

REVUE DE STATISTIQUE APPLIQUÉE

S. BOTTARD

Application de la méthode du Bootstrap pour l'estimation des valeurs extrêmes dans les distributions de l'intensité des séismes

Revue de statistique appliquée, tome 44, n° 4 (1996), p. 5-17

http://www.numdam.org/item?id=RSA_1996__44_4_5_0

© Société française de statistique, 1996, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

APPLICATION DE LA MÉTHODE DU BOOTSTRAP POUR L'ESTIMATION DES VALEURS EXTRÊMES DANS LES DISTRIBUTIONS DE L'INTENSITÉ DES SÉISMES

S. Bottard*

*Institut de Protection et de Sécurité Nucléaire, Département de Protection de la santé
de l'Homme et de Dosimétrie, Service d'Evaluation et de Gestion des Risques,
IPSN, BP6, F-92265 Fontenay-aux-Roses Cedex*

RÉSUMÉ

Dans la plupart des méthodes d'analyse probabiliste du risque sismique, le modèle qui décrit la sismicité d'une source suppose que l'intensité des séismes est une variable aléatoire de distribution exponentielle. Cependant, la qualité de l'ajustement d'une distribution exponentielle repose sur l'exhaustivité des observations, qui n'est pas assurée sur la totalité de l'historique des données de sismicité concernant les petits événements et les événements majeurs. Les distributions issues de la théorie des valeurs extrêmes de Gumbel ont l'avantage de reposer exclusivement sur les valeurs maximales prises par l'intensité des séismes, plus fiable sur la majeure partie de l'historique que l'information concernant tout le spectre d'intensités et notamment les plus faibles. De plus, la théorie des valeurs extrêmes permet d'obtenir directement des estimations fiables des grands événements, pour lesquels le petit nombre de réalisations induit de fortes incertitudes dans l'analyse probabiliste de l'aléa sismique. Une méthode originale d'inférence statistique semiparamétrique, fondée sur la théorie des valeurs extrêmes et une procédure Bootstrap est proposée, afin de déterminer les intensités maximales susceptibles de se produire sur le territoire français.

Mots-clés : Théorie des valeurs extrêmes, statistique non paramétrique du Bootstrap, analyse probabiliste de l'aléa sismique, France.

ABSTRACT

In most of the probabilistic analysis of seismic hazard the model describing the seismicity of the source assumes that the earthquake intensity behaves like a random variable with exponential distribution. But the fit depends on the exhaustivity of the observations which is not achieved on the whole historical seismicity data, especially for small and major events. The distributions issued from Gumbel's theory for extreme values have the advantage to rely only on the maximum values of the earthquake intensity. These values are more reliable on the greater part of the historical data than the information concerning the whole spectrum of

* Adresse actuelle : Laboratoire de Détection et de Géophysique, Commissariat à l'Energie Atomique, B.P. 12, 91680 Bruyères-le-Châtel.

intensities and especially the small ones. The theory of extreme values provides straight accurate estimates of great events for which the low number of observations entails high uncertainties in the probabilistic analysis of seismic hazard. A convenient statistical method is proposed to determine maximum intensities likely to occur in the French territory. This semiparametric method associates the theory of extreme values and a Bootstrap procedure.

Keywords : *Theory of extreme values, Bootstrap nonparametric statistic, probabilistic analysis of seismic hazard, France.*

1. Introduction

L'évaluation de l'aléa sismique pour les installations à risque pose aux statisticiens un certain nombre de problèmes spécifiques nécessitant le recours à des types de résolution particuliers. L'une des difficultés majeures dans ce type d'études provient du fait que les événements considérés sont souvent «rares» et souvent à conséquences «graves». Ces événements ne s'étant pratiquement jamais produits, il est impossible de les étudier en appliquant directement les méthodes statistiques classiques. Dans le domaine de l'évaluation de l'aléa sismique, le problème consiste à estimer les probabilités d'occurrence associées aux séismes majeurs auxquels les structures doivent résister. Du fait du caractère modéré de la sismicité en France, il est difficile de déterminer l'intensité (ou la magnitude) maximale susceptible de se produire sur un site. Le séisme maximal peut, ne jamais avoir eu lieu par le passé, ou bien ne pas avoir eu lieu pendant la période historique d'observation. En France, les données de sismicité historique concernent le dernier millénaire; le problème consiste donc à estimer de façon fiable l'aléa sismique lié aux événements dont les périodes de retour sont supérieures à 1 000 ans (c'est-à-dire de probabilité inférieure à 10^{-3}).

Dans l'analyse de l'aléa sismique, l'intensité est généralement supposée suivre une loi exponentielle. Dans la pratique, les ingénieurs introduisent une borne inférieure et une borne supérieure dans la distribution (Cornell, 1968), ceci pour deux raisons distinctes : 1) les événements de faible intensité ne sont pas représentés de façon exhaustive dans les données de sismicité historique sur lesquelles sont réalisés les ajustements; 2) dans des contextes sismotectoniques comme celui de la France, caractérisés par une sismicité intraplaque diffuse et relativement modérée, il est généralement admis qu'il existe un maximum au potentiel sismique. Cependant, la discontinuité abrupte qui caractérise une distribution tronquée ne reflète pas forcément la réalité, et il demeure nécessaire d'estimer l'intensité maximale, si l'on ne veut pas considérer par défaut une borne supérieure à l'échelle des intensités, comme le degré XII de l'échelle MSK¹. La précision des estimations obtenues reste soumise à la pertinence de la borne supérieure choisie.

¹ L'intensité est un indice qualitatif de la sévérité des secousses subies en un point donné lors d'un séisme. L'échelle des intensités MSK, mise au point par V. Medvedev, W. Sponheuer et V. Karnik en 1964, est une échelle discrète qui s'étend du degré I au degré XII. Elle classe les séismes de façon hiérarchique selon les effets ressentis par les personnes, les dommages causés aux bâtiments et à l'environnement naturel. La magnitude est une mesure quantitative de la taille d'un séisme, liée à l'amplitude maximale du mouvement du sol par une relation log-linéaire. La magnitude se mesure sur l'échelle établie par C.F. Richter en 1935, qui compte neuf degrés en pratique, mais qui n'est a priori pas bornée. Sur cette échelle, un saut de un degré de magnitude équivaut à une énergie libérée à la source 30 fois supérieure.

Bien que des méthodes de recherche de limites physiques fondées sur le moment sismique ou le cumul des déformations estimées à partir des données historiques aient été proposées, le choix d'une intensité maximale absolue reste assez arbitraire (généralement on majore l'intensité maximale observée), surtout en France où le caractère modéré et diffus de la sismicité ne permet pas d'estimer avec certitude un tel paramètre. C'est pourquoi il est préférable de considérer l'intensité maximale, non plus comme un paramètre de la distribution des intensités, mais directement comme une variable aléatoire. Ce choix, qui permet de tenir compte de l'incertitude due au manque de données sur les événements majeurs, est également le plus approprié à l'analyse de l'incertitude liée à ce paramètre. Les données concernant les grandes intensités sont généralement insuffisantes pour justifier avec certitude l'hypothèse d'une loi paramétrique particulière. Plutôt que de choisir parmi les distributions de probabilité classiques celle qui ajuste le mieux les valeurs maximales observées, on s'intéresse aux lois issues de la théorie des valeurs extrêmes, qui permettent d'obtenir directement des prévisions des intensités maximales.

2. La théorie des valeurs extrêmes

2.1. Les distributions des valeurs extrêmes

La théorie des valeurs extrêmes a pour but d'étudier la distribution du maximum d'une suite de variables aléatoires même si, et spécialement si, la distribution du phénomène n'est pas connue. Formellement, si n est le nombre de séismes qui se sont produits pendant une période donnée et $X_1 \dots X_n$ les variables aléatoires représentant leurs intensités, la variable aléatoire dont on cherche à déterminer la distribution est l'intensité maximale $X = \max(X_1 \dots X_n)$. Les intensités étant supposées être des variables aléatoires indépendantes et identiquement distribuées, de fonction de répartition F_M , la distribution F de l'intensité maximale est définie par :

$$F(x) = P(X \leq x) = \prod_{i=1}^n P(X_i \leq x) = [F_M(x)]^n \quad \forall x \quad (1)$$

L'étude de la distribution asymptotique de X quand n tend vers l'infini (Gumbel, 1967) conduit à trois types de lois limites :

2.1.1. Distribution de type I

$$F(x) = \exp(-e^{-\alpha(x-u)}) \quad \forall x, \quad (\alpha, u) > 0 \quad (2)$$

où la valeur caractéristique u est égale à $m - \gamma/\alpha$ (m désignant la moyenne de la distribution et γ la constante d'Euler) et où le paramètre d'échelle α est égal à $\pi/(\sigma\sqrt{6})$ (σ étant l'écart-type de la distribution). La distribution de type I est généralement appelée loi de Gumbel ou double exponentielle.

En effectuant le changement de variable $Y = \alpha(X - u)$, où Y est appelée variable de Gumbel standard, la distribution de type I a pour expression :

$$F(y) = \exp(-e^{-y}) \quad \forall y \quad (2bis)$$

Cette distribution a pour espérance mathématique γ et pour variance $\pi^2/6$.

2.1.2. Distribution de type II

$$F(x) = \exp\left(-\left(\frac{\nu}{x}\right)^k\right) \quad \forall x > 0, (k, \nu) > 0 \quad (3)$$

Le domaine de définition de la distribution de type II est borné à gauche. Cette loi est parfois associée au nom de Fréchet et plus rarement à ceux de Fischer et Tipett. Les moments de la distribution de type II n'ont pas d'expression analytique simple, mais la distribution peut se ramener à la loi de Gumbel standard en effectuant le changement de variable $Y = -\ln(\nu/X)^k$.

2.1.3. Distribution de type III

$$F(x) = \exp\left(-\left(\frac{x}{\nu}\right)^k\right) \quad \forall x \leq 0, k > 0, \nu < 0 \quad (4)$$

Le domaine de définition de la distribution de type III admet 0 pour borne supérieure. Cette distribution peut se ramener à la loi de Gumbel standard en effectuant le changement de variable $Y = -\ln(X/\nu)^k$. Le mode de la distribution de type III, qui a pour valeur $\nu((k-1)/k)^{1/k}$, $k > 1$, est proche du paramètre caractéristique ν dès que $k > 3$.

Les conditions théoriques d'application des lois des valeurs extrêmes sont très rarement satisfaites lors des applications pratiques. Malgré ces conditions d'application très restrictives (l'indépendance des événements notamment), les distributions des valeurs extrêmes fournissent souvent de bons ajustements aux données expérimentales. Un certain nombre d'auteurs ont utilisé une distribution de type I ou II des valeurs extrêmes pour décrire le processus d'occurrence des séismes de grande intensité dans une région à forte sismicité comme la Californie (Kijko, 1987). Aucune étude des lois des valeurs extrêmes pour l'évaluation de l'aléa sismique n'a été réalisée en France jusqu'à présent.

2.2. Modélisation de la sismicité régionale française : estimation paramétrique de la distribution de type III

Les données de sismicité régionales sont extraites du fichier SIRENE qui recense plus de 5 000 séismes caractérisés par leur épicentre et leur intensité

macrosismique. L'ajustement des lois des valeurs extrêmes a été effectué sur la série des intensités historiques maximales annuelles observées dans les régions du zonage sismotectonique français élaboré par Grellet *et al.* (Grellet, 1993) (figure 1). Le degré d'intensité minimal considéré est de V (MSK), pour des raisons d'exhaustivité. Les ajustements ont été réalisés sur l'historique 1343-1993. La période d'une année d'observation semble être la période la plus appropriée pour l'analyse de la série des intensités maximales selon des études réalisées par Schenkova et Karnik sur la période d'observation (Schenkova, 1970). Une période plus petite introduit un grand nombre de données manquantes.

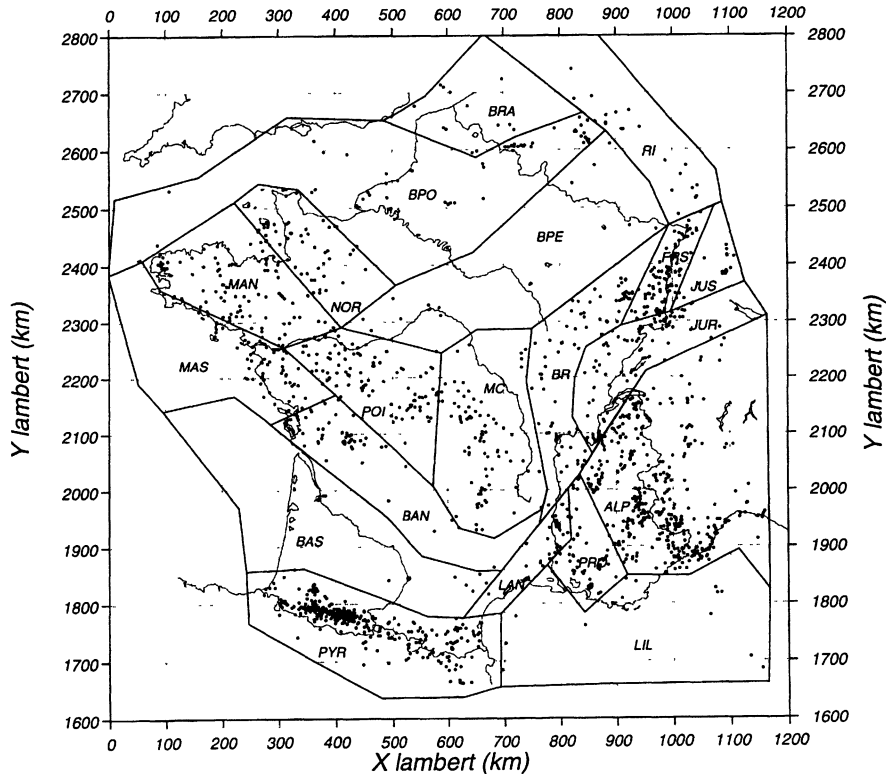


FIGURE 1
Zonage sismotectonique français

On peut d'abord s'intéresser à la distribution de type I qui extrapole les probabilités associées aux intensités supérieures à l'intensité maximale observée. Mais avec un tel type de distribution, il reste le problème de la définition d'une intensité maximale. Certains auteurs définissent l'intensité maximale potentielle comme la valeur $x_{\max}$ dont la probabilité de dépassement soit égale à 1%, c'est-à-dire telle que $F(x_{\max}) = 0,99$. L'autre alternative est l'ajustement de la distribution de type III des valeurs extrêmes qui estime une borne supérieure à la distribution des intensités à partir de la distribution observée.

La loi limite de type III a été ajustée par la méthode d'estimation non linéaire de Gauss-Newton. Cette méthode consiste à linéariser la fonction de répartition en utilisant son développement de Taylor au premier ordre, puis à minimiser la somme des carrés des erreurs du problème linéarisé (c'est-à-dire régresser les résidus sur l'espace vectoriel engendré par les dérivées partielles du premier ordre du modèle linéarisé). On utilisera la forme suivante de la distribution :

$$P(I \leq i) = \exp \left(- \left(\frac{w - i}{w - u} \right)^k \right) \quad \forall i \leq w, \quad (k, u, w) > 0 \quad (5)$$

où w désigne la borne supérieure de la distribution, k le paramètre d'échelle et u la valeur caractéristique. Le domaine de définition de la distribution, associe une contrainte à chaque paramètre :

- une contrainte de positivité en ce qui concerne le paramètre k ;
- une contrainte de positivité et une limite supérieure imposée par son domaine de définition pour le paramètre u : $0 < u < w$;
- une contrainte double pour le paramètre w : $I_{\max} < w \leq 12$; la limite inférieure étant imposée par définition, car w doit être strictement supérieur à l'intensité maximale observée $I_{\max} = \max(x_1 \dots x_n)$. L'introduction de la limite supérieure, qui est celle de l'échelle MSK, provient du fait que l'estimation de la borne supérieure de la distribution, w , doit avoir un sens du point de vue de la physique, c'est-à-dire doit être interprétable en termes d'intensité maximale.

Il s'agit donc d'effectuer la régression du problème linéarisé sous contrainte. Soient β^* l'estimateur du problème de Gauss-Newton linéarisé non contraint où $\beta = (k, u, w)$, et C_β le domaine défini par les contraintes sur β . Deux cas se présentent : 1) si β^* appartient à C_β , alors la solution de la régression sous contrainte est β^* lui-même; 2) sinon, la solution de la régression sous contrainte est la projection de β^* sur la frontière de C_β . Autrement dit, si β^* satisfait la contrainte, il est solution du problème contraint, sinon la solution optimale sature une des contraintes.

En pratique, les contraintes de positivité ne posent pas de réel problème, les données permettant aux estimateurs de les satisfaire sans qu'elles soient intégrées explicitement au problème d'optimisation (cas 1). Il en est de même pour la contrainte sur u qui s'écrit finalement $u < I_{\max}$. Remarquons que les estimations de u sont comprises entre 4,60 et 6,07 (tableau 1), et que u prend une valeur proche du mode de la distribution des intensités, égal à $u((k-1)/k)^{1/k} + w[1 - ((k-1)/k)^{1/k}]$, $k > 1$, dès que $k > 3$ (si k tend vers 1, le mode est plus proche de la borne supérieure w). On aurait pu imposer une contrainte inférieure plus forte que celle de positivité sur ce paramètre en posant $5 \leq u$, puisque seules les intensités supérieures au degré V sont considérées. Mais on observe que, d'une part relâcher cette contrainte fait diminuer les écarts-types, et d'autre part les estimations de u restent très proches de 5,0, ce qui est réaliste puisque la distribution de type III extrapole les petites intensités non prises en compte. Par contre la contrainte $I_{\max} < w \leq 12$, doit être introduite explicitement dans le critère. On s'aperçoit en effet que les estimateurs de ce paramètre viennent systématiquement saturer la contrainte, soit par limite inférieure, soit par limite supérieure (cas 2).

TABLEAU 1
Estimation de la distribution de type III, pour les intensités maximales annuelles, dans les régions françaises

Source	$I_{\max}$	w	$\sigma(w)$	k	$\sigma(k)$	u	$\sigma(u)$	n
Alpes	10,5	10,50	1,14	4,20	1,17	5,58	0,03	131
Artois et Brabant	8,0	8,46	0,73	2,01	0,67	4,97	0,13	11
Bassin Aquitain Nord	7,5	12,00	40,06	12,42	75,79	4,81	0,15	27
Bassin Aquitain Sud	7,5	7,66	0,13	2,17	0,16	5,11	0,01	6
Bassin Parisien Est et Sud	6,0	6,00	—	0,44	—	5,20	—	3
Bassin Parisien Ouest	8,0	10,30	3,01	5,68	3,10	5,41	0,05	32
Bresse	8,0	9,34	3,63	6,78	6,53	4,87	0,05	38
Fossé Rhénan Nord et Rhin Inférieur	8,0	8,00	0,48	2,10	0,69	6,07	0,06	25
Fossé Rhénan Sud	7,5	7,74	0,53	2,05	0,65	4,82	0,11	47
Jura	8,5	8,94	0,45	3,44	0,55	5,23	0,02	63
Jura Souabe	8,5	8,51	0,78	3,37	1,25	5,98	0,04	14
Languedoc	7,5	8,12	0,67	2,28	0,82	5,40	0,07	26
Lion, Ligurie et Maures	7,0	7,00	0,74	1,58	1,14	5,42	0,11	9
Massif Armoricaïn Nord	7,5	12,00	13,12	12,57	25,08	4,66	0,07	45
Massif Armoricaïn Ouest et Sud	7,5	12,00	20,97	15,47	49,36	5,03	0,04	30
Massif Central	8,5	12,00	12,64	8,73	17,64	4,91	0,10	45
Normandie	6,5	6,60	0,06	2,06	0,13	4,86	0,01	28
Poitou	8,0	12,00	9,57	7,13	11,14	4,60	0,17	53
Provence	9,0	12,00	6,80	8,38	9,31	5,24	0,05	48
Pyrénées	10,5	12,00	3,31	6,62	3,50	5,16	0,04	132

Les estimations obtenues après convergence de la procédure de Gauss-Newton sont résumées dans le tableau 1. On peut juger graphiquement la qualité des ajustements de la loi des valeurs extrêmes majorée, comme sur la figure 2 pour la région des Alpes. On constate que les estimations des paramètres de la distribution des valeurs extrêmes respectent bien les valeurs caractéristiques des distributions observées. Toutefois les écarts-types des paramètres w et k sont élevés dans certaines régions. L'analyse des matrices des corrélations asymptotiques révèle que le coefficient de corrélation entre ces deux paramètres est pratiquement égal à 1. Si le paramètre u est donc bien déterminé, l'estimation des paramètres w et k est moins précise, cette instabilité provenant de leur forte corrélation.

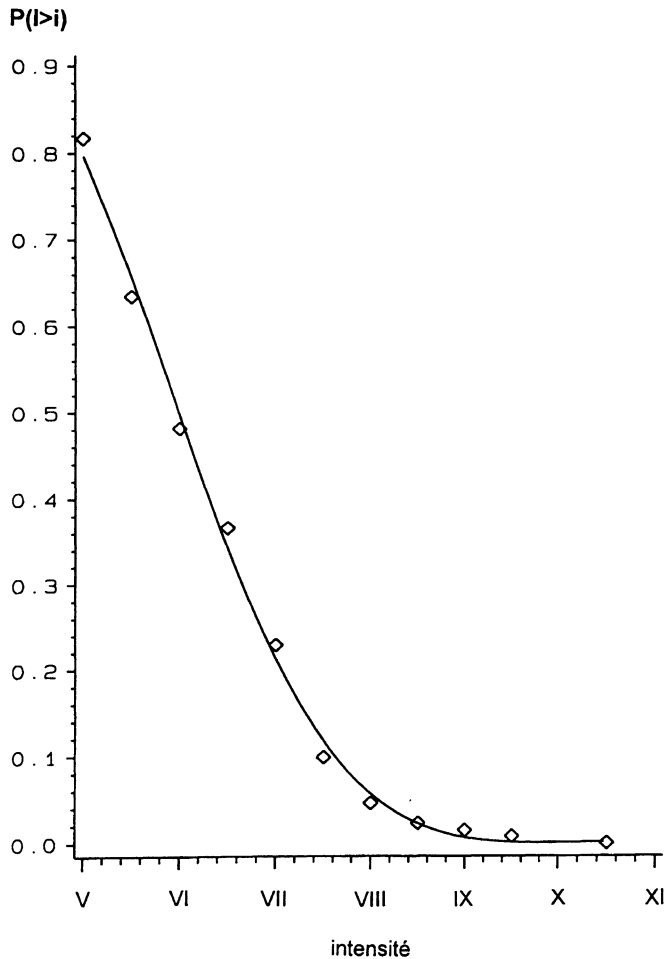


FIGURE 2

*Probabilité de dépassement de l'intensité maximale annuelle
dans la région des Alpes – Distribution de type III*

En fait, les écarts-types sont particulièrement élevés lorsque l'estimation du paramètre w tend vers une borne de son domaine de définition. En effet, on distingue deux cas :

1) Dans sept régions (Bassin Aquitain Nord, Massif Armorica Nord et Sud, Massif Central, Poitou, Provence et Pyrénées), la valeur estimée de w est égale à 12. Ces régions sont donc caractérisées par une distribution de l'intensité épiscopale qui s'étend entièrement sur l'échelle des intensités du degré V au degré XII. La queue de distribution est extrapolée pour ce qui concerne les événements d'intensité supérieure à l'intensité maximale observée, c'est-à-dire caractérisés par des périodes de retour ne permettant pas leur observation. Ce résultat est intéressant pour les régions à fort potentiel sismique concernées, comme les Pyrénées. Le fait que la distribution atteigne la borne supérieure de l'échelle MSK contraint la probabilité de dépassement à tendre vers zéro en cette limite.

2) Pour les autres régions, la borne supérieure de la distribution w tend vers l'intensité maximale observée $I_{\max}$ (Alpes, Bassin Parisien Est et Sud, Fossé Rhénan Nord et Rhin inférieur, Lion-Ligurie et Maures) ou reste dans son voisinage (comme pour le Jura Souabe où $I_{\max} = 8,5$ et $w = 8,51$). Dans ces régions, où la sismicité est faible à modérée (excepté les Alpes), on peut donc admettre que l'intensité maximale observée définit effectivement la limite supérieure du potentiel sismique de la région.

Donc, dans la moitié des régions du zonage, la contrainte sur le paramètre w est saturée soit par limite supérieure, soit par limite inférieure. Ceci est la cause des écart-types parfois anormalement élevés des paramètres w et k . Afin d'obtenir un estimateur de la borne supérieure de la distribution plus fiable, on propose une méthode d'inférence semiparamétrique fondée sur la technique non paramétrique du Bootstrap.

3. Une méthode semiparamétrique pour l'estimation des valeurs extrêmes

3.1. Méthode non paramétrique du Bootstrap

Dans les situations où l'on ne peut pas faire l'hypothèse que la distribution des observations ou de certains paramètres appartient à une famille connue, il existe des méthodes d'inférence statistique adaptées, dites non paramétriques ou «libres de distribution». Ces méthodes reposent sur des postulats moins contraignants que ceux des méthodes paramétriques qui dépendent de façon cruciale des hypothèses émises sur la distribution de l'échantillon de base. On peut appliquer les méthodes non paramétriques à des problèmes d'estimation de paramètres caractéristiques ou de tests d'hypothèses.

La technique du Bootstrap a été introduite par B. Efron (Efron, 1979). C'est la méthode de réplcation des échantillons la mieux fondée théoriquement. Elle consiste à créer, à partir d'un échantillon de base, un grand nombre d'échantillons par tirage aléatoire avec remise. Sur chaque échantillon, les statistiques auxquelles on s'intéresse sont calculées, ce qui permet d'approcher leur dispersion. On peut ainsi estimer la variance ou la loi des paramètres caractéristiques de la distribution de l'échantillon et construire des intervalles de confiance, lorsque la distribution des paramètres est analytiquement complexe. Cette méthode semble donc être adéquate

pour la résolution du problème de l'intensité maximale pour lequel les méthodes paramétriques restent insatisfaisantes.

La méthode du Bootstrap se déroule de la manière suivante :

Première étape : Soient X une variable aléatoire de fonction de répartition F dépendant du paramètre β , et l'échantillon $\varepsilon = (x_1 \dots x_n)$. Soit F_n la distribution empirique associée à ε (i.e. la loi qui à tout x_i de ε associe la masse $1/n$); ε muni de la loi F_n sert de population de base.

Seconde étape : On procède à n tirages équiprobables dans ε , avec remise. On obtient le pseudo-échantillon $\varepsilon^* = (x_1^* \dots x_n^*)$ tel que $\forall i = 1 \dots n, \exists j \in (1 \dots n) / x_i^* = x_j$; c'est-à-dire que, dans ε^* , certaines valeurs initiales sont présentes avec répétition et que d'autres ne sont pas représentées. Puis, on calcule β^* issu de ε^* et F_n^* la distribution empirique associée à ε^* .

Troisième étape : On approche la distribution de β par la distribution empirique de β^* , obtenue par simulation, en itérant N fois l'étape précédente pour obtenir $(\beta_p^*(\varepsilon_p^*, F_n^*))_{p=1 \dots N}$, à partir des N réalisations indépendantes et identiquement distribuées $(\varepsilon_p^*)_{p=1 \dots N}$. On peut donc estimer la variance Bootstrap du paramètre et son intervalle de confiance.

Efron propose une méthode, appelée de façon assez naturelle «méthode des fractiles», pour construire des intervalles de confiance sur les paramètres (Efron, 1987). Les fractiles 5% et 95% de la distribution du Bootstrap donnent l'intervalle de confiance à 90% du paramètre β^* . Il est clair que cette méthode conduit à l'obtention d'intervalles de confiance asymétriques, ce qui est particulièrement pertinent dans le cas de petits échantillons.

Le nombre N d'échantillons détermine la précision de la technique du Bootstrap. Efron a observé que, dans un certain nombre d'applications où les vraies valeurs des paramètres sont connues, les estimations obtenues pour N compris entre 100 et 1 000 sont suffisamment précises et que les intervalles de confiance sont également proches des valeurs exactes, si les extrémités de la distribution simulée sont assez précises.

3.2. Une application semiparamétrique pour les valeurs extrêmes

Dans le cas où l'on cherche à approcher la distribution du maximum d'un échantillon, il est clair que la technique usuelle du Bootstrap pour le rééchantillonnage des données ne convient pas, puisqu'il est impossible d'obtenir des pseudo-échantillons comportant des valeurs supérieures à la valeur maximale observée. On aurait donc ce problème pour l'estimation de la borne supérieure d'une distribution en appliquant directement la méthode du Bootstrap. C'est pourquoi, on utilisera une méthode développée par D. Zeltermann afin d'approximer la distribution des $q + 1$ plus grandes statistiques des valeurs extrêmes d'une distribution F inconnue à partir d'un échantillon ordonné $X_1 > \dots > X_n$ où $n \gg q$ (Zeltermann, 1993). L'auteur propose de rééchantillonner, non plus les valeurs elles-mêmes, mais les intervalles entre les valeurs adjacentes de l'échantillon. Il s'appuie pour cela sur le théorème de Weissman (1978, cf. (Zeltermann, 1993)) :

Pour q fixé, quand $n \rightarrow \infty$, les intervalles des plus grandes statistiques des valeurs extrêmes normalisés $d_j = ja_n^{-1}(X_j - X_{j+1})$, $j = (1 \dots q)$, sont indépendants et identiquement distribués selon une loi exponentielle.

L'exponentialité est valable pour la plupart des distributions communément utilisés. L'auteur suppose de manière générale que $a_n = 1$, cette hypothèse ne semblant pas très restrictive. Les valeurs bootstrap ($d_1^* \dots d_q^*$) des intervalles sont ensuite «dénormalisées» puis sommées, afin de simuler la série des plus grandes statistiques des valeurs extrêmes définies par :

$$X_q^* = X_{q+1} + d_q^*/q \quad \text{et} \quad X_j^* = X_{j+1}^* + d_j^*/j, j = q - 1 \dots 1 \quad (6)$$

Remarquons que la méthode du Bootstrap n'utilise pas le résultat d'exponentialité du théorème mais seulement celui d'indépendance et d'identité de distribution des intervalles. Cette méthode est dite semiparamétrique car la forme de la distribution F n'est pas spécifiée, mais sa queue de distribution est supposée suivre une loi de type exponentiel. Le problème restant est l'estimation de q , ce qui équivaut à déterminer où la queue de distribution de F commence. L'auteur propose un test séquentiel sur les intervalles normalisés $d_1 \dots d_q$ fondé sur l'hypothèse d'exponentialité et la statistique de Gini ou la distance de Kolmogorow-Smirnov.

Dans le problème des intensités maximales, la variable aléatoire X représente l'intensité maximale. On fait l'hypothèse que la fonction de répartition théorique F est la distribution du type III des valeurs extrêmes (de type exponentiel), c'est pourquoi on va appliquer la méthode de rééchantillonnage des intervalles successifs à toute l'étendue des échantillons ($q = n$). Les paramètres estimés pour chacun des N échantillons par l'algorithme de Newton sont le facteur d'échelle k^* , la valeur caractéristique u^* et la borne supérieure w^* de la distribution de type III. Les estimateurs bootstrap (k_B, u_B, w_B) de ces trois paramètres sont les moyennes de leurs distributions simulées respectives. Dans le paragraphe suivant on s'intéressera particulièrement au paramètre w_B , la précision ajoutée à l'estimation des deux autres paramètres par la méthode du Bootstrap n'étant pas suffisamment significative.

3.3. Estimation des intensités maximales régionales

La méthode semiparamétrique précédente a été appliquée aux 20 régions définies par le zonage sismotectonique français. Pour chaque région, 1 000 séries d'intensités maximales annuelles ont été simulées, les unes étant caractérisées par une valeur maximale inférieure à l'intensité maximale observée, les autres par une valeur maximale supérieure. Les estimations des intensités maximales obtenues et arrondies (notées w_B) accompagnées de leur écart-type et de leur intervalle de confiance à 90% (notées I.C. 90%) sont présentées dans le tableau 2 et comparées aux intensités maximales observées ($I_{\max}$).

Les intervalles de confiance sont relativement larges. En dehors de cela, les résultats obtenus sont satisfaisants puisque, dans la majorité des régions, l'intensité maximale estimée par la méthode du Bootstrap est de 0,5 degré à 1,5 degré supérieure à l'intensité maximale observée. Seule la région des Pyrénées a une intensité maximale estimée égale à l'intensité maximale observée. En moyenne, la valeur bootstrap de l'intensité maximale en France est de un degré supérieure à l'intensité maximale observée.

4. Conclusion

Si on émet l'hypothèse que l'intensité des séismes susceptibles de se produire dans une région donnée est limitée physiquement, ce qui semble être le cas notamment dans les régions à faible sismicité, l'utilisation d'une distribution majorée telle la loi de type III de la théorie des valeurs extrêmes se justifie et permet de respecter au mieux la queue de distribution des intensités observées. De plus, l'approche semiparamétrique proposée pour la détermination des intensités maximales régionales, montre que l'on peut extrapoler les probabilités d'occurrence des séismes supérieurs à l'événement historique majeur, dans un intervalle de 0,5 degré à 1,5 degré (selon la région). Cette démarche est fiable dans la mesure où la distribution de type III des valeurs extrêmes reste très fidèle aux données disponibles concernant les événements majeurs, et dans la mesure où l'on pourra, par la suite, évaluer l'influence des événements extrêmes sur l'estimation de l'aléa sismique pour les périodes de retour en jeu dans chaque étude. L'inférence statistique non paramétrique constitue un bon outil d'aide à la décision dans le choix de certaines lois et de certains paramètres de l'analyse probabiliste.

TABLEAU 2
Intensités maximales régionales bootstrap

<i>Source</i>	$I_{\max}$	w_B	$\sigma(w_B)$	I.C. 90%
Alpes	10,5	11,0	1,5	[8,0;12,0]
Artois et Brabant	8,0	9,0	2,0	[6,5;12,0]
Bassin Aquitain Nord	7,5	9,0	2,0	[6,5;12,0]
Bassin Aquitain Sud	7,5	8,5	1,5	[6,5;12,0]
Bassin Parisien Est et Sud	6,0	7,0	1,0	[6,5; 9,5]
Bassin Parisien Ouest	8,0	9,5	2,0	[6,5;12,0]
Bresse	8,0	9,0	2,0	[6,5;12,0]
Fossé Rhénan Nord et Rhin Inférieur	8,0	9,5	2,0	[6,5;12,0]
Fossé Rhénan Sud	7,5	9,0	2,0	[6,5;12,0]
Jura	8,5	9,5	2,0	[7,0;12,0]
Jura Souabe	8,5	9,5	2,0	[7,0;12,0]
Languedoc	7,5	9,0	2,0	[6,5;12,0]
Lion, Ligurie et Maures	7,0	8,0	1,5	[6,5;12,0]
Massif Armoricaïn Nord	7,5	9,0	2,0	[6,5;12,0]
Massif Armoricaïn Ouest et Sud	7,5	9,0	2,0	[6,5;12,0]
Massif Central	8,5	9,5	2,0	[6,5;12,0]
Normandie	6,5	8,5	2,0	[6,5;12,0]
Poitou	8,0	9,5	2,0	[7,0;12,0]
Provence	9,0	10,0	2,0	[7,0;12,0]
Pyrénées	10,5	10,5	1,5	[7,0;12,0]

Remerciements : Je tiens à remercier le Professeur P. Cazes pour les conseils qu'il m'a apportés pendant la réalisation et la rédaction de cette étude. Je remercie également le Professeur E. Diday pour les orientations bibliographiques qu'il m'a suggérées au cours des recherches.

Références bibliographiques

- [1] CORNELL C.A. (1968). Engineering seismic risk analysis. *Bulletin of the Seismological Society of America*. Vol. 58, No. 5, p. 1583-1606.
- [2] EFRON B. (1979). Bootstrap methods : another look at the Jackknife. *The Annals of Statistics*. Vol. 7, No. 1, p. 1-26.
- [3] EFRON B. (1987). Better Bootstrap confidence intervals. *Journal of the American Statistical Association*. Theory and Methods, Vol. 76, No. 374, p. 312-319.
- [4] GUMBEL E.J. (1967). Statistics of extremes. *New York and London : Columbia University Press*. Fourth Printing.
- [5] GRELLET B. *et al.* (1993). Sismotectonique de la France métropolitaine dans son cadre géologique et géophysique. *Mémoires de la Société Géologique de France*. Mémoires N.S. No. 164, Vol. 1 et 2.
- [6] KIJKO A., DESSOKEY M.M. (1987). Application of the extreme magnitude distributions to incomplete earthquake files. *Bulletin of the Seismological Society of America*. Vol. 77, No. 4, p. 1429-1436.
- [7] SCHENKOVA Z., KARNIK V. (1970). The probability of occurrence of largest earthquakes in the european area. *Pure and Applied Geophysics*. Vol. 80, No. 3, p. 152-161.
- [8] ZELTERMAN D. (1993). A semiparametric Bootstrap technique for simulating extreme order statistics. *Journal of the American Statistical Association*. Theory and Methods, Vol. 88, No. 422, p. 477-485.