# Projet de Machine Learning
# Prévoir les séismes : une vanité à l'ère de l'IA ?

Clara Carlier Lucas De Lara

Mars 2020

## Table des matières

# 1 Introduction

Le 19 septembre 1985, et exactement 32 ans plus tard, le 19 septembre 2017, la ville de Mexico fut frappée par deux séismes meurtriers. Bien que les sismologues s'attendaient depuis longtemps à ce qu'un tremblement de terre de grande ampleur se produise, le lieu et la date du choc de 2017 a surpris tout le monde.

C'est un consensus : nous sommes aujourd'hui incapable de prédire les tremblements de terre, ou du moins pas avec une assez grande précision pour anticiper les plus destructeurs. L'Indonésie et le Japon à l'été 2018 ou la Turquie en ce début d'année sont quelques exemples parmi beaucoup qui montrent notre incapacité à prédire les séismes dans des zones notoirement à risque.

Parmi les méthodes qui tentent de prédire de nouveaux évènements, nous distinguons l'analyse des risques sismiques et l'analyse des précurseurs. Le risque sismique est cartographié par notre connaissance de la tectonique des plaques, et par la récurrence et l'intensité des tremblements de terre dans le monde. Ce risque est un indicateur d'occurrences sur de grandes fenêtres temporelles. Les précurseurs sont des signaux avant-coureurs, comme par exemple des ondes se propageant dans la croûte terrestre, dont la compréhension permettrait d'alerter à court-terme du déclenchement d'un tremblement de terre.

Face au manque de théorie fiable sur la sismicité, et en raison de l'essor de l'intelligence artificielle, de plus en plus de méthodes analysent risques et précurseurs à l'aide du machine learning. Depuis quelques années, sismologues, statisticiens et amateurs proposent de nombreux data challenges pour prédire les séismes à différentes échelles spatio-temporelles.

C'est un dataset trouvé sur Kaggle qui lança ce projet. Notre étude se limite à l'évaluation du risque sismique à partir d'historiques de tremblements de terre dans le monde. Son intérêt est double : sur le plan pédagogique, l'étude de séismes demandent une analyse en temps et en espace qui exigent une approche plus originale que l'application de méthodes de régression classique ; sur le plan des enjeux, ce projet permet de tester et d'évaluer les performances de modèles statistiques récents, dans l'espoir de développer des méthodes pratiques et efficaces.

Plus précisément, nous allons tenter de répondre à la question suivante : combien de séismes sont à prévoir dans le futur ?

## 2 Analyse descriptive en temps et en espace

Nous commençons par réaliser une description des données que nous avons à notre disposition. Sont répertoriés tous les séismes de magnitude supérieure à 5.5, de Janvier 1965 à Décembre 2016 à l'échelle mondiale. Les informations fournies sont : l'heure, la date, la magnitude, la latitude, la longitude, la profondeur ainsi que d'autres variables moins importantes. Notons que les séismes de magnitude supérieure ou égale à 5.5 sont dits *modéré* et causent des dommages visibles sur les constructions (voir échelle de Richter en annexe A), il est donc assez aisé de les percevoir dans des **zones habitées**. De plus, la magnitude maximale dans ce jeu de données est de 9.1.

### 2.1 Planifier les séismes

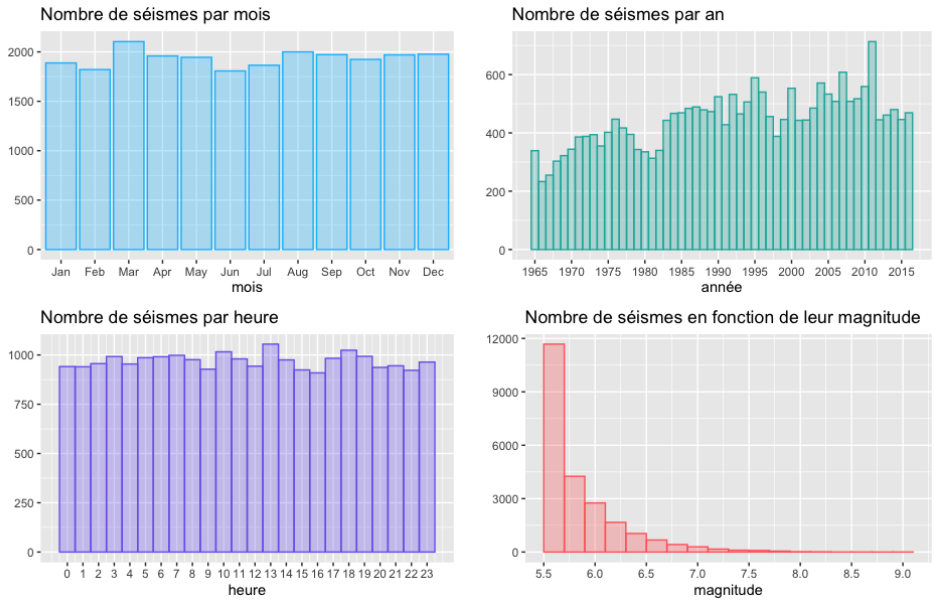



FIGURE 1 – Histogrammes liés au nombre de séismes

Dans les trois premiers histogrammes, nous avons représenté l'évolution du nombre de séismes en fonction de variables temporelles (heure, mois, année) afin d'identifier une saisonnalité. La fréquence des séismes est uniforme pour les heures et les mois. L'analyse selon le mois de l'année ou le jour de la semaine se prête naturellement à l'étude de phénomènes liés aux activités humaines, mais pas nécessairement à des phénomènes géologiques comme les séismes. C'est pourquoi nous procéderons à une analyse plus générale à l'aide de méthodes pour séries temporelles.

Par ailleurs, nous notons une légère tendance croissante pour le nombre de séismes comptabilisés par année. Ce phénomène peut s'expliquer de deux façons différentes :

1. L'interprétation la plus probable serait que l'amélioration et l'augmentation du nombre de capteurs au cours des époques permettraient de capter un plus grand nombre de séismes, par exemple, se produisant dans des **zones non habitées**.
2. L'autre possibilité serait que le nombre de séismes a effectivement augmenté au cours du temps.

Dans le cadre du projet, nous travaillons sur des données récoltées à des périodes différentes, dans des lieux différents par des personnes différentes. Par conséquent, il existe une inconsistance dans les

données qu'il n'est pas facile de gérer en l'absence de méthodes statistiques robustes. Par précaution, il sera bon de ne pas considérer, ou d'accorder une faible confiance, aux données les plus anciennes.

Nous allons par la suite étudier plus en détails la répartition du nombre de séismes par an en fonction de leur magnitude.

Le quatrième et dernier histogramme décrit bien une évidence : plus la magnitude du séisme est élevée, plus il est rare. Nous verrons plus loin que cette tendance est prise en compte par une densité à décroissance exponentielle dans le modèle théorique ETAS.

Nous étudions désormais la répartition du nombre de séismes par an plus en détails. Pour cela, dans les figures 2 et 3, nous distinguons les séismes par leur magnitude. Du bas vers le haut sur les barres de l'histogramme, il s'agit de la magnitude appartenant à : $[5.5, 5.6]$, $(5.6, 5.7]$, $(5.7, 5.8]$, $(5.8, 6.0]$ et $(6.0, 9.2]$.

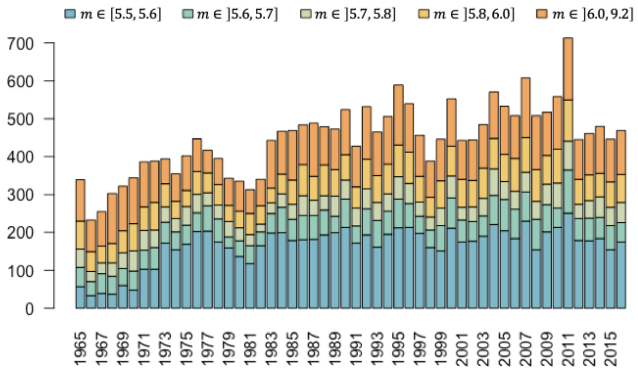



FIGURE 2 – Nombre de séismes par années en fonction de leur magnitude

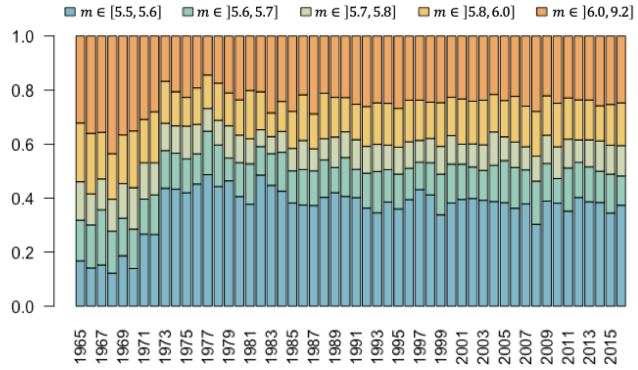



FIGURE 3 – Pourcentage de séismes par années en fonction de leur magnitude

Dans la figure 2, nous remarquons une augmentation des séismes de magnitude comprise entre 5.5 et 5.6 durant les huit premières années, cependant par la suite leur nombre est assez stable et semble varier de façon régulière. Cela rejoint notre précédente remarque : il était plus difficile par le passé de répertorier les séismes d'intensités les plus faibles, notamment loin des zones habitées. Quant aux

autres tranches de magnitude, le nombre d'occurrences ne semble pas changer significativement.

La figure 3 nous confirme tout d'abord qu'il y a une réelle différence durant les huit premières années par rapport aux années suivantes. De plus, les proportions des séismes par rapport à leur magnitude semblent se conserver de 1973 à 2016. Le nombre total de séismes augmente donc de façon proportionnelle depuis 1973.

Nous concluons que les distributions du nombre total de séismes et du nombre de séisme par magnitudes sont stationnaires. Nous interprétons la phase de stabilisation sur les huit premières années comme un phénomène lié à l'hétérogénéité des données plutôt que comme une tendance sismique. Nous n'avons pas de raison de penser que le nombre de séisme serait resté bas durant toute l'histoire de la croûte terrestre, avant de quadrupler sur quelques années au XXe siècle.

Cette première analyse en temps nous a permis de dégager l'absence de tendance des séismes. Avant de réaliser une analyse plus fine de la périodicité, nous allons d'abord aborder l'analyse en espace de ce phénomène extrêmement localisé à l'échelle mondiale.

## 2.2 Cartographier les séismes

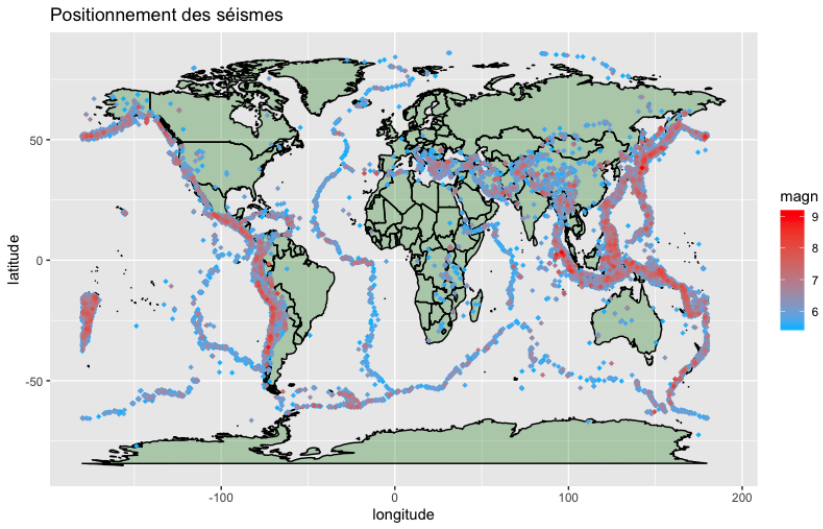



FIGURE 4 – Positionnement des séismes de nos données en fonction de leur magnitude

Pour l'étude géographique, nous commençons par représenter la position des séismes sur un planisphère en fonction de leur magnitude. Sans surprise, nous remarquons qu'une grande partie se situe le long des frontières des plaques. Nous avons un total de 23 232 séismes répertoriés ce qui constitue un nombre assez élevé d'entrées. De plus, prédire les séismes à trop grande échelle n'a pas beaucoup d'intérêt. Les zones à risques sont déjà connues et ne constituent qu'un étroit réseau au milieu de vastes zones inhabitées et/ou peu sensibles. Pour ces raisons, nous allons par la suite distinguer des zones d'intérêt et réaliser nos prédictions sur ces sous-groupes. Sur la carte de la figure 4, nous remarquons certaines zones où les séismes de magnitude élevée (en rouge) sont plus présents. Nous décidons alors de délimiter ces zones-ci afin de constituer nos sous-groupes.

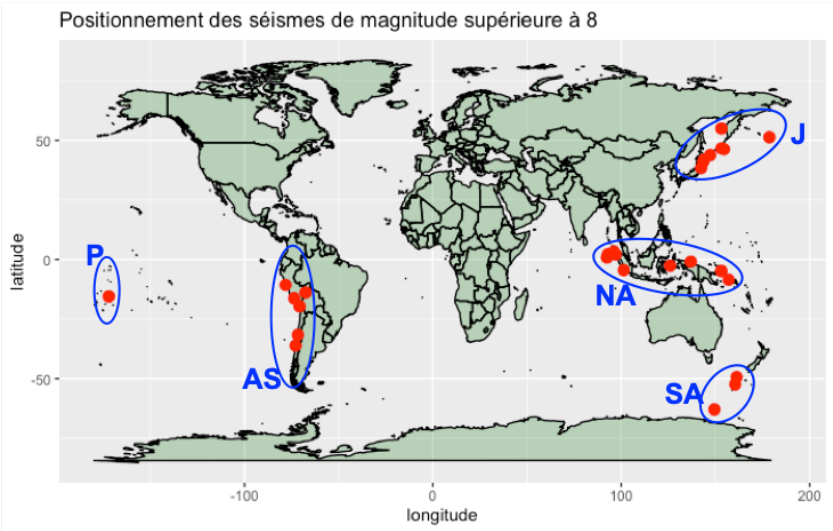



FIGURE 5 – Positionnement des séismes ayant une magnitude supérieure à 8

Dans la figure 5, nous représentons la position des séismes ayant une magnitude supérieure à 8. Nous distinguons alors cinq zones principales : une dans l'Océan Pacifique (**P**), la côte Ouest de l'Amérique du Sud (**AS**), le Sud de l'Australie (**SA**), le Japon (**J**) et une zone constituée des îles au nord de l'Australie (Indonésie, Papouasie-Nouvelle-Guinée, ...) (**NA**).
Notons que les zones **P** et **SA** sont des zones très peu voir pas habitées, prédire des séismes dans ces endroits a donc un intérêt limité.

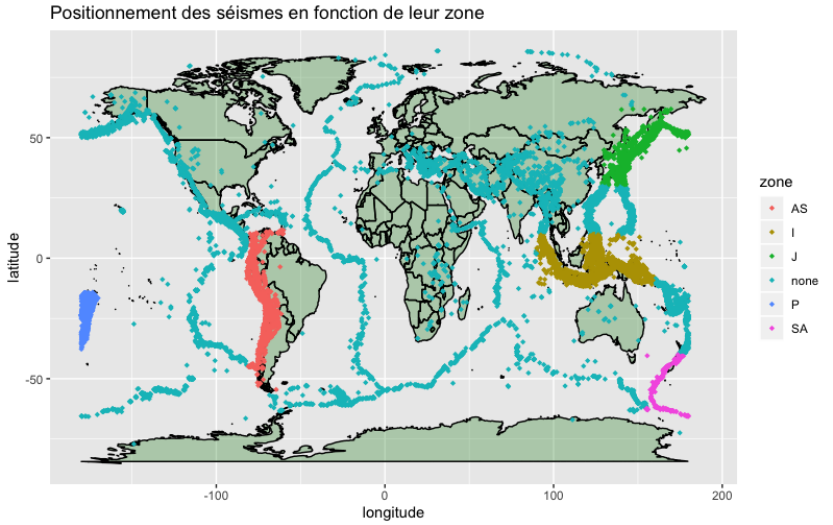


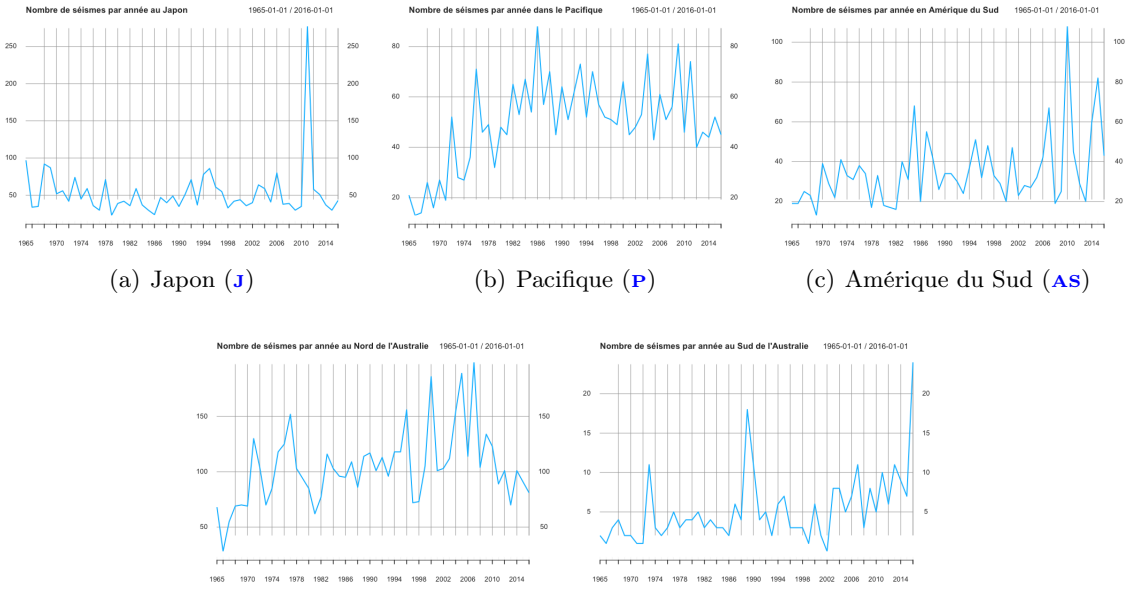

FIGURE 6 – Positionnement des séismes en fonction des zones attribuées

Nous délimitons les zones d'intérêt, elles sont représentées dans la figure 6. Ainsi, nous pouvons constituer cinq bases de données distinctes sur lesquelles nous allons réaliser nos prédictions.

Nous allons désormais réaliser une analyse plus fine à partir des séries temporelles du nombre de séismes par année pour chaque zone.

## 2.3 Étude des données en séries temporelles

Il pourrait sembler naturel de tenter de modéliser l'évolution du nombre de séismes par des séries temporelle - ce que suggèrent certains data challenges. Cependant, la communauté scientifique est quasi-unanime sur le fait que les séismes ne présentent pas de périodicité ni de tendance. Nos données confirment ce point. En effet, nous avons construit des vecteurs comptant le nombre de séismes répertoriés par an pour chaque zone d'intérêt puis nous avons calculé l'ACF et la PACF associées à chaque zone.




(a) Japon (**J**) (b) Pacifique (**P**) (c) Amérique du Sud (**AS**)

(d) Nord de l'Australie (**NA**) (e) Sud de l'Australie (**SA**)

FIGURE 7 – Évolution du nombre de séismes par année en fonction de la zone

Dans la figure 7, quelque soit la zone d'intérêt, nous ne remarquons pas la présence d'une quelconque saisonnalité. De même pour la tendance : hormis pour les zones **P** et **NA**, les séries sont plutôt stables dans le temps. Dans la figure 8, pour le Japon (**J**), nous remarquons qu'il n'y pas d'auto-corrélation élevée : elles sont toutes non significatives. De même pour les auto-corrélations partielles : elles sont toutes non significatives. Cette assertion est également vérifiée pour les zones **AS** et **SA**.

Nous réalisons également un test de Ljung-Box sur nos cinq séries temporelles. Ce dernier a comme hypothèses :

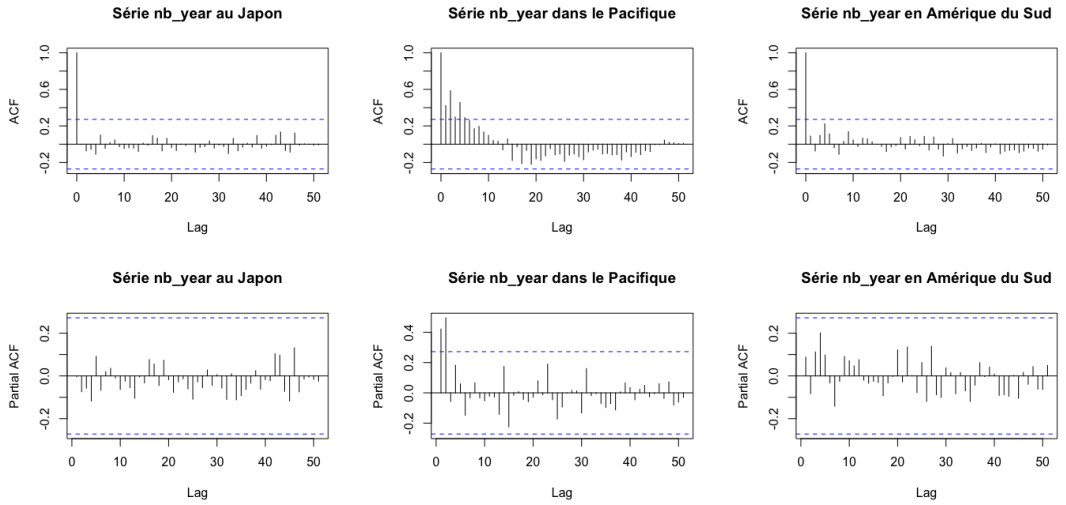
$$\mathcal{H}_0 : \text{il n'y a pas d'auto-corrélation entre les observations}$$
$$\mathcal{H}_1 : \text{il y a auto-corrélation entre les observations}$$

Les $p$-valeurs obtenues sont données dans le tableau 1. Elles confirment nos premiers résultats : les observations des zones **J**, **AS** et **SA** ne sont pas auto-corrélées, en effet avec ces $p$-valeurs nous conservons $\mathcal{H}_0$. Cependant, pour **P** et **NA**, leur $p$-valeurs sont basses et nous rejetons $\mathcal{H}_0$. Ce test doit être appliqué à des séries stationnaires or celle de ces deux zones ne le sont pas. Nous allons alors par la suite leur soustraire leur tendance.

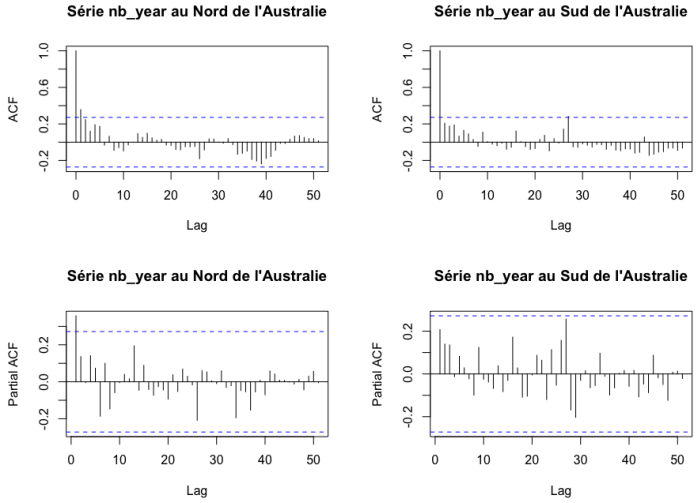
Nous cherchons alors à vérifier que les auto-corrélations restantes dans les zones **P** et **NA** sont présentes à cause de la tendance de leur série. Pour cela, nous allons utiliser `ksmooth` afin d'estimer leur tendance puis la soustraire afin d'étudier l'ACF et la PACF des nouvelles séries. Dans la figure 9, nous avons représenter les séries soustraites par leur tendance ainsi que les nouvelles ACF et PACF associées. Pour le Nord de l'Australie, il n'y a désormais plus d'auto-corrélation. Pour le Pacifique, il reste deux auto-corrélations significatives mais petites.

| Zone | **J** | **P** | **AS** | **NA** | **SA** |
|---|---|---|---|---|---|
| $p$-valeurs | 0.980 | 0.002 | 0.514 | 0.008 | 0.124 |

TABLE 1 – $p$-valeurs des tests de Ljung-Box pour chaque zone

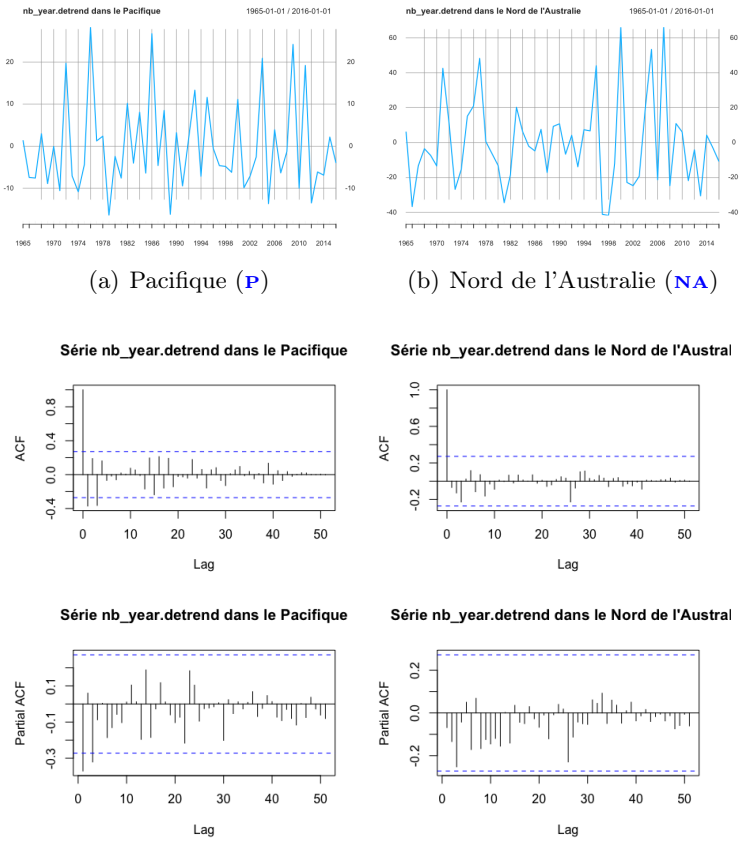



(a) Japon (**J**) (b) Pacifique (**P**) (c) Amérique du Sud (**AS**)




(d) Nord de l'Australie (**NA**) (e) Sud de l'Australie (**SA**)

FIGURE 8 – ACF et PACF du nombre de séismes par année en fonction de la zone




(a) Pacifique (**P**) (b) Nord de l'Australie (**NA**)

(c) Pacifique (**P**) (d) Nord de l'Australie (**NA**)

FIGURE 9 – Séries du nombre de séismes par an corrigées par leur tendance pour les zones **P** et **NA**

De cette analyse en temps, nous retenons que les séismes ne présentent pas d'auto-corrélation ni de saisonnalité. Nous retrouverons cette caractéristique dans le modèle que nous utiliserons pour faire nos prédictions.

Après ces analyses et cette restructuration des données, nous allons désormais présenter notre choix de modèle pour la prédiction.

# 3 Quels modèles pour la prédiction ?

## 3.1 Des méthodes classiques inadaptées

Nous avons déjà écarté la modélisation par des séries temporelles, mais il s'avère par ailleurs que l'étude d'historiques sismiques ne peut se faire à l'aide de méthodes classiques d'apprentissage machine comme de la régression par exemple.
Rappelons que nos observations sont des vecteurs de la forme :

(`date`, `latitude`, `longitude`, `magnitude`)

Il y a une observation par séisme. Fatalement, comme un historique n'encode pas le fait qu'un séisme n'a pas eu lieu, nous ne pouvons pas opérer de régression pour prédire le nombre d'évènements. De la même façon, prédire la magnitude de futurs séismes en fonction de la date et du lieu interpolerait les données et conduirait à une surestimation de la sismicité dans des zones peu sensibles.

Sinon, nous pourrions transformer nos données et notre problème en agrégeant le nombre de séismes par période et par zone, mais cela revient à penser en terme de séries temporelles.

Nous tenons à souligner ici que nous ne sommes pas dans un problème en grande dimension. Aucune sélection de variables n'est à prévoir, car toutes peuvent expliquer l'occurrence sismique comme nous l'avons vu lors de notre analyse descriptive. En l'absence de plus de covariables possiblement explicatives (nature du sol par exemple), il faudra utiliser des modèles qui puissent capter une structure spécifique à la sismicité.

Le support couramment utilisé pour l'étude des séismes est le support des processus ponctuels, au sein duquel le modèle dit ETAS se distingue.

## 3.2 Le modèle ETAS

Le *epidemic-type aftershock sequence* [1] s'est imposé comme le modèle de référence au cours des dernières années. Il s'agit d'un processus de Poisson que l'on caractérise par son intensité conditionnellement à un historique de séismes. Nous nous intéressons ici à la version la plus complète de ETAS, dans laquelle le processus est inhomogène en temps et en espace. L'idée essentielle du modèle est que les séismes sont issus de deux phénomènes : d'une sismicité stationnaire fonction de l'espace, sensé rendre compte de la géologie et de la topographie du lieu ; d'un risque de réplique qui dépend de l'apparition des précédents séismes proches en temps et en espace. L'estimation du modèle consiste à approcher cette intensité conditionnelle par maximum de vraisemblance. Des packages `R` tel que le package `ETAS` permettent d'entraîner un modèle ETAS sur un historique.

### 3.2.1 Principe et théorie

Nous disposons d'un $N$-échantillon $\{(t_i, x_i, y_i, m_i) : i = 1, \dots, N\}$ (ou catalogue) d'un processus $X$ comptant le nombre de séismes selon la date $t$, la position $(x, y)$ et la magnitude $m$. L'historique à la date $t$ est défini par : $H_t = \{(t_i, x_i, y_i, m_i) \in X : t_i < t\}$. L'intensité conditionnelle de $X$ est alors la fonction : $(t, x, y, m) \mapsto \lambda(t, x, y, m|H_t)$.

Sans entrer dans les détails, le modèle ETAS paramétrise l'intensité à l'aide d'un réel $\beta$ et d'un vecteur $\theta = (\mu, A, \alpha, c, p, D, \gamma, q)$. La fonction se décompose alors :

$$\lambda_{\beta,\theta}(t, x, y, m|H_t) = \nu_\beta(m)\lambda_\theta(t, x, y|H_t) \tag{1}$$

où $\nu_\beta(m) = \beta \exp(-\beta(m - m_0))$ est la densité du nombre de séisme de magnitude $m$. Comme uniquement les séismes les plus dangereux nous intéressent, on peut choisir un seuil $m_0$ au dessous duquel on ignore les évènements.
Le terme $\lambda_\theta(t, x, y|H_t)$ est la somme de deux termes qui traduisent le principe de ETAS :

- Le *background sismicity rate* $\tilde{u}(x, y)$ ne dépend que de la location et prend la forme semi-paramétrique $\mu u(x, y)$ où $u$ est régulière.
- Le *aftershock sismicity rate* $\sum_{t_i<t} \kappa_{A,\alpha}(m_i) g_{c,p}(t - t_i) f_{D,\gamma,q}(x - x_i, y - y_i, m_i)$, où $\kappa_{A,\alpha}(m_i)$ est le nombre moyen de répliques générées par un séisme de magnitudes $m_i$, $g_{c,p}(t - t_i)$ est la densité du temps d'occurrence d'une réplique issue d'un séisme à la date $t_i$ et $f_{D,\gamma,q}(x - x_i, y - y_i, m_i)$ est la densité d'occurrence d'une réplique d'un séisme de magnitude $m_i$ qui a eu lieu en $(x_i, y_i)$.
- Pour des raisons de lisibilité, on notera $v(t, x, y|H_t)$ l'intensité des répliques, de sorte que : $\lambda_\theta(t, x, y|H_t) = \tilde{u}(x, y) + v(t, x, y|H_t)$.

Le lecteur intéressé peut se reporter à l'article [1] pour connaître le détail de la fonction $v(t, x, y|H_t)$. Dans le cadre de ce projet il faut retenir que chacun de ses termes décroît rapidement lorsque $(t, x, y)$ s'éloigne de $(t_i, x_i, y_i)$, et qu'elle n'inclut aucune périodicité ou de tendance a priori.

Par ailleurs, l'estimation des paramètres du modèle ETAS sur un catalogue de séismes permet aussi d'estimer la probabilité qu'un évènement soit une réplique ou non. Nous utiliserons cette probabilité afin de classifier les séismes.

L'estimation se fait par maximum de vraisemblance avec plusieurs approximations, notamment un maillage de l'espace. Naturellement, de par sa nature semi-paramétrique et multidimensionnelle, la procédure d'estimation est extrêmement coûteuse computationnellement. La procédure `R` que nous utilisons pour le projet converge généralement en plusieurs dizaines de minutes sur des petits jeux de données (quelques centaines de séismes) et est extrêmement sensible au choix du vecteur initial des paramètres $\theta_0$.

### 3.2.2 Ce que permet le package `ETAS`

Le package `ETAS` [1] permet à partir d'un historique de séismes, encodé sous la forme d'un objet appelé catalogue de séismes, d'estimer un modèle ETAS. Concrètement, une fonction implémente l'estimation par maximum de vraisemblance pour renvoyer un objet de type `etas.fit`. Différentes méthodes permettent alors d'en extraire :

- Les paramètres $\beta$ et $\theta$.
- La fonction qui pour $(t, x, y, \theta)$ renvoie :

$$v(t, x, y|H_t) = \sum_{t_i<t} \kappa_{A,\alpha}(m_i) g_{c,p}(t - t_i) f_{D,\gamma,q}(x - x_i, y - y_i, m_i) \tag{2}$$

- La fonction qui, à partir d'une discrétisation choisie de l'espace, renvoie $\tilde{u}(x, y)$ sous la forme d'une matrice.
- Les probabilités de *clustering* des évènements du catalogue. Autrement dit la probabilité pour chaque séisme d'être une réplique ou au contraire un *background*.
- D'autres méthodes statistiques, des `plot` de graphiques...

Notons que **la méthode ne permet pas de simuler un processus ETAS** à partir d'une fonction $\tilde{u}$ et de paramètres $\beta$ et $\theta$ connus, et ne contient pas de méthode `predict`. Nous devrons donc

mettre en place notre propre procédure.

Par ailleurs, la procédure d'estimation s'avère compliquée à utiliser dans la pratique. Outre son temps de convergence long, elle est extrêmement sensible à la valeur initiale $\theta_0$. Il faut donc chercher, dans la littérature par exemple, des valeurs plausibles des paramètres ETAS pour chaque région du monde que nous considérons.

# 4 Notre méthode de prédiction

Nous suivons la démarche suivante pour prédire des séismes : pour une zone géographique donnée, nous divisons le catalogue en une période passée qui sert à l'entraînement (autrement dit l'estimation des paramètres) et une période future sur laquelle nous testerons nos prédictions.

Nous étudions deux risques : le risque global sur de grandes périodes de prédiction et le risque de réplique sur de courtes périodes. Le modèle ETAS sera utilisé pour toutes ces prédictions.

Notons que nous ne nous intéresserons ici qu'aux séismes de magnitude supérieure ou égale à 5.0. Cela nous semble être un bon compromis entre avoir suffisamment de données (rappelons que le nombre de séismes décroît exponentiellement avec la magnitude), et faire des prédictions utiles (le lecteur intéressé peut se reporter à l'échelle de Richter en annexe). De plus, considérer un seuil minimum permet de limiter les effets d'incomplétion des données. Nous tenons à le rappeler, par le passé et dans certaines régions du globes inhabitées, tous les séismes ne pouvaient probablement pas être détectés.

## 4.1 Méthode d'estimation

Nous disposons d'un catalogue (ou historique) de séismes noté $H_T$ sur une période $[t_0, T]$. Considérons une zone $S$ et une date $t^*$ en amont de laquelle nous avons entraîné un modèle ETAS. Autrement dit, le modèle est entraîné sur les séismes appartenant à $[t_0, t^*]$. Soient de plus des dates initiale $t_i$ et finale $t_f$ de sorte que $[t_i, t_f]$ définisse la période de prédiction, nécessairement postérieures à $t^*$. En résumé on a $t_0 < t^* \leq t_i < t_f \leq T$ (nous pouvons envisager $t^* = t_i$ et $t_f = T$). Notre objectif est de déterminer le nombre de séismes qui aura lieu durant l'intervalle $[t_i, t_f]$.

Selon le modèle ETAS, nous obtenons le nombre $n$ d'occurrences sur la zone $S$ durant la période $[t_i, t_f]$ de la façon suivante :

$$n = \int_{m_0}^{+\infty} \int_{t_i}^{t_f} \int_S \lambda(t, x, y, m)\, dt\, dx\, dy\, dm \tag{3}$$

La difficulté de l'implémentation vient du fait que l'on dispose uniquement de la densité conditionnellement à un historique passé $(H_t)$ et non pas la densité elle-même. Le calcul de :

$$\int_{m_0}^{+\infty} \int_{t_i}^{t_f} \int_S \lambda(t, x, y, m|H_t)\, dt\, dx\, dy\, dm \tag{4}$$

n'a pas de sens dès lors que $(H_t)$ est une variable aléatoire et qu'en pratique on ne connaît au mieux que jusqu'à $H_{t^*}$ avant de faire nos prédictions.

Afin de mener l'implémentation malgré tout, nous ferons l'approximation :

$$\forall t \geq t^*, \quad \lambda(t, x, y, m) \approx \lambda(t, x, y, m|H_{t^*}) \tag{5}$$

où $H_{t^*}$ contient les séismes numérotés de 1 à $N$. Elle est légitime dans deux régimes :

- Pour $t \to +\infty$, afin d'estimer uniquement le nombre de séismes *background*, car dans ce cas :

$$\sum_{i \leq N, t_i < t} \kappa_{A,\alpha}(m_i) g_{c,p}(t - t_i) f_{D,\gamma,q}(x - x_i, y - y_i, m_i) \longrightarrow 0 \tag{6}$$

et donc :

$$\lambda(t, x, y, m|H_{t^*}) \approx \nu_\beta(m)\tilde{u}(x, y) \tag{7}$$

Mais comme dans la pratique $\tilde{u}(x, y)$ est estimé indépendamment de l'intensité conditionnelle totale, on peut en déduire le nombre de séismes *background* sur n'importe qu'elle période sans avoir à faire d'approximation.

- Pour $t$ proche de $t^*$ à droite, si nous choisissons $t^*$ et $S$ au voisinage d'un séisme de $H_{t^*}$. Nous pouvons alors espérer capter le nombre moyen de répliques attendues en ne tenant pas compte du *background sismicity rate*.

En résumé, sachant que $\int_{m_0}^{+\infty} \nu_\beta = 1$, on a en pratique :

- Pour tout $t$, le nombre de séismes *background* $n_b$ est estimé par :

$$n_b = (t_f - t_i) \int_S \tilde{u}(x, y)\, dx\, dy \tag{8}$$

- Pour $t_i = t^*$, où $S$ et $t^*$ sont au voisinage d'un séisme de $H_{t^*}$, le nombre de répliques $n_a$ est estimé par :

$$n_a = \int_{t_i}^{t_f} \int_S \sum_{i \leq N, t_i < t} \kappa_{A,\alpha}(m_i) g_{c,p}(t - t_i) f_{D,\gamma,q}(x - x_i, y - y_i, m_i)\, dt\, dx\, dy \tag{9}$$

### 4.2 Méthode de test

Rappelons-nous que nous sommes partis d'un historique de séismes jusqu'à une date $T$ noté $H_T$ et que nous avons entraîné notre modèle ETAS sur un sous-catalogue $H_{t^*}$. Naturellement, l'idée serait désormais d'évaluer les prédictions que nous avons faites sur l'intervalle $[t_i, t_f] \subset [t^*, T]$ en les comparant à ce qui s'est effectivement passé entre $t^*$ et $T$. Cependant, une nouvelle difficulté se pose : nous ne savons pas a priori si un séisme est un évènement *background* ou *aftershock*. Autrement dit, nous n'avons pas accès à la *ground truth* qui permettrait de classifier nos séismes. Par conséquent, nous n'avons accès qu'au nombre total $n$ de séismes dans une zone et sur une période donnée. Or nos estimations portent sur $n_a$ et $n_b$, qui sont estimés dans des régimes différents, si bien que jamais nous ne pourrons faire l'opération : $n_a + n_b = n$.

Une façon de procéder est d'entraîner une seconde fois un modèle ETAS sur l'ensemble du catalogue $H_T$ cette fois. Cela permet d'obtenir les probabilités pour chaque séisme d'être une réplique. Un seuil à 50% de confiance permet de classifier les séismes de tout le catalogue $H_T$ et donc de pouvoir effectuer des comparaisons. Une lacune de cette méthode est que les probabilités de *clustering*, issues d'une estimation par le modèle ETAS, ne peuvent elles pas être testées. En l'absence de meilleure solution, nous utiliserons malgré tout cette méthode.

# 5 Mise en pratique

Nous avons montré que `ETAS` ne permettait pas de prédire le nombre total de séismes, mais seulement le nombre de séismes de chaque catégorie dans certains régimes. Une approche plus simple qui donnerait directement le nombre total de séismes serait préférable, c'est pourquoi nous allons dans un premier temps appliquer des méthodes plus élémentaires de prédictions qui ne discriminent pas les séismes en *background* et *aftershock*.

## 5.1 Des approches simples pour prédire le risque global

### 5.1.1 Régression à noyau gaussien à grande échelle temporelle

Bien que nous ayons considéré qu'une modélisation en séries temporelles serait inadaptée, plusieurs personnes prétendent obtenir des résultats. Des commandes `R` permettent d'estimer notre série à l'aide de noyaux gaussiens.

Nous cherchons à prédire le nombre total $n$ de séismes par année, sans identifier si un évènement est une réplique ou non. Pour cela nous réalisons une régression à noyau gaussien classique à l'aide de la commande `ksmooth`. Nous allons systématiquement retirer les 9 premières années des données qui nous paraissent inadaptées au reste du dataset. Pour chaque zone, nous allons tester la méthode pour des paramètres $k$ variant de 1 à 100 et conserver celui qui minimise la RMSE dont nous rappelons la formule ci-après.

$$RMSE(u, v) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(u_i - v_i)^2} \tag{10}$$

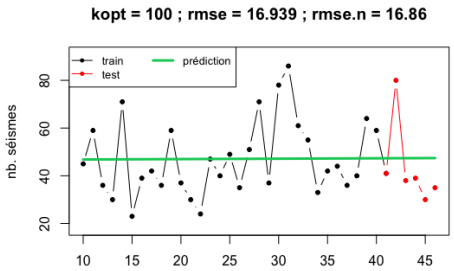
Dans le tableau 2 nous répertorions les informations associées à chaque prédiction. Puis dans la figure 10, nous représentons en noir les données utilisées comme base d'entraînement, en rouge celles utilisées comme échantillon *test* et en vert la prédiction proposée par `ksmooth` pour un paramètre $k$ optimal minimisant la RMSE. Pour chaque zone, nous comparons la prédiction à un prédicteur naïf, à savoir le nombre moyen de séismes sur la période *train*.

L'approche par `ksmooth` ne se distingue pas de notre référence naïve. Cela corrobore notre analyse descriptive en série temporelle : les séismes étant un phénomène stationnaire, l'information se réduit à la moyenne lorsque nous considérons des échelles temporelles relativement grandes.

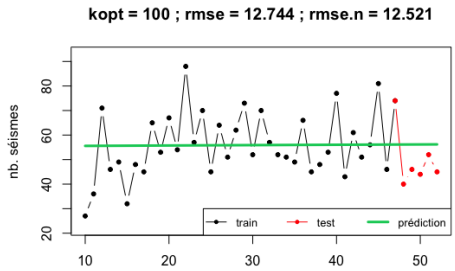
Cela ne signifie pas pour autant que l'on ne pourrait pas capter de la structure à plus petite échelle. Les répliques par exemple arrivent généralement dans les jours qui suivent un séismes. A priori `ksmooth` devrait se distinguer du prédicteur naïf sur des périodes plus courtes.

| Zone | J | P | AS | NA | SA |
|---|---|---|---|---|---|
| années *train* | 1974-2004 | 1974-2010 | 1974-2010 | 1974-2010 | 1974-2010 |
| années *test* | 2005-2010 | 2011-2016 | 2011-2016 | 2011-2016 | 2011-2016 |
| $k$ optimal | 100 | 100 | 5 | 100 | 7 |
| RMSE | 16.939 | 12.744 | **18.543** | 28.420 | **7.609** |
| RMSE naïf | **16.860** | **12.521** | 23.122 | **27.190** | 8.690 |

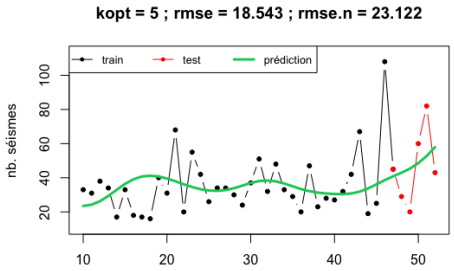
TABLE 2 – Informations et résultats liés à la méthode `ksmooth`



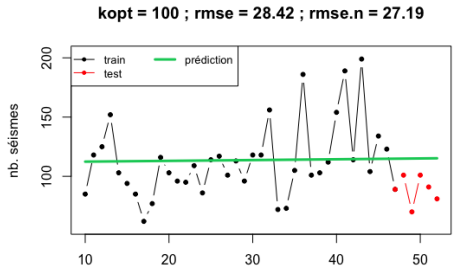

(a) Japon (**J**)



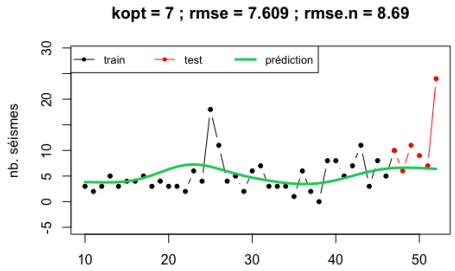

(b) Pacifique (**P**)




(c) Amérique du Sud (**AS**)




(d) Nord de l'Australie (**NA**)




(e) Sud de l'Australie (**SA**)

FIGURE 10 – Prédictions réalisées à l'aide `ksmooth` sur les différentes zones

### 5.1.2 Régression à noyau gaussien à petite échelle temporelle

Nous réalisons les mêmes démarches que précédemment sur une période de *train* de 10 ans et de *test* de 2 ans. Cette fois-ci, les résultats sont plus encourageants. Ils sont résumés dans le tableau 3. Dans le tableau 4, nous comparons les RMSE sur une courte période à celles sur une longue période.

| Zone | J | P | AS | NA | SA |
|---|---|---|---|---|---|
| années *train* | 1999-2008 | 2005-2014 | 2005-2014 | 2005-2014 | 2005-2014 |
| années *test* | 2009-2010 | 2015-2016 | 2015-2016 | 2015-2016 | 2015-2016 |
| $k$ optimal | 1 | 6 | 1 | 3 | 2 |
| RMSE | **6.891** | **3.343** | **21.258** | **6.424** | **10.411** |
| RMSE naïf | 15.997 | 6.689 | 26.402 | 36.742 | 11.673 |

TABLE 3 – Informations et résultats de la méthode `ksmooth` sur des périodes plus courtes

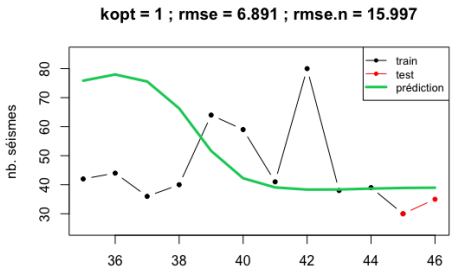
| Zone | J | P | AS | NA | SA |
|---|---|---|---|---|---|
| RMSE courte période | **6.891** | **3.343** | 21.258 | **6.424** | 10.411 |
| RMSE longue période | 16.939 | 12.744 | **18.543** | 28.420 | **7.609** |

TABLE 4 – Comparaisons des RMSE obtenus sur des périodes longues et courtes

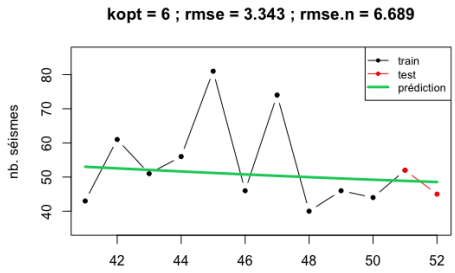
Le prédicteur `ksmooth` réalise des scores bien meilleurs que précédemment : sa RMSE est systématiquement plus petite que celle du prédicteur naïf. De plus, les zones pour lesquelles la première prédiction `ksmooth` ne fonctionnait pas (le $k$ optimal était maximal et la RMSE n'était pas meilleure que celle du prédicteur naïf) ont désormais une meilleure prédiction. Cela corrobore notre intuition sur l'importance de l'échelle considérée : à grande échelle la sismicité ne présente pas de tendance particulière ; à courte échelle nous arrivons à capter des variations.

Cependant, les zones pour lesquelles la prédiction avec `ksmooth` était meilleure que celle naïve, admettent maintenant une RMSE plus élevée que précédemment. Pour les zones **AS** et **SA** le prédicteur semble mal s'adapter à des variations brusques. Autrement dit, il ne capte qu'une tendance localisée dans le temps et s'avèrera donc forcément inutile pour anticiper une période soudainement violente. Pour **J**, **P** et **NA**, comme sur les données considérées le nombre de séismes oscille peu, la prédiction fonctionne raisonnablement bien.

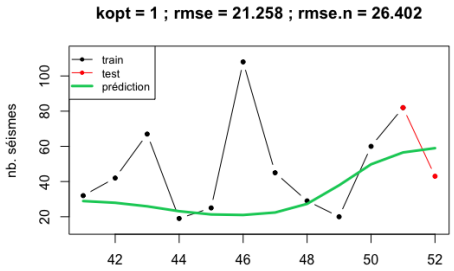
En conclusion, une approche simple comme `ksmooth` ne capte que des tendances locales ou générales selon l'échelle considérée, mais ne permet pas de détecter une structure spécifique qui permettrait d'anticiper des périodes à risques. C'est pourquoi le modèle ETAS semble plus prometteur : en classifiant les séismes et en considérant les répliques comme un phénomènes épidémiques, nous espérons gagner en information.



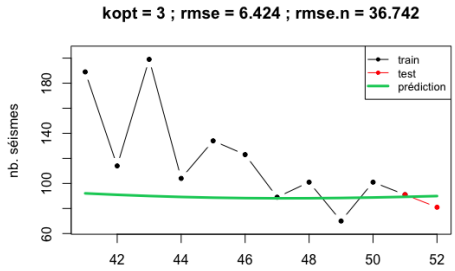

(a) Japon (**J**)



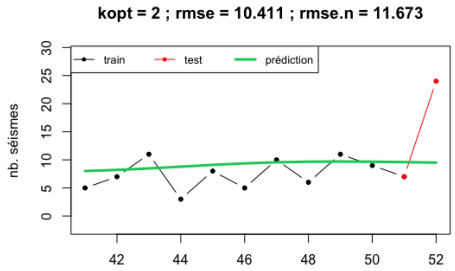

(b) Pacifique (**P**)




(c) Amérique du Sud (**AS**)




(d) Nord de l'Australie (**NA**)




(e) Sud de l'Australie (**SA**)

FIGURE 11 – Prédictions réalisées à l'aide `ksmooth` sur les différentes zones pour des périodes plus courtes

### 5.1.3 Mélange de lois de Poisson

À présent, nous allons nous intéresser à la répartition du nombre de séismes par an en cherchant à ajuster un mélange de lois de Poisson à l'histogramme des nos données pour chaque zone. Il s'agit d'un premier pas vers les processus ponctuels que nous poursuivrons avec ETAS. Nous avons utilisé la fonction `flexmix`.

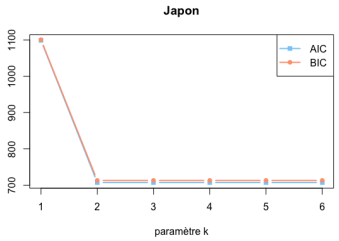
Cette fois-ci, il faut déterminer le nombre $k$ de lois de Poisson à utiliser dans le mélange. Pour cela, nous procédons par minimisation des critères AIC et BIC. Nous avons représenté leur évolution en fonction de $k$ pour chaque zone dans la colonne de gauche de la figure 12. Les deux critères ne sont pas systématiquement minimaux pour le même paramètre $k$, nous en choisissons donc un arbitrairement (à partir des courbes) qui semble minimiser au mieux les deux. Dans la colonne de droite de la figure 12, nous avons représenté les histogrammes de la répartition du nombre de séismes par année superposés par la somme des lois de Poisson. Dans le tableau 5 nous fournissons les informations associées à chaque cas.

| Zone | J | P | AS | NA | SA |
|---|---|---|---|---|---|
| $k$ optimal | 2 | 3 | 3 | 3 | 2 |
| AIC | 707.31 | 444.06 | 439.80 | 532.93 | 278.75 |
| BIC | 713.17 | 453.82 | 449.55 | 542.68 | 284.61 |

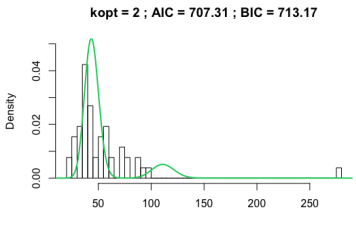
TABLE 5 – Informations et résultats liés à la méthode `flexmix`

Pour chaque zone, les critères décroissent assez rapidement puis se stabilisent. Nous prenons alors les premiers $k$ pour lesquelles l'AIC et le BIC semblent atteindre cette valeur stable. À vu d'œil, l'ajustement des courbes semblent correct et bien correspondre aux histogrammes.

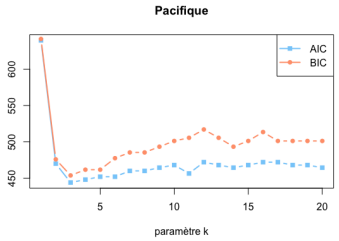
Afin d'exploiter cette méthode, il faudrait ajuster les modèles à des données d'entraînement et évaluer la probabilité d'être dans un certain mode de la distribution. Sans cela, le mélange de Poisson ne permettrait de récupérer qu'une moyenne. Nous laissons cette approche comme une extension possible du projet. Sachant que les processus de Poisson semblent en effet bien modéliser les séismes, nous passons à une version plus avancées de ces derniers : le modèle ETAS.



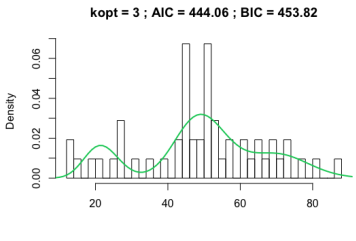

(a) Japon (J)



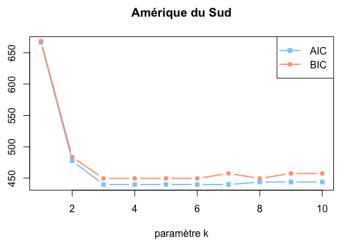

(b) Japon (J)



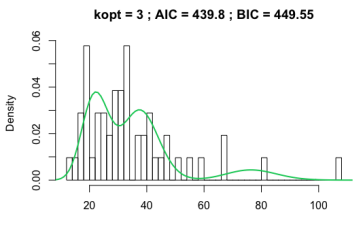

(c) Pacifique (P)



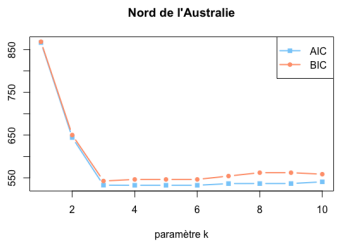

(d) Pacifique (P)



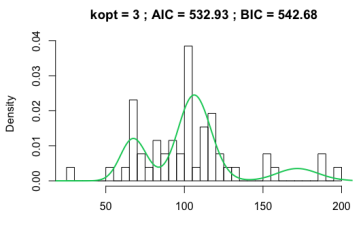

(e) Amérique du Sud (AS)



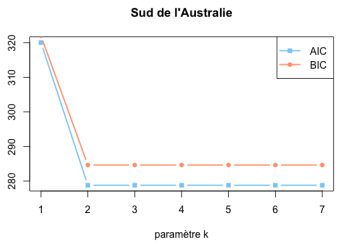

(f) Amérique du Sud (AS)



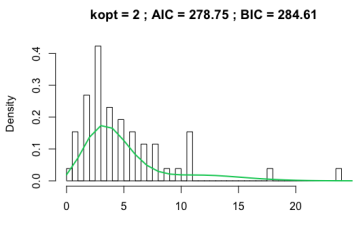

(g) Nord de l'Australie (NA)




(h) Nord de l'Australie (NA)




(i) Sud de l'Australie (SA)




(j) Sud de l'Australie (SA)

FIGURE 12 – Prédictions réalisées à l'aide `ksmooth` sur les différentes zones

## 5.2 Le risque *background* selon le modèle ETAS

Comme nous l'avons vu précédemment, le modèle ETAS se prête mal à la prédiction des répliques en dehors de régimes très particuliers, alors que l'on est toujours en mesure d'accéder à l'intensité des *background*. Notre objectif sera avant tout de prédire le nombre de séismes *background* $n_b$, mais nous garderons à titre indicatif la prédiction de réplique $n_a$. Nous prendrons toujours $[t_i, t_f] = [t^*, T]$ comme intervalle de prédiction. Nous définirons alors $t^*$ selon le ratio test/entraînement souhaité. De façon concrète, nous nous fixons une fraction $\delta$ de la durée totale du catalogue qui sera dédiée à l'apprentissage. La date limite d'entraînement est alors définie par : $t^* = t_0 + \delta \times (T - t_0)$. Nous avons alors l'entraînement sur $[t_0, t^*]$ et la prédiction sur $[t^*, T]$. Chacune des dates de format `AAAA-MM-JJ HH:MM:SS` aura une représentation numérique que l'on arrondira au jour prés pour faciliter la lecture des données.

### 5.2.1 Le Nord de l'Australie

Nous travaillons sur un jeu de 800 séismes tel que $t_0 = 0$ représente la date `1965-02-16 12:24:10`, $t^* = t_0 + 0.80 \times (T - t_0)$ la date `1978-10-18 23:30:20`, et $T$ la date `1982-03-21 10:22:33`. Le tableau ci-dessous résume la division en temps du dataset :

| Split pour $N = 800$ séismes | Calendrier | Numérique |
|---|---|---|
| Date initiale $t_0$ | `1965-02-16 12:24:10` | 0 |
| Date limite d'entraînement $t^*$ ($\delta = 0.80$) | `1978-10-18 23:30:20` | 4994 |
| Date finale $T$ | `1982-03-21 10:22:33` | 6242 |

Table 6 – Chronologie choisie pour le Nord de l'Australie

Nous entraînons le modèle ETAS sur $[t_0, t^*]$ et nous faisons des prédictions sur $[t_i, t_f] = [t^*, T]$ de durée 1248 jours en environ. Nous obtenons les résultats suivants :

| Résultats | $n_a$ | $n_b$ |
|---|---|---|
| Valeur test | 17 | 85 |
| Valeur prédite | $5 \times 10^{-4}$ | 13.3 |
| Erreur absolue | 17 | 71.7 |
| Erreur relative | 100.0 % | 84.4 % |

Table 7 – Prédictions sur les données d'entraînement du Nord de l'Australie

Nos résultats souffrent d'une **grande imprécision** qui était prévisible au moins pour la prédiction des répliques. Présentons les résultats obtenus sur les autres zones géographiques avant d'analyser les résultats dans leur globalité.

### 5.2.2 Le Japon

Nous procédons de même sur un jeu de données contenant 400 séismes au Japon.

| Split pour $N = 400$ séismes | Calendrier | Numérique |
|---|---|---|
| Date initiale $t_0$ | `2000-02-13 02:57:09` | 0 |
| Date limite d'entraînement $t^*$ ($\delta = 0.80$) | `2006-05-22 13:08:03` | 2298 |
| Date finale $T$ | `2007-12-25 14:04:35` | 2872 |

Table 8 – Chronologie choisie pour le Japon

Nous entraînons le modèle ETAS sur $[t_0, t^*]$ et nous faisons des prédictions sur $[t_i, t_f] = [t^*, T]$ de durée 574 jours environ. Nous obtenons les résultats suivants :

| Résultats | $n_a$ | $n_b$ |
|---|---|---|
| Valeur test | 2 | 15 |
| Valeur prédite | 0.8 | 6.8 |
| Erreur absolue | 1.2 | 8.2 |
| Erreur relative | 60.0 % | 54.7 % |

TABLE 9 – Prédictions sur les données d'entraînement du Japon

Notons que les prédictions sont **remarquablement meilleures** que pour le Nord de l'Australie.

### 5.2.3 Le Pacifique

À nouveau, nous réalisons la même démarche pour la zone du Pacifique.

| Split pour $N = 800$ séismes | Calendrier | Numérique |
|---|---|---|
| Date initiale $t_0$ | `1998-07-07 14:17:51` | 0 |
| Date limite d'entraînement $t^*$ ($\delta = 0.80$) | `2009-10-04 17:09:53` | 4113 |
| Date finale $T$ | `2012-08-03 07:22:31` | 5141 |

TABLE 10 – Chronologie choisie pour le Pacifique

Nous entraînons le modèle ETAS sur $[t_0, t^*]$ et nous faisons des prédictions sur $[t_i, t_f] = [t^*, T]$ de durée 1028 jours environ. Nous obtenons :

| Résultats | $n_a$ | $n_b$ |
|---|---|---|
| Valeur test | 26 | 139 |
| Valeur prédite | 0.01 | 4.3 |
| Erreur absolue | 26 | 134.7 |
| Erreur relative | 100.0 % | 96.9 % |

TABLE 11 – Prédictions sur les données d'entraînement du Pacifique

### 5.2.4 L'Amérique du Sud

Nous réalisons la même démarche pour l'Amérique du Sud.

| Split pour $N = 200$ séismes | Calendrier | Numérique |
|---|---|---|
| Date initiale $t_0$ | `1965-04-11 00:11:11` | 0 |
| Date limite d'entraînement $t^*$ ($\delta = 0.70$) | `1996-03-23 10:46:46` | 11331 |
| Date finale $T$ | `2009-08-05 08:31:40` | 16187 |

TABLE 12 – Chronologie choisie pour l'Amérique du Sud

Nous entraînons le modèle ETAS sur $[t_0, t^*]$ et nous faisons des prédictions sur $[t_i, t_f] = [t^*, T]$ de durée 4856 jours environ. Nous obtenons :

| Résultats | $n_a$ | $n_b$ |
|---|---|---|
| Valeur test | 23 | 41 |
| Valeur prédite | 0.001 | 0.6 |
| Erreur absolue | 23 | 40.4 |
| Erreur relative | 100.0 % | 98.6 % |

TABLE 13 – Prédictions sur les données d'entraînement de l'Amérique du Sud

### 5.2.5 Analyse des résultats

Finalement, nous échouons à prédire le risque sismique à l'aide du modèle ETAS. Bien que nous ne nous attendions pas à des résultats extrêmement précis, nos estimations se trompent souvent d'une puissance de dix. La zone région du monde pour laquelle nous retrouvons au moins le bon ordre de grandeur est le Japon.

Afin de vérifier la pertinence du modèle ETAS, nous allons contrôler la stabilité de ses paramètres au cours du temps.

## 5.3 Stabilité et réplicabilité des paramètres ETAS

Nous envisageons deux façons d'interpréter l'imprécision de nos résultats. La première serait un sur-ajustement aux données d'entraînement ; la seconde remettrait en cause la stationnarité du modèle. Le modèle ETAS repose sur l'hypothèse que ses paramètres ne varient pas en fonction du temps. Or, la communauté scientifique s'intéresse aussi à des modèles ETAS non-stationnaires. Cela irait à l'encontre de notre conclusion sur la stationnarité des séismes, mais encore une fois tout dépend de l'échelle temporelle considérée.

Pour évaluer le sur-ajustement du modèle, nous allons calculer l'erreur d'entraînement. Nous comparerons aussi les valeurs des paramètres estimés du modèle sur la période d'entraînement et sur la période de test, afin d'observer une éventuelle variation.

Dans notre configuration, les unités des paramètres sont les suivantes :

| Paramètres | $\mu$ | $A$ | $c$ | $\alpha$ | $p$ | $D$ | $q$ | $\gamma$ |
|---|---|---|---|---|---|---|---|---|
| Unités | Sans | Nombre de séismes | Jour | Magnitude$^{-1}$ | Sans | Degré$^2$ | Sans | Magnitude$^{-1}$ |

TABLE 14 – Unités

Reprenons les mêmes catalogues que précédemment :

### 5.3.1 Estimation des paramètres

| Période | $\mu$ | $A$ | $c$ | $\alpha$ | $p$ | $D$ | $q$ | $\gamma$ |
|---|---|---|---|---|---|---|---|---|
| $[t_0, t^*]$ | 1.08 | 0.05 | 0.05 | 1.87 | 1.22 | 0.05 | 3.26 | 0.97 |
| $[t^*, T]$ | 0.30 | 1.29 | 0.02 | 0.94 | 1.02 | 0.30 | 7.72 | 0.08 |

TABLE 15 – Paramètres ajustés sur les données d'entraînement de l'Amérique du Sud

| Période | $\mu$ | $A$ | $c$ | $\alpha$ |
|---|---|---|---|---|
| $[t_0, t^*]$ | 1.01 | $4.73 \times 10^{-1}$ | $4.00 \times 10^{-3}$ | 1.41 |
| $[t^*, T]$ | $1.28 \times 10^{-1}$ | $8.67 \times 10^{-2}$ | $1.04 \times 10^{4}$ | $8.50 \times 10^{-1}$ |
| Période | $p$ | $D$ | $q$ | $\gamma$ |
| $[t_0, t^*]$ | 1.03 | $5.34 \times 10^{-2}$ | 8.69 | 1.10 |
| $[t^*, T]$ | $5.46 \times 10^{4}$ | $2.63 \times 10^{7}$ | $6.81 \times 10^{9}$ | $2.21 \times 10^{-4}$ |

TABLE 16 – Paramètres ajustés sur les données d'entraînement du Japon

| Période | $\mu$ | $A$ | $c$ | $\alpha$ | $p$ | $D$ | $q$ | $\gamma$ |
|---|---|---|---|---|---|---|---|---|
| $[t_0, t^*]$ | 1.02 | 0.12 | 0.02 | 1.35 | 1.13 | 0.24 | 9.48 | 0.96 |
| $[t^*, T]$ | 0.20 | 0.08 | 0.01 | 1.80 | 1.16 | 0.02 | 2.79 | 1.24 |

TABLE 17 – Paramètres ajustés sur les données d'entraînement du Pacifique

| Période | $\mu$ | $A$ | $c$ | $\alpha$ | $p$ | $D$ | $q$ | $\gamma$ |
|---|---|---|---|---|---|---|---|---|
| $[t_0, t^*]$ | 1.04 | 0.30 | $4.46 \times 10^{-3}$ | 1.50 | 1.05 | 0.01 | 2.80 | 1.73 |
| $[t^*, T]$ | 0.18 | 0.26 | $1.47 \times 10^{-2}$ | 1.67 | 1.16 | 0.07 | 4.15 | 1.20 |

TABLE 18 – Paramètres ajustés sur les données d'entraînement du Nord de l'Australie

### 5.3.2 Erreur d'entraînement

| Résultats | $n_a$ | $n_b$ |
|---|---|---|
| Valeur test | 173 | 306 |
| Valeur prédite | 0.1 | 79.7 |
| Erreur absolue | 172.9 | 226.4 |
| Erreur relative | 100 % | 74.0 % |

TABLE 19 – Prédictions sur les données d'entraînement du Nord de l'Australie

| Résultats | $n_a$ | $n_b$ |
|---|---|---|
| Valeur test | 98 | 537 |
| Valeur prédite | 0.54 | 17 |
| Erreur absolue | 97.5 | 520 |
| Erreur relative | 99.5 % | 96.8 % |

TABLE 20 – Prédictions sur les données d'entraînement du Pacifique

| Résultats | $n_a$ | $n_b$ |
|---|---|---|
| Valeur test | 105 | 198 |
| Valeur prédite | 17.9 | 27.3 |
| Erreur absolue | 87.1 | 170.7 |
| Erreur relative | 83.0 % | 86.2 % |

TABLE 21 – Prédictions sur les données d'entraînement du Japon

| Résultats | $n_a$ | $n_b$ |
|---|---|---|
| Valeur test | 31 | 105 |
| Valeur prédite | 0.07 | 1.4 |
| Erreur absolue | 30.9 | 103.6 |
| Erreur relative | 99.8 % | 98.7 % |

TABLE 22 – Prédictions sur les données d'entraînement de l'Amérique du Sud

### 5.3.3 Analyse des résultats

Une chose est absolument claire : il n'y a aucun problème de sur-ajustement puisque l'erreur d'entraînement est elle aussi extrêmement grande. Les performances pour le Japon sont mêmes moins bonnes sur le training set que le testing set. Par ailleurs, les paramètres estimés ne semblent absolument pas stationnaires. Cependant, comme le contrôle de la stationnarité n'a de sens que si l'on a des estimations correctes, nous ne nous pouvons pas conclure à ce sujet.

Comment comprendre cet échec du modèle ? Nous proposons quelques pistes, qui ne demeurent néanmoins que des spéculations.

D'un point de vue pratique, les prédictions sur des historiques de séismes sont naturellement peu robustes à de nombreuses perturbations, avec en tête l'incomplétion des données. Si l'on veut prédire des séismes de magnitudes significativement grandes, on se retrouve à étudier des évènements relativement rares. Outre le problème du faible nombre de données sur des périodes resserrées, le moindre évènement non reporté dans le dataset modifiera fortement le résultat : chaque séisme est représentatif de l'intensité *background*, mais il peut aussi donner naissance à plusieurs répliques. Lorsque le modèle ETAS s'ajuste à un historique, c'est sous l'hypothèse que 100% des séismes qui ont eu lieu s'y trouvent. C'est pourquoi il est absolument nécessaire de travailler sur un dataset particulièrement fiable. Or, un jeu de données à l'échelle mondiale et sur un siècle nous semble plus sujet à des hétérogénéités qu'un jeu focalisé dans l'espace et dans le temps.

Selon les normes choisies (pour la définition de la localisation d'un séisme par exemple), les datasets du domaine public ne semblent pas équivalents. Afin de voir si la méthode ETAS peut fonctionner dans des cas favorables, nous allons tenter de l'appliquer sur un dataset spécifique au Japon, utilisé à titre d'exemple par les auteurs du package `ETAS`. En considérant la même période et le même découpage, ce second dataset contient un peu moins de données que le précédent (900 contre 800 environ). Nous allons aussi comparer les résultats obtenus sur les données de test à ceux d'un estimateur naïf : la moyenne sur les données d'entraînement.

## 5.4 L'influence du jeu de données

La chronologie pour ce nouveau jeu de donnée est la suivante :

| Split pour $N = 800$ séismes | Calendrier | Numérique |
|---|---|---|
| Date initiale $t_0$ | `2001-10-05 01:56:46` | 0 |
| Date limite d'entraînement $t^*$ ($\delta = 0.80$) | `2005-04-20 10:09:04` | 1294 |
| Date finale $T$ | `2006-03-10 17:55:10` | 1618 |

TABLE 23 – Chronologie choisie pour les données du Japon (dataset `ETAS`)

Nous entraînons le modèle ETAS sur $[t_0, t^*]$ et nous faisons des prédictions sur $[t_0, t^*]$ pour évaluer l'erreur d'entraînement, puis sur $[t_i, t_f] = [t^*, T]$ pour évaluer l'erreur de généralisation.

Sur les données d'entraînement, nous obtenons :

| Résultats | $n_a$ | $n_b$ |
|---|---|---|
| Valeur test | 303 | 344 |
| Valeur prédite | 440.5 | 107 |
| Erreur absolue | 137.5 | 237 |
| Erreur relative | 45.4 % | 68.9 % |

TABLE 24 – Prédictions sur les données d'entraînement du Japon (dataset `ETAS`)

Puis sur les données de *test*, nous obtenons :

| Résultats | $n_a$ | $n_b$ |
|---|---|---|
| Valeur test | 65 | 62 |
| **Valeur naïve** | 79 | 86 |
| Valeur prédite | 22.0 | 27.8 |
| Erreur absolue | 43 | 35.2 |
| Erreur relative | 66.2 % | 56.8 % |

TABLE 25 – Prédictions sur les données test du Japon (dataset `ETAS`)

Les résultats obtenus sur ces données ne sont certes pas précis, mais sont **sensiblement meilleurs** que pour le dataset précédent, particulièrement sur le training set.

Cependant, l'approche naïve est beaucoup plus précise. Il est possible que comme pour `ksmooth`, la procédure `ETAS` se distingue à petite échelle temporelle, là où les répliques jouent un rôle. Nous allons lancer la procédure une nouvelle fois sur une période de *test* de l'ordre de quelques semaines.

### 5.5 ETAS à petite échelle temporelle

Nous gardons le dernier dataset du Japon utilisé, a priori plus fiable. Nous prenons une valeur de $t^*$ définie par $\delta = 0.95$ afin de nous placer dans le régime voulu. La chronologie est la suivante :

| Split pour $N = 400$ séismes | Calendrier | Numérique |
|---|---|---|
| Date initiale $t_0$ | `2001-10-05 01:56:46` | 0 |
| Date limite d'entraînement $t^*$ ($\delta = 0.95$) | `2004-02-17 01:30:12` | 867 |
| Date finale $T$ | `2004-04-04 09:01:22` | 912 |

TABLE 26 – Chronologie choisie pour les données du Japon (dataset `ETAS`)

Nous entraînons le modèle ETAS sur $[t_0, t^*]$ et nous faisons des prédictions sur $[t_i, t_f] = [t^*, T]$ de durée 45 jours environ. Nous obtenons les résultats suivants :

| Résultats | $n_a$ | $n_b$ |
|---|---|---|
| Valeur test | 2 | 10 |
| Valeur naïve | 0.1 | 0.5 |
| Valeur prédite | 0 | 2.9 |
| Erreur absolue | 2 | 7.1 |
| Erreur relative | 99.6 % | 71.2 % |

TABLE 27 – Prédictions sur les données d'entraînement du Japon (dataset `ETAS`)

Le modèle ETAS ne fait toujours pas preuve d'une précision remarquable. Il s'avère malgré tout plus précis que le prédicteur naïf à cette petite échelle temporelle.

Il est sûrement possible d'utiliser ETAS avec une meilleure approche afin d'améliorer les prévisions. Nous pensons cependant que le modèle ne pourra jamais permettre de prédire les séismes de façon fiable : même sur les datasets donnés à titre d'exemple dans le package, l'erreur d'entraînement s'avère très grande.

ETAS pouvait sembler prometteur au sens où il permettait de récupérer de la dépendance temporelle par le biais des répliques. Cependant, l'intensité des répliques, conditionnées par le passé, devient mal définie dans les temps futurs. De plus, même l'intensité *background*, a priori plus élémentaire dans son estimation, est mal évaluée.

De façon générale, que ce soit en utilisant la procédure `ETAS` ou une méthode plus élémentaire, nous échouons sur le long-terme à faire mieux que prendre des moyennes, et nous n'arrivons pas non plus à anticiper les variations à court-terme. Par conséquent, nous pensons que les approches qui étudient en temps réel les signaux précurseurs sont plus légitimes pour deux raisons. En jouant le rôle d'avertisseur de catastrophe en temps réel, elles sont plus utiles que la prévision de la sismicité dans le futur ; en reposant sur de nouvelles covariables en des lieux précis du globe, elles pourraient fournir des modèles plus interprétables et plus précis.

# 6 Conclusion

En commençant ce projet, nous n'ambitionnions pas de réussir en quelques mois là où les scientifiques échouent depuis des années. La connaissance de la tectonique des plaques ainsi que de la répartition des séismes permet de connaître les zones à risques, sans pouvoir prédire précisément la date et le lieu des prochaines secousses. Nous n'avons pas réussi à faire mieux à l'aide du modèle ETAS. La méthode associée est utilisée par la communauté des sismologues pour faire des statistiques descriptives sur des catalogues de séismes, et notamment pour vérifier la stationnarité des phénomènes ou identifier les répliques. Systématiquement, la littérature se restreint à des zones géographiques précises et cherche rarement à utiliser le modèle ETAS pour faire des prédictions. Les résultats que nous obtenons sur notre jeu de données de très grande échelle ne sont dès lors pas très étonnants. Nous aurions pu malgré tout envisager une autre approche. La nature stationnaire et apériodique de la sismicité fait, qu'à grande échelle, nous ne retenons que la moyenne. C'est pourquoi une première possibilité aurait été de se limiter à de très petites échelles spatio-temporelles afin de capter les variations locales liées aux répliques d'un évènement majeur par exemple. Cependant, nous nous serions probablement heurtés au problème de la rareté des données.

De façon générale, nous pensons qu'une méthode qui n'exploite que la fréquence en temps et en espace des séismes ne pourra jamais s'avérer performante. Il faudrait ajouter des covariables qui rendent compte de l'état de la croûte terrestre par exemple. Dès lors, les approches de court-termes qui étudient en des lieux précis de multiples signaux précurseurs en temps réel nous semblent plus prometteuses.

## Références

[1] Abdollah Jalilian. ETAS : An R package for fitting the space-time ETAS model to earthquake data. *Journal of Statistical Software, Code Snippets*, 88(1) :1–39, 2019.

# A Échelle de Richter

| Magnitude | Description | Fréquence moyenne | Effets |
|---|---|---|---|
| $< 1.9$ | Micro | 8 000 par jour | Micro tremblement de terre, non ressenti |
| 2.0 à 2.9 | Très mineur | 1 000 par jour | Généralement non ressenti, mais détecté et enregistré |
| 3.0 à 3.9 | Mineur | 150 par jour | Souvent ressenti, mais cause rarement des dommages |
| 4.0 à 4.9 | Léger | 20 par jour | Secousses d'objets à l'intérieur des maisons, bruits d'entre-chocs, dommages importants peu communs |
| 5.0 à 5.9 | Modéré | 2-3 par jour | Peut causer des dommages majeurs à des édifices mal conçus, cause de légers dommages aux édifices bien construits |
| 6.0 à 6.9 | Fort | 120 par an | Peut être destructeur dans des zones allant jusqu'à 180 km à la ronde |
| 7.0 à 7.9 | Majeur | 18 par an | Peut provoquer des dommages modérés à sévères dans des zones plus vastes |
| 8.0 à 8.9 | Important | 1 par an | Peut causer des dommages sérieux dans des zones à des centaines de kilomètres à la ronde |
| $> 9$ | Dévastateur | 1 à 5 par siècle | Dévaste des zones de plusieurs milliers de kilomètres à la ronde |