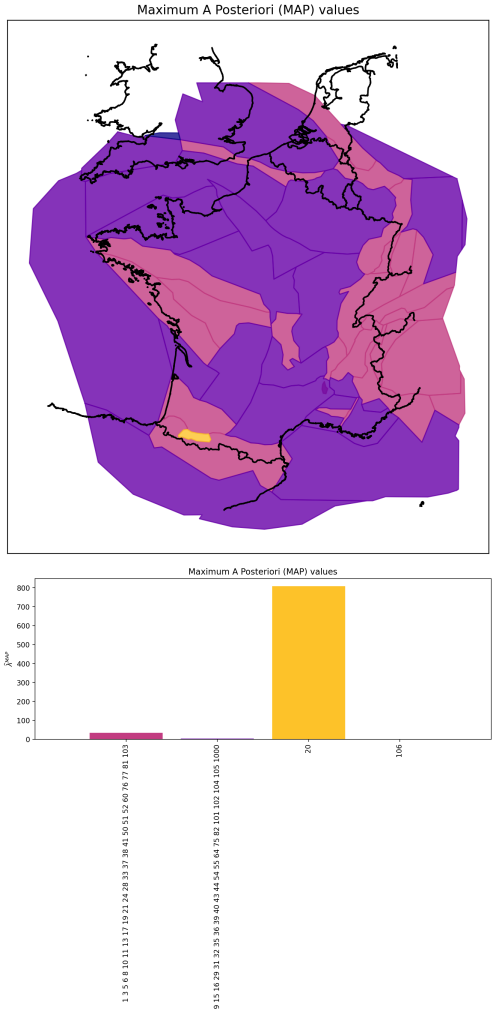


Fig. 9: Normalized λ_i MAP values estimated within each large domain of the optimallyu clustered EDF model, using the FCat-D20 catalog, for a reference area of $10^6 km^2$, and for magnitudes over $M_{\min} = 3.0$.

a posterior weight of 1. Figure 9 shows the resulting "most probable" map of yearly earthquake rates.

However, note that we have only considered so far the *a posteriori* most probable clustering $\widehat{\boldsymbol{\gamma}}_{MAP}(\mathcal{M}) = \arg\max_{\boldsymbol{\gamma}} P(CAT|\mathcal{M}_{\boldsymbol{\gamma}})$ for each source model $\mathcal{M}$. If we focus on the EDF source model, remember that our Gibbs algorithm outputs a full sample $(\boldsymbol{\gamma}^{(t)})_{1\leq t\leq T}$ from $\boldsymbol{\gamma}$'s posterior distribution. The associated posterior weights $P(CAT|\mathcal{M}_{\boldsymbol{\gamma}^{(t)}})$ are much less contrasted than in the previous setting, as can be seen in Figure 10 by sorting them in decreasing order and taking cumulative sums: 250 distinct models, corresponding each to

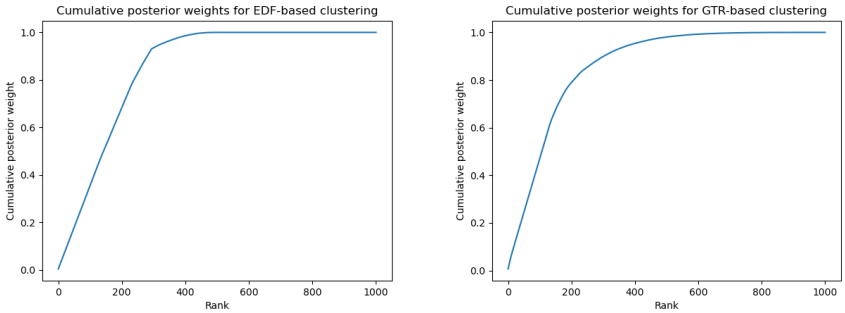


Fig. 10: Cumulative sum of posterior probabilities associated to each possible clustering of the EDF (left) and GTR (right) source model, ordered in decreasing order.

a different clustering of the EDF seismotectonical zones, are needed to cover the full support of the posterior distribution.

As discussed previously, we can then easily obtain the BMA posterior distribution of recurrence parameters per seismotectonical zones, that is, averaged over all possible clusterings. Figure 11 shows this BMA posterior distribution for the recurrence parameters from two distinct zones of the EDF model, characterized by high (Zone 20), and low (zone 15) respective sismicity levels. Interestingly, the marginal posteriors have similar spreads in both cases (visual differences being due to the scales), demonstrating that our clustering approach enables to estimate recurrence parameters accurately, even in zones with limited data. On the other hand, the joint posterior distribution in Zone 15 is bimodal, illustrating the uncertainty on the clustering in such low sismicity regions.

Figure 12 shows the resulting map of normalized λ values (earthly earthquake rates) over France, when considering the 95-th percentile of the BMA distribution of normalized λ values per zone of the EDF and GTR models. Choosing a quantile resulted in conservative upper bounds for λ. As can be seen, the EDF MAP seems to better localize a specific seismic source in the Pyrenean, resulting in a locally more contrasted map, which may explain why it outperforms the GTR model. On the other hand, GTR seems to better localize the sources in the Alpean region, which may exlpain why it scored second to best. Hence, it seems fair to assume that there is still space for improvement here, and that one could obtain yet higher marginal likelihood values, by combining in some way the different zoning models, rather than treating them separately.

6 Conclusion and perspectives

We have introduced a fully Bayesian methodology to update prior distributions on the choice of seismic source models as well as the associated parameters,

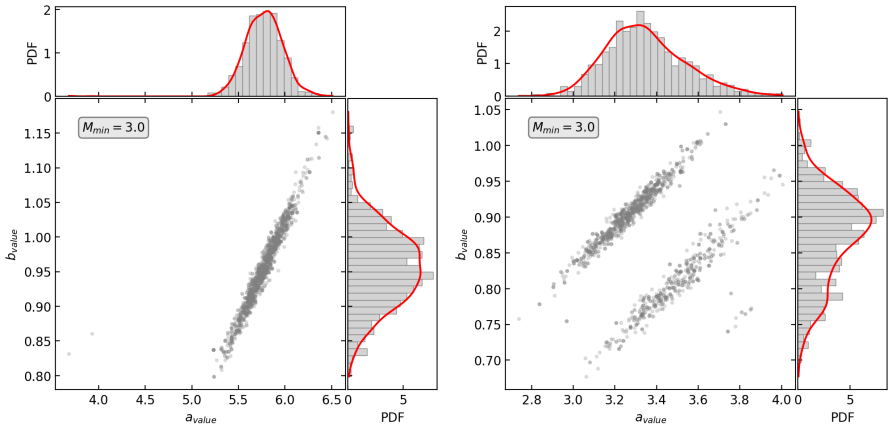


Fig. 11: Scatter plots of the BMA joint posterior distributions for the λ and β values of Zones 20 (left) and 15 (right) of the EDF source model, averaged over all possible clusterings

based on a catalog of observed earthquakes within a certain geographical domain of interest. We then extended the model choice by including all possible clusterings of a given source model's seismotectonical zones.

We illustrated the workings of our approach on a toy problem, which consisted in choosing between a single-zone source model and a two-zones source model. We then applied it to a real-life case study involving four different source models adapted to Metropolitan France.

Our results highlight the conservatism of Bayes factor are in the context of the considered seismotectonical source models, leading to systematically choose the simplest model (with less zones), except in the ideal case where the data is explicitly simulated according to one of the candidate models. in the latter case, given enough data, the "true" model is systematically chosen, in accordance with the well known fact that Bayes factors are consistent Schwarz (1978).

In our real-like application, the weights allocated to the four difference source models where strictly ranked according to their number of zones, the domains based-source model, comprising six large domains, ends up being preferred above all other parametric source models. This was a clear sign that a compromise needed to be found between the current source models, containing too many zones compared to the available data to be correctly estimated, and the overly-simplistic domains-based source model.

This is why we used a probabilistic clustering approach to optimally merge regions with similar seismicity. This resulted in an optimal number of regions between 2 and 7, far below the 32 to 58 regions of current source models, while achieving significantly higher values of the associated log marginal likelihood. Furthermore, the clustering changed the ranking of the different models, for instance the GTR model with its 7 optimal clusters outperformed both the

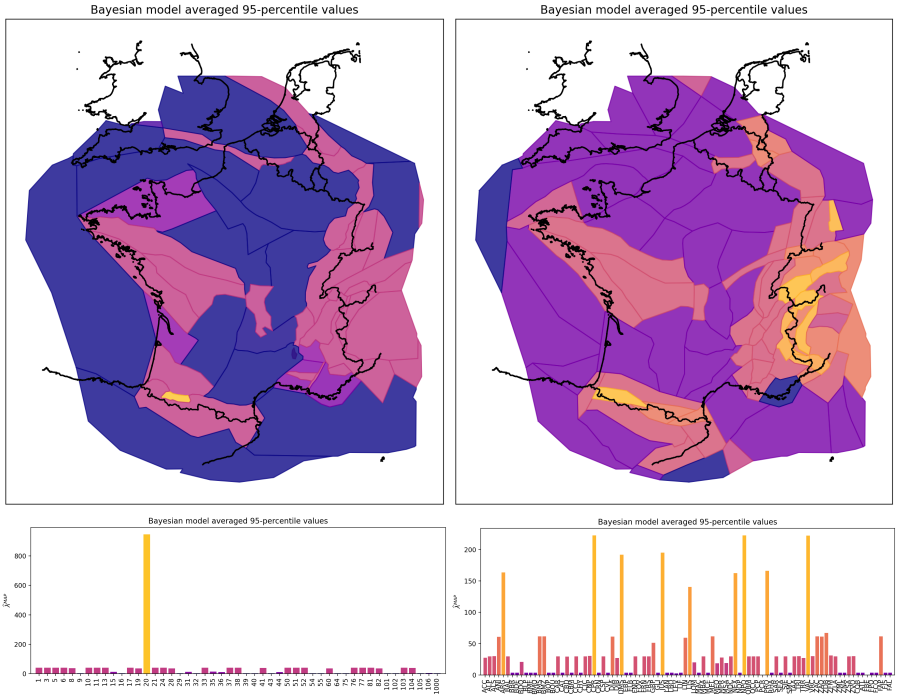


Fig. 12: Posterior, Bayesian-model averaged (BMA) 95% quantile of lambda values, per zone for the EDF (left) and GTR (right) source models.

IRSN with its 3 optimal clusters, and CEA with its 3 optimal clusters. At the end of the day, the EDF model, reduced to 4 optimal clusters, is seen to outperform all other models. A possible explanation is that it identifies with most accuracy the highly active zone in the Pyrenean. However, we have also seen that the GTR model, which ranked second after clustering, seems to detect seismic activity more accurately in the Alpine region.

As discussed previously, these results suggests that there is still some space for improvement in the design and/or selection of zoning models. Building upon the present work, a natural research direction would be to further improving existing source models. A simple idea would be to perform clustering on the intersection of all source models, define by intersecting all possible combinations of zones from the candidate models.

Another path for the improvement of a given seismotectonical zoning, could also consist in considering other recurrence models than the simplistic Gutenberg-Richter law Gutenberg and Richter (1944). A promising candidate could be the recent contribution of Dutfoy (2021), which estimates the maximum magnitude using extreme value theory.

However, whatever the improvement, zoning models still rely on the rather strong assumption that earthquake rates are piecewise constant. Hence, another promising research avenue would be to investigate so-called "zoneless"

source models, which try to fit more general, usually nonparametric, point process models to the earthquake catalog. Many methods have been proposed, from the kernel density estimate approach in Drouet et al (2018), to the full nonparametric Bayesian approach in kolev2020. The latter work also uses self-exciting (Hawkes) processes to model the earthquakes' occurences. This could be another promising research avenue, which allow to analyse the full raw earthquake catalog, with replicas included.

Finally, recent advances on generative modeling through diffusion processes, such as in Cardoso et al (2024), suggest yet another strategy to perform purely data-driven predictions of the seismic hazard. But it is important to keep in mind that such advanced models usually work when applied to large datasets, which is not the case in regions of low seismic rates sur as the French continental domain. This is why leveraging the additional information provided by the zoning models will always be necessary. Hence we advocate the development of a *hybrid* model, which would combine the flexibility of data-driven approaches based, with more informed strategies, based on area source models. The Bayesian framework appears as a natural framework for combining both sources of information.

Acknowledgments. We would like to thank Pr. Jairo Cagliari for fruitful discussions, and more specifically for providing the seminal idea of mixture-model based clustering of seismotectonical zones. This work has been supported by the Nuclear Power Plant of the Future initiative of the French Tripartite Institute for Nuclear Energy.

Appendix A Gutenberg-Richter law and the exponential distribution

The Gutenberg-Richter law is usually written as:

$$log_{10}(N) = a - b.M_w, \tag{A1}$$

where:

- N is the number of earthquakes above magnitude M_w in a given domain D over a certain period;
- $a = Log_{10}(N_0)$, where N_0 is the total number of earthquakes in D;
- b is the slope of the Gutenberg-Richter law for domain D.

Note that Equation (A1) can be re-written as:

$$log_{10}\left(\frac{N}{N_0}\right) = -b.M_w.$$

In the above equation, the left-hand term $log_{10}\left(\frac{N}{N_{tot}}\right)$ is random, since it is the observed proportion of earthquakes with magnitude no less than M_w in D

over a certain period (on a log-scale), whereas the right-hand term $-b.M_w$ is constant.

Hence the equality must not be taken literally, but in an averaged sense, meaning that:

$$Log_{10}\left(\mathbb{E}\left[\frac{N}{N_0}\right]\right) = -b.M_w,$$

where the expectation is taken with respect to the joint sampling distribution of N_0, the total number of earthquakes, and $(m_1, \ldots, m_{N_0})$, their magnitudes, from which N can be computed as:

$$N := N(M_w) = \sum_{i=1}^{N_0} \mathbf{1}_{\{m_i \geq M_w\}}.$$

Hence, conditional on N_0, we have:

$$\begin{aligned}\mathbb{E}\left[\frac{N}{N_0} \mid N_0\right] &= \mathbb{E}\left[\frac{\sum_{i=1}^{N_0} \mathbf{1}_{\{m_i \geq M_w\}}}{N_0} \mid N_0\right] \\ &= \frac{\sum_{i=1}^{N_0} \mathbb{E}\left[\mathbf{1}_{\{m_i \geq M_w\}}\right]}{N_0}.\end{aligned}$$

Now suppose that the magnitudes m_i are independent and identically distributed (iid), with common cumulative distribution function (cdf) $F(M_w) := P[m_i \leq M_w] = 1 - P[m_i \geq M_w]$. Hence:s

$$\begin{aligned}\mathbb{E}\left[\frac{N}{N_0} \mid N_0\right] &= \frac{\sum_{i=1}^{N_0} 1 - F(M_w)}{N_0} \\ &= 1 - F(M_w).\end{aligned}$$

Hence, Gutenberg-Richter is equivalent to stating that:

$$F(M_w) = 1 - 10^{-b.M_w} = 1 - \exp\left(-\log 10.b.M_w\right),$$

meaning that the magnitudes are distributed according to the exponential distribution $\mathscr{E}(\beta)$, with $\beta = \log(10) \times b$.

Appendix B Laplace-guided Importance Sampling

We provide below the mathematical details regarding the importance sampling estimation of the marginal likelihood factors $P(CAT_i|\mathcal{M})$ in Equation (4),

defined by:

$$P(CAT_i|\mathcal{M}) := \int_{\lambda_i,\beta_i} P(CAT_i|\lambda_i,\beta_i,\mathcal{M})\pi(\lambda_i)\pi(\beta_i)d\lambda_i d\beta_i,$$

where $P(CAT_i|\lambda_i,\beta_i,\mathcal{M})$ is the likelihood of the earthquake counts for zone i of source model $\mathcal{M}$, as defined by Equation (1), and $\pi(\lambda_i)\pi(\beta_i)$ are the prior densities defined in Section 2.2.

As shown in Section 2.2, integration with respect to λ_i can be performed analytically, so that the computation of the marginal likelihood actually boils down to the one-dimensional integration problem:

$$P(CAT_i|\mathcal{M}) := \int_{\beta_i} P(CAT_i|\beta_i,\mathcal{M})\pi(\beta_i)d\lambda_i d\beta_i,$$

where the expression of $P(CAT_i|\beta_i,\mathcal{M})$ is given by Equation 5.

The key point in any importance sampling strategy lies in the choice of an instrumental distribution $q(\beta_i)$, according to the two basics requirements:

- q should be "as close as possible" (according to a given metric) to the posterior distribution $\pi(\beta_i|CAT_i,\mathcal{M})$, since the ideal choice $q = \pi$ would lead to an *exact* estimator (with zero variance);
- $q(\beta_i)$ should have fatter tails, and larger support, than $\pi(\beta_i|CAT_i,\mathcal{M})$.

A popular method is Laplace's approximation, which consists in approaching the posterior density by a Gaussian whose expectation is the posterior mode $\widehat{\beta}_i := \arg\max_{\beta_i} \pi(\beta_i|CAT_i,\mathcal{M})$, and whose variance is given by:

$$\widehat{\mathbb{V}}(\beta_i|CAT_i,\mathcal{M})^{-1} := -\frac{\partial^2}{\partial\beta_i^2}\log\pi(\beta_i|CAT_i,\mathcal{M})|_{\beta_i=\widehat{\beta}_i}.$$

Note that, to compute the posterior mode and variance estimates, one only requires to know the expression $\pi(\theta|CAT_i,\mathcal{M})$ up to a multiplicative constant, which is the case here. In practice, a symbolic calculus code was used to perform exact differentiation.

However, Laplace's approximation is Gaussian, which is ill-fitted for β_i's posterior, whose support is a compact interval contained in $\mathbb{R}^+$. Instead, we chose to approach β_i's posterior distribution by a Gamma distribution, whose mode and variance matched $\widehat{\beta}_i$ and $\widehat{\mathbb{V}}(\beta_i|CAT_i,\mathcal{M})$, respectively. Finally, this Gamma distribution was truncated between the chosen bounds for β_i (the default choice being: $\beta_i \in [0.1, 10]$). In cases when the above calculations failed, for instance when the second order derivative was equal to zero, $q(\beta_i)$ was simply taken equal to the prior density $\pi(\beta_i)$, which is also a truncated Gamma distribution.

In all cases, our Gamma-Laplace approximation already provides an estimate of the marginal likelihood, as shown in Kass and Raftery (1995), under

the form of:

$$\widehat{P}(CAT_i|\mathcal{M})^{GL} = \frac{P(CAT_i|\widehat{\beta}_i, \mathcal{M})\pi(\widehat{\beta}_i)}{q(\widehat{\beta}_i)},$$

where $q(\widehat{\beta}_i)$ is the density of the truncated Gamma-Laplace approximation to the posterior distribution, evaluated at the posterior mode. A more accurate importance sampling estimate is then given by:

$$\widehat{P}(CAT_i|\mathcal{M})^{IS} = \frac{1}{K}\sum_{k=1}^{K} \frac{P(CAT_i|\beta_i^{(k)}, \mathcal{M})\pi(\beta_i)^{(k)}}{q(\widehat{\beta}_i^{(k)})},$$

with $(\beta_i^{(k)})_{k=1}^K$ a sample from the instrumental density $q(\beta_i)$.

In practice, the two marginal likelihood estimates $\widehat{P}(CAT_i|\mathcal{M})^{GL}$ and $\widehat{P}(CAT_i|\mathcal{M})^{IS}$ gave almost identical numerical values, showing that our truncated Gamma-Laplace approximation was very accurate indeed. Plus, the effective sample size associated to our importance sampling algorithm Kong (1992) was always very close to the number K of simulated particles.

Appendix C Correcting the likelihood for model selection

We show below that the likelihood of the raw catalog CAT for a given seismotectonic model $\mathcal{M}$ can be factored into:

$$P(CAT|\mathcal{M}, \lambda_{1:I_\mathcal{M}}, \beta_{1:I_\mathcal{M}}) = \frac{e^{-\sum_i[\lambda_i\{\sum_j t_j p_{ij}(\beta_i)\}]}}{n!} \prod_i \left(\frac{\lambda_i}{|R_i|}\right)^{n_i} \prod_{i,j} (t_j p_{ij}(\beta_i))^{n_{ij}} \quad (C2)$$

with $|R_i|$, the area of region R_i. Note that the R_i, λ_i and β_i depend on the chosen model, but we have dropped the k index from the notations for clarity's sake.

In comparison, the likelihood of the reduced catalog $CAT(\mathcal{M})$ for model $\mathcal{M}$ is given by:

$$P(CAT(\mathcal{M})|\lambda_{1:I_\mathcal{M}}, \beta_{1:I_\mathcal{M}}) = \frac{e^{-\sum_i[\lambda_i\{\sum_j t_j p_{ij}(\beta_i)\}]}}{\prod_{i,j} n_{ij}!} \prod_i (\lambda_i)^{n_i} \prod_{i,j} (t_j p_{ij}(\beta_i))^{n_{ij}}$$

It is interesting to note that, though both formulations are statistically equivalent, in the sense that they yield strictly the same posterior distribution for the model parameters $\lambda_{1:I_\mathcal{M}}, \beta_{1:I_\mathcal{M}}$, taking the ratio of the above display with respect to (C2) yields an 'imbalance' factor equal to:

$$\frac{n! \prod_i |R_i|^{n_i}}{\prod_{i,j} n_{ij}!}$$

This factor can be used to correct the likelihood values, as well as the posterior expression in Equation (5), used to compute model weights.

The key to Weichert's model reformulation lies in the fact that it can be interpreted as the aggregation of homogeneous marked spatial Poisson processes attached to each region R_i , with intensity measure $\mathcal{M}_i$ given by (dropping for clarity the seismotectonical model index k):

$$d\mathcal{M}_i(x,y) = \lambda_i \left\{ \sum_j t_j p_{ij}(\beta_i) \right\} \mathbf{1}_{R_i}(x,y) d\mu(x,y),$$

where:

- λ_i is re-defined as the average number of earthquakes per surface unit in region i;
- $d\mu(x,y)$ is the Lebesgue measure on $\mathbb{R}^2$.

The above characterization can easily be justified from (1), using the stability property of the Poisson distribution by variable summation. Likewise, stability of Poisson point processes by superposition means that the complete catalog is a non-homogeneous Poisson point process with intensity measure $\mathcal{M}$ given by:

$$d\mathcal{M}(x,y) = \sum_i \left[\lambda_i \left\{ \sum_j t_j p_{ij}(\beta_i) \right\} \mathbf{1}_{R_i}(x,y) \right] d\mu(x,y)$$

The mark attached to each point (x_k, y_k) of above the Poisson point process is the associated magnitude m_k, whose distribution is defined conditional on each region R_i:

$$\mathbb{P}[m_k = M_j | (x_k, y_k) \in R_i] = \frac{t_j p_{ij}(\beta_i)}{\sum_\ell t_\ell p_{i\ell}}. \tag{C3}$$

Accordingly, the total number n of points, *ie* the length of the catalog, is a Poisson variable with intensity given by the measure $\mathcal{M}$ of the overall region R, yielding:

$$n \sim \mathcal{P}\left(\sum_i \left[\lambda_i \, |R_i| \, \left\{ \sum_j t_j p_{ij}(\beta_i) \right\} \right] \right), \tag{C4}$$

where $|R_i|$ is the Lebesgue measure (*ie* the area) of region i.

Next, each point (x_k, y_k) belongs to, and is uniformly distributed within, regions i, with probability $\frac{\mathcal{M}_i(R_i)}{\mathcal{M}(R)}$, hence follows the following mixture of

uniform distributions:

$$x_k, y_k \overset{iid}{\sim} \frac{\sum_i \lambda_i |R_i| \left\{\sum_j t_j p_{ij}(\beta_i)\right\} \mathcal{U}(R_i)}{\sum_i \left[\lambda_i |R_i| \left\{\sum_j t_j p_{ij}(\beta_i)\right\}\right]}. \tag{C5}$$

Hence, the likelihood of the complete dataset $CAT = \{x_k, y_k, m_k\}_{1 \leq k \leq n}$ can be factored into:

$$\begin{aligned}
P(CAT|\lambda_{1:I}, \beta_{1:I}) &= P(n|\lambda_{1:I}, \beta_{1:I}) \prod_{k=1}^{n} P(x_k, y_k|\lambda_{1:I}, \beta_{1:I})\, P(m_k|x_k, y_k, \lambda_{1:I}, \beta_{1:I}) \\
&= \frac{e^{-\mathcal{M}(R)} \mathcal{M}(R)^n}{n!} \prod_i \left(\frac{\mathcal{M}_i(R_i)}{\mathcal{M}(R)\,|R_i|}\right)^{n_i} \prod_{i,j} \left(\frac{t_j p_{ij}(\beta_i)}{\sum_\ell t_\ell p_{i\ell}(\beta_i)}\right)^{n_{ij}} \\
&= \frac{e^{-\mathcal{M}(R)}}{n!} \prod_i \left(\frac{\mathcal{M}_i(R_i)}{|R_i|}\right)^{n_i} \prod_{i,j} \left(\frac{t_j p_{ij}(\beta_i)}{\sum_\ell t_\ell p_{i\ell}(\beta_i)}\right)^{n_{ij}} \\
&= \frac{e^{-\sum_i \left[\lambda_i |R_i| \left\{\sum_j t_j p_{ij}(\beta_i)\right\}\right]}}{n!} \prod_i \lambda_i^{n_i} \prod_{i,j} \left(t_j p_{ij}(\beta_i)\right)^{n_{ij}}.
\end{aligned}$$

In comparison, the likelihood of the clustered catalog is given by:

$$\begin{aligned}
P(CAT_1, \ldots, CAT_I|\lambda_{1:I}, \beta_{1:I}) &= \prod_{i,j} P(n_{ij}|\lambda_i, \beta_i) \\
&= \prod_{i,j} \frac{e^{-\lambda_i |R_i| t_j p_{ij}(\beta_i)} (\lambda_i |R_i| t_j p_{ij}(\beta_i))^{n_{ij}}}{n_{ij}!} \\
&= \frac{e^{-\sum_i \left[\lambda_i |R_i| \left\{\sum_j t_j p_{ij}(\beta_i)\right\}\right]}}{\prod_{i,j} n_{ij}!} \prod_i (\lambda_i |R_i|)^{n_i} \prod_{i,j} (t_j p_{ij}(\beta_i))^{n_{ij}}.
\end{aligned}$$

References

Aki K (1965) Maximum likelihood estimate of b in the formula log n= a-bm and its confidence limits. Bull Earthquake Res Inst, Tokyo Univ 43:237–239

Baize S, Cushing EM, Lemeille F, et al (2013) Updated seismotectonic zoning scheme of metropolitan france, with reference to geologic and seismotectonic data. Bulletin de la Société Géologique de France 184(3):225–259

Beauval C (2003) Analysis of uncertainties in a probabilistic seismic hazard estimation, example for france. PhD thesis, Grenoble University Joseph Fourier

Beauval C, Bard PY (2021) History of probabilistic seismic hazard assessment studies and seismic zonations in mainland france. Comptes Rendus Géoscience 353(S1):413–440

Beauval C, Scotti O (2004) Quantifying sensitivities of psha for france to earthquake catalog uncertainties, truncation of ground-motion variability, and magnitude limits. Bulletin of the Seismological Society of America 94(5):1579–1594

Bender B, Perkins D (1993) Treatment of parameter uncertainty and variability for a single seismic hazard map. Earthquake Spectra 9:165–195

Berge-Thierry C, Cotton F, Scotti O, et al (2003) New empirical response spectral attenuation laws for moderate European earthquakes. Journal of Earthquake Engineering 7(2):193–222

Bernier J (2003) Décisions et comportement des décideurs face au risque hydrologique / Decisions and attitude of decision makers facing hydrological risk. Hydrological Sciences Journal 43(3):301–316

Bodda SS, Keller, M. AGupta, et al (2021a) A Bayesian approach to estimate weights for GMPE models in a logic tree. Bulletin of Earthquake Engineering (under revision)

Bodda SS, Keller, M. AGupta, et al (2021b) A methodological approach to update Ground Motion Prediction Models using Bayesian Inference. Pure and Applied Geophysics (submitted)

Cardoso G, el idrissi YJ, Corff SL, et al (2024) Monte carlo guided denoising diffusion models for bayesian linear inverse problems. In: The Twelfth International Conference on Learning Representations, URL https://openreview.net/forum?id=nHESwXvxWK

Cornell C (1968) Engineering seismic risk analysis. Bulletin of the Seismological Society of America 58(1):1583–1606

Damblin G, Keller M, Pasanisi A, et al (2014) Approche décisionnelle bayésienne des inceertitudes dans un contexte industriel. Application aux courbes de fragilité sismique. Journal de la Société Française de Statistiques (accepted)

Drouet S, Ameri G, Senfaute G (2018) Seismic hazard maps for the French metropolitan territory. 16th European Conference in Earthquake Engineering, Thessaloniki

Drouet S, Ameri G, Le Dortz K, et al (2020) A probabilistic seismic hazard map for the metropolitan france. Bulletin of Earthquake Engineering 18(5):1865–1898

Dutfoy A (2021) A probabilistic seismic hazard map for the metropolitan france. Pure Appl Geophys 178:1549–1561

Eckert N, Parent E, Faug T, et al (2009) Bayesian optimal design of an avalanche dam using a multivariate numerical avalanche model. Stochastic Environmental Research and Risk Assessment 23(8):1123–1141

Gutenberg B, Richter CF (1944) Frequency of earthquakes in California. Bulletin of the Seismological Society of America 34(4):185–188

Kass RE, Raftery AE (1995) Bayes factors. Journal of the american statistical association 90(430):773–795

Keller M, Pasanisi A, Marcilhac M, et al (2014) A bayesian methodology applied to the estimation of earthquake recurrence parameters for seismic hazard assessment. Quality and Reliability Engineering International (submitted)

Kijko A (1988) Maximum likelihood estimation of gutenberg-richter b parameter for uncertain magnitude values. Pure and Applied Geophysics 127:573–579

Kijko A, Sellevoll MA (1989) Estimation of earthquake hazard parameters from incomplete data files. part i. utilization of extreme and complete catalogs with different threshold magnitudes. Bulletin of the Seismological Society of America 79(3):645–654

Kijko A, Sellevoll MA (1992) Estimation of earthquake hazard parameters from incomplete data files. part ii. incorporation of magnitude heterogeneity. Bulletin of the Seismological Society of America 82(1):120–134

Kijko A, Smit A (2012) Extension of the aki-utsu b-value estimator for incomplete catalogs. Bulletin of the Seismological Society of America 102(3):1283–1287

Kijko A, Vermeulen PJ, Smit A (2022) Estimation techniques for seismic recurrence parameters for incomplete catalogues. Surveys in Geophysics 43(2):597–617

Kolev AA (2020) Extensions of self-exciting point processes with applications in seismology and ecology. PhD thesis, UCL (University College London)

Kong A (1992) A note on importance sampling using standardized weights. Tech. rep., Department of Statistics, University of Chicago

Kramer S (1996) Geotechnical Earthquake Engineering. Prentice Hall

Le Goff B, Fitzenz D, Beauval C (2011) Towards a bayesian seismotectonic zoning for use in probabilistic seismic hazard assessment (psha). In: AIP Conference Proceedings, American Institute of Physics, pp 242–249

Lombardi A, Akinci A, Malagnini L, et al (2005) Uncertainty analysis for seismic hazard in Northern and Central Italy. Annals of Geophysics 48(6):853–865

Lominashvili G, Patsatsia M (2013) On the Estimation of a Maximum Likelihood of Truncated Exponential Distributions. Bulletin of the Georgian National Academy of Sciences 7(1):21–24

Manchuel K, Traversa P, Baumont D, et al (2018) The french seismic catalogue (fcat-17). Bulletin of Earthquake Engineering 16(6):2227–2251

Marin J, Robert C (2007) Bayesian Core: A Practical Approach to Computational Bayesian Statistics. Springer

Marin S, Avouac JP, Nicolas M, et al (2004) A probabilistic approach to seismic hazard in metropolitan france. Bulletin of the Seismological Society of America 94(6):2137–2163

Martin C, Combes P, Secanell R, et al (2002) Révision du zonage sismique de la France. Etude probabiliste. Tech. Rep. GTR/MATE/0701-150, GEOTER

McGuire R (1976) Fortran computer program for seismic risk analysis. Tech. Rep. Open-File Report 76-67, United States Department of the Interior, Geological Survey

Mosca I, Baptie B, Villani M, et al (2019) Objective quantification of the seismic source model for nuclear sites

O'Hagan A (1995) Fractional bayes factors for model comparison. Journal of the royal statistical society series b-methodological 57:99–118

Pasanisi A, Keller M, Parent E (2012) Estimation of a quantity of interest in uncertainty analysis: some help from Bayesian decision Theory. Reliability Engineering and System Safety 100:93–101

Reiter L (1990) Earthquake Hazard Analysis: Issues and Insights. Columbia University Press

Robert C (2007) he Bayesian Choice: From Decision Theoretic Foundations to Computational Implementation. Springer

Robert C, Casella G (2004) Monte Carlo statistical methods. Springer Verlag

Schwarz G (1978) Estimating the Dimension of a Model. The Annals of Statistics 6(2):461 – 464

Selva J, Sandri L (2013) Probabilistic seismic hazard assessment: Combining cornell-like approaches and data at sites through bayesian inference. Bulletin of the Seismological Society of America 103(3):1709–1722

Solomos G, Pinto A, Dimova S (2008) A Review of the seismic hazard zonation in national building codes in the context of Eurocode 8. Tech. Rep. EUR 23563 EN-2008, JRC European Commission

Stucchi M, Rovida A, Gomez Capera A, et al (2013) The share european earthquake catalogue (sheec) 1000–1899. Journal of Seismology 17(2):523–544

Taroni M, Selva J (2021) Gr_est: an octave/matlab toolbox to estimate gutenberg–richter law parameters and their uncertainties. Seismological Research Letters 92(1):508–516

van der Vaart AW (2000) Asymptotic Statistics (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge University Press

Viallet E, Humbert N, Mottier P (2019) Updating a probabilistic seismic hazard assessment with instrumental and historical observations based on a bayesian inference. Nuclear Engineering and Design 350:98–106

Wasserman L (2000) Bayesian model selection and model averaging. J Math Psychol 44(1):92–107

Weichert D (1980) Estimation of the earthquake recurrence parameters for unequal observation periods for different magnitudes. Bulletin of the Seismological Society of America 70:1337–1356

Wiemer S (2001) A software package to analyze seismicity: Zmap. Seismological Research Letters 72(3):373–382

Yaghmaei-Sabegh S, Ostadi-Asl G (2022) Bayesian estimation of b-value in gutenberg–richter relationship: a sample size reduction approach. Natural Hazards 110(3):1783–1797

Yang Y (2005) Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation. Biometrika 92(4):937–950

Zannane S, Keller M (2020) Zoning model selection in PSHA testing. Tech. Rep. 6125-3119-2019-04348-FR, EDF R& D

Zentner I (2010) Numerical computation of fragility curves for NPP equipment. Nuclear Engineering and Design 240(6):1614–1621