Hybrid Models for seismic hazard quantification
PhD subject (2025-2028)

Merlin Keller, Sophie Donnet, Joseph Muré,
Gloria Senfaute, Luis Guillermo Alvarez Sanchez

January 15, 2025

Industrial Context

This subject originates in the fundamental need for EDF to assess the safety of its critical industrial assets (eg. NPP, dam, dyke, . . .) with respect to seismic hazard. A key step of such a risk assessment, known as Probabilistic Seismic Hazard Analysis (PSHA), aims at quantifying the distribution of a certain measure of earthquake intensity, such as the annual peak ground acceleration (PGA), in a given site of interest. Integration with respect to the fragility curve, *ie* the conditional probability of failure of a given industrial component given the PGA, yields the final risk measure.

Available for PSHA analysis are the so-called *instrumental data*, containing the records from the existing seismographic stations. These are available on various periods according to: magnitude range, location, and how the accelerometric networks has grown, from the first experimental seismographes of the XIX-th century to the domestic and international netwroks starting in the 60's.

Thankfully, Instrumental data tend to be richer in frequent earthquakes with small magnitudes (no more than 3), than much less frequent earthquakes with large earthquakes (larger than 6). However, industrial standards for safety analysis typically concern the second kind, evidence of which must be searched from a second source, containing what are collectively termed *historical data*. These span a much higher period, depending on whether we look at: written records (especially since the beginning of the middle ages, circa 179 AC), impact on precarious structures, whether mand-made (churches, bridges, cities, . . .) or natural (trees, rocks, stalactites, stalagmites, . . .), ...

Issues in seismic hazard modeling

The current FCat17 catalog, built in 2017 by EDF and partners within the SIGMA1-2-3 international consortium on PSHA, combines both sources of data, totalizing 10 000 plus seismic events. However, these are plagued with many uncertainties tainting the magnitude, location and depth od each seismic event. In particular, the available magnitudes values are doubly-censored due to limited precision. Additionally, data are missing (according to a poorly known mechanism) outside of so-called *complete observation periods*. These periods quantify the fact that large earthquakes are rarer, but have been observed exhaustively since much longer, than small ones.

To simplify the modeling task, it is common for PSHA to use only data from complete observation periods, and with magnitudes above a minimal threshold, in which case the analysis only concerns a small fraction of the original dataset (about 2 000 events).

At the end of the day, in such a low seismic region as Metropolitean France, only few earthquakes have been measured instrumentally at extreme magnitudes (above 7) corresponding to the hazard that needs to be assessed. Hence, the information provided by the data must necessarily be complemented by the geological expertise underlying the physically-inspired models of the spatio-temporal earthquake occurence, and ground motion propagation, processes. Thus, the resulting seismic hazard assessment is necessarily hybrid, since it must be informed by both data and physical knowledge.

In this PhD work program, we will be focusing on how to build such a hybrid estimation of earthquake occurences, based on both the available earthquake catalog and the geological expert knowledge embedded in current seismic source models.

Previous works and open questions

The classical approach to PSHA model, grounded in the seminal works of [GR44], [Cor68] and [Wei80], depends crucially on a so-called *seismotectonical zoning*, which is a partition of the territory under study into a fixed number of neighbouring regions, assumed to be *seismotectonically homogeneous*. More precisely, the annual number of earthquakes observed within each region, are modelled as Poisson variates. Earthquake magnitudes within each region are then modelled as exponential variates, truncated between user-defined minimal and maximal magnitude values.

This methodology was applied to the French context by [DALD+20], for the needs of the nuclear industry, with the specific difficulty that, instead of a single seismotectonical zoning model, several candidate models where provided by the different institutions involved (EDF, CEA, IRSN, GEOTER). This raised the question of selecting or averaging the predictions arrising from the different candidate models, after there estimation based on the available earthquake catalog.

Meanwhile, in [KPM+14], we have developed a full Bayesian approach to the estimation of the parameters of this Poisson-exponential model based on the available catalog. This was based on an importance sampling algorithm that yielded the posterior distribution within each seismotectonical model, together with an estimate of the marginal likelihood, or evidence, which forms the basis of Bayesian model selection techniques [KR95]. We followed this lead and, in [KZD+24], demonstrated that current seismotectonical models, all comprising over 30 distinct zones, are much too detailed to be estimated accurately based on the available catalog. Extending the model selection framework to allow the clustering of zones sharing common parameter values, we found out that no more than 4 clusters could be accurately estimated from the catalog.

We could simply select the simplified zoning corresponding to the *a posteriori* most probable clusters, and compute the corresponding seismic hazard curve. However the result would still heavily depend on the original partition of the search region according to one or another seismotectonical zoning model. Hence, the main questions we want to answer are:

- Based on the available data, can we validate, or invalidate, the strong assumptions of spatially piecewise homogeneous Poisson underlying current source models, and of time-homogeneous Poisson process assumed for the earthquake occurence models ?
- can we relax these assumptions one way or the other ?

This PhD subject aims at answering the above questions by testing the feasibility of alternative models to describe the spatio-temporal occurence of earthquakes throughout the French continental territory.

Research program

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. Below are several promising research avenues we have identified as being of primary concern for the PhD research effort.

Improving recurrence modeling using Hawkes processes

Another path for the improvement of a given seismotectonical zoning, could also consist in considering other recurrence models than the simplistic Gutenberg-Richter law [GR44]. A promising candidate could be the recent contribution of [Dut21], which estimates the maximum magnitude using extreme value theory. However, such estimates are still based on the key assumption that earthquake occurences are independant from one another. In particular, this entails that all earthquake replicas must be removed from the catalog prior to the analysis. This pre-treatment necessarily reduces the amount of available data, while presenting the risk of removing, either too many or too few, earthquakes from the catalog.

Hence, a first direction we wish to investigate would consist in explicitly modeling earthquake replicas, rather than trying to remove them. Self-exciting point processes, in particular Hawkes processes [HAW71], would be good candidates. These have been used

However, one would need to account for the incomplete and censored nature of the available data, as previously described. Several solutions have been proposed in the litterature, such as in [DSL19],

[DR21] or [SQS18], to cite but a few. a first objective for the PhD candidate would be to select, or develop if necessary, an earthquake recurrence model tailored to the available catalog, replicas included.

Spatial occurence modeling using Dirichlet process

However, whatever the improvement on modeling the temporal reccurence of earthquakes, 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 [DALD+20], to the full nonparametric Bayesian approach in [Kol20]. The latter work is based on the formalism of Dirichlet processess [Fer73], which can be used as a Bayesian extension to kernel density estimation, allowing in particular to include prior information, such as contained in the classical zoning models. As previously, learning such a process in the incomplete data setting of PSHA will be one of the main technical challenges.

Combine and compare different strategies

Having conducted research in the two above-described directions to propose novel seismic source models, the last task of the PhD project will be to compare these to current standard models, while at the same time trying to combine them in a single coherent spatio-temporal source model if possible, as done previously in [Kol20]. The task of model comparison, or model selection, viewed from the traditional Bayesian viewpoint of Bayes factor comparison, is in itself a challenging task [KR95], and has at our knowledge not yet been applied to point process model selection, especially in the presence of censored data. MCMC-based methods, exploiting the link between model selection and mixture modeling, we recently tested in [KMRR18, KK18], could be tested to adress this problem.

Practical aspects

Applications

As discussed above, this work contrtibutes to strengthen the theoretical and methodological foundations of the safety demonstration of critical infrastructures of the French energy production industry. More specifically, the PhD candidate will be working with the French earthquake catalog (FCat) and current source models, with the goal of validating and / or improving these models based on the information provided by the data.

However, it is highly recommended not to limit the applied part of this work to a single dataset, but to consider alternative, either real-life or simulated, datasets. In particular, we consider that for a PhD in applied maths, simulation studies are an essential part of the research process: they allow to demonstrate the validity and practical advantages of the proposed approaches, perform a sanity check of their implementation, and help identify their limitations and possible future developments.

Work environment

This PhD is funded by a CIFRE contract of ANRt between AgroParisTech and EDF RD. The PhD director, Sophie Donnet, Professor in Statistics at the AgroParisTech school, is part of the MIA Paris-Saclay research team, depending of the Jacques Hadamard doctoral school. The supervising team at EDF RD will be constituted of four research engineers, Merlin Keller and Joseph Muré, two statisticians from the PRISME department, as well as Gloria Senfaute and Luis Guillermo Alvarez-sanchez, two seismologists from the ERMES department. As such, the phD candidate would benefit from a desk in AgroParisTech, EDF Lab Saclay (both in Palaiseau), and EDF Lab Chatou (Chatou).

This work also benefit from the support of the French Commission for Atomic Energy (CEA), where it will be followed by Clément Gauchy, as it arises from a collaborative work program of the Nuclear Tripartite Institute (I3P), regrouping EDF, CEA and Framatome.

Finally, it will also be part of the Scientific Interest Group (GIS) Lartisste initiative for uncertainty quantificatyion, regrouping major industrial and academic actors centered on the Saclay region concerned with machine learning and artificial intelligence, of which AgroParisTech, EDF and CEA are active members.

Requested skills

This PhD is open to all candidates with a master degree (M2) in applied maths with a strong theoretical and / or computational background in probability and statistics. Familiarity with Bayesian methods, as well as stochastic point processes modeling, would be an advantage.

As is increasingly common in industrial studies, the methods will be developed in the Python language. Previous experience with the Python language, collaborative development (using, *eg*, GIT) and various operating systems (especially Linux) would be an advantage.

Finally, an appetite for applications, data analysis, and working in strong interaction with other disciplines, is also strongly desired, given the centrality of the case-study in the motivation of this work.

References

[Cor68] C.A. Cornell. Engineering seismic risk analysis. *Bulletin of the Seismological Society of America*, 58(1):1583–1606, 1968.

[DALD$^+$20] Stéphane Drouet, Gabriele Ameri, Kristell Le Dortz, Ramon Secanell, and Gloria Senfaute. A probabilistic seismic hazard map for the metropolitan france. *Bulletin of Earthquake Engineering*, 18(5):1865–1898, 2020.

[DR21] Isabella Deutsch and Gordon J. Ross. Abc learning of hawkes processes with missing or noisy event times, 2021.

[DSL19] J. Derek Tucker, Lyndsay Shand, and John R. Lewis. Handling missing data in self-exciting point process models. *Spatial Statistics*, 29:160–176, 2019.

[Dut21] Anne Dutfoy. A probabilistic seismic hazard map for the metropolitan france. *Pure Appl. Geophys.*, 178:1549–1561, 2021.

[Fer73] Thomas S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209 – 230, 1973.

[GR44] B. Gutenberg and C F. Richter. Frequency of earthquakes in California. *Bulletin of the Seismological Society of America*, 34(4):185–188, 1944.

[HAW71] ALAN G. HAWKES. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 04 1971.

[KK18] Merlin Keller and Kaniav Kamary. Bayesian model averaging via mixture model estimation, 2018.

[KMRR18] Kaniav Kamary, Kerrie Mengersen, Christian P. Robert, and Judith Rousseau. Testing hypotheses via a mixture estimation model, 2018.

[Kol20] Aleksandar Atanasov Kolev. *Extensions of self-exciting point processes with applications in seismology and ecology.* PhD thesis, UCL (University College London), 2020.

[KPM$^+$14] Merlin Keller, Alberto Pasanisi, Marine Marcilhac, Thierry Yalamas, Ramòn Secanell, and Gloria Senfaute. A bayesian methodology applied to the estimation of earthquake recurrence parameters for seismic hazard assessment. *Quality and Reliability Engineering International*, 30:921–933, 2014.

[KR95] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[KZD$^+$24] Merlin Keller, Sanaa Zannane, Clara Duverger, Gloria Senfaute, and Jessie Mayor. Bayesian estimation and selection of seismic source models. *Pure and Applied Geophysics (submitted)*, 2024.

[SQS18] Christian Shelton, Zhen Qin, and Chandini Shetty. Hawkes process inference with missing data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32, 04 2018.

[Wei80] D.H. Weichert. Estimation of the earthquake recurrence parameters for unequal observation periods for different magnitudes. *Bulletin of the Seismological Society of America*, 70:1337–1356, 1980.