Société de Calcul Mathématique, S. A.
Algorithmes et Optimisation

Analyse probabiliste de la séismologie

Construction d'une loi de probabilité :

prise en compte des incertitudes sur les mesures

Document adressé au

Commissariat à l'Energie Atomique

Direction de l'Energie Nucléaire

Par la

Société de Calcul Mathématique S. A.

en application du contrat no 4000288990/P5B61, 26 mai 2007

rédaction Bernard Beauzamy

Décembre 2007

Siège social et bureaux : 111, Faubourg Saint Honoré, 75008 Paris. Tel. : 01 42 89 10 89. Fax : 01 42 89 10 69
Société Anonyme au capital de 56 200 Euros. RCS : Paris B 399 991 041. SIRET : 399 991 041 00035. APE : 731Z

Pour de très nombreux phénomènes, les informations sur les mesures sont de moins en moins précises à mesure qu'on recule dans le passé : les instruments n'avaient pas la même qualité qu'aujourd'hui. C'est typiquement le cas pour la météorologie ou la séismologie : les informations relatives aux grands séismes du passé sont souvent très vagues. Mais on ne peut les ignorer : si on se limite aux données récentes, on ne dispose pas d'un échantillon suffisant pour incorporer des phénomènes dont la durée de retour est grande.

Si on fait l'hypothèse que les événements suivent tous la même loi (par exemple de Poisson) et sont indépendants (pas d'influence de l'un d'eux sur l'apparition des autres), alors certainement observer 20 sites sur 10 ans donnera le même résultat, en nombre de séismes d'une magnitude donnée, qu'observer 2 sites sur 100 ans, ou bien 200 sites sur un an. Toutefois, la précision des mesures ne sera pas la même : dans le cas de mesures récentes, elle sera meilleure.

On est donc amené à se poser la question : comment définir une loi de probabilité tenant compte des incertitudes sur les mesures ? Ces incertitudes, éventuellement, peuvent être différentes d'une mesure à l'autre. C'est cette question qui est résolue dans ce document.

1. Construction traditionnelle d'une loi de probabilité à partir d'un échantillon

Si l'on dispose d'un échantillon $x_1, x_2, ..., x_N$ de valeurs observées (par exemple des températures en un point, des magnitudes de séisme en un point), on constitue une loi de probabilité (tout simplement un histogramme) en affectant la probabilité $1/N$ à chacune des valeurs prises. Si les valeurs prises sont distinctes, chacune aura donc la même probabilité ; s'il y a des répétitions, elles sont ainsi prises en compte : si une valeur est répétée 7 fois, elle aura la probabilité $\frac{7}{N}$.

La loi de probabilité associée au phénomène est donc, de manière formelle, définie par :

$$P = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i} \qquad (1)$$

où δ_{x_i} est la mesure de Dirac associée au i-ème point de l'échantillon. Rappelons que, pour un point x quelconque, δ_x est une mesure, définie par $\delta_x(A) = 1$ si l'ensemble A contient x, 0 sinon.

Le choix d'une mesure de Dirac dans la formule (1) correspond à une précision infinie : il n'y a pas d'incertitude sur la mesure.

Si on décide de construire un histogramme, on divise l'intervalle complet en petits sous-intervalles, et on met la valeur $1/N$ à chaque fois qu'un point de l'échantillon tombe dans l'un des sous-intervalles. Cette manière de procéder est très simple, mais le découpage est à la fois régulier et arbitraire : il ne permet pas de rendre compte des incertitudes variables sur toute la gamme de mesure. Voir Beauzamy-Zeydina [2] pour d'autres méthodes de construction d'un histogramme.

2. Prise en compte de l'incertitude sur les mesures

Supposons maintenant que chacune des mesures soit affectée d'une incertitude. Dans le cas le plus simple, ce sera une plage de valeurs $[x_i - \varepsilon_i, x_i + \varepsilon_i]$: on fait alors l'hypothèse d'une loi uniforme dans cette plage. De manière pratique, cela signifie que la mesure donne x_i, mais que l'on considère que la vraie valeur peut être n'importe où dans l'intervalle $[x_i - \varepsilon_i, x_i + \varepsilon_i]$. L'hypothèse "uniforme" signifie qu'aucune valeur n'est privilégiée dans cet intervalle.

On peut vouloir faire d'autres hypothèses quant à l'incertitude sur la mesure, par exemple dire qu'elle est gaussienne, de moyenne x_i et de variance dépendant de la gamme de mesure. En effet, la précision de la mesure, donc la variance de la loi, se détériorent aux deux extrémités de la gamme (voir [1] pour ce qu'on appelle « effet d'échelle »). Dans le cas présent, on mesure plus mal les séismes très faibles et très forts que les séismes moyens.

De manière générale, admettons donc que nous ayons une densité de probabilité f_i relative à la i-ème mesure. Cette densité de probabilité est relative à la qualité de la i-ème mesure, et n'a rien à voir avec les autres. Pour les séismes, elle caractérisera par exemple le fait que la i-ème mesure est ancienne ou récente, prise en un point bien appareillé ou en un point mal surveillé, etc.

La densité f_i n'a pas de raison d'être symétrique, mais elle vérifie :

$$\int_{-\infty}^{+\infty} f_i(t)dt = x_i \,, \tag{2}$$

hypothèse qui signifie que la valeur moyenne de la densité (l'espérance) est égale à la valeur observée.

Cette densité permet d'avoir la probabilité que la valeur vraie diffère de la valeur observée :

$$P\left\{\left|X_i - x_i\right|\right\} > \varepsilon = \int_{-\infty}^{-\varepsilon} f_i(t)dt + \int_{\varepsilon}^{+\infty} f_i(t)dt \tag{3}$$

Il est important de noter que la probabilité f_i n'est jamais véritablement connue : elle est évaluée par calibration de l'instrument (voir B. Beauzamy [1]). Si on ne sait rien, pour se pénaliser on met une loi uniforme entre deux bornes extrêmes.

Une fois chacune des lois f_i choisies, la probabilité globale P s'obtient par une formule analogue à la formule (1) :

$$P = \frac{1}{N}\sum_{i=1}^{N} f_i \tag{4}$$

C'est bien une densité de probabilité, mais ce n'est plus une moyenne de masses de Dirac.

3. Exemples pratiques

a) Cas de lois d'erreurs uniformes

Commençons par un exemple très simple : dix mesures, avec un intervalle de taille 10^{-2} autour de chacune, et une loi uniforme pour les incertitudes. Voici la liste des valeurs obtenues, rangées par ordre croissant :

0,014
0,290
0,302
0,533
0,580
0,706
0,709
0,761
0,775
0,814

Si l'on tient compte des incertitudes, ± 0.01 de part et d'autre de chaque valeur, on obtient le tableau suivant :

0,004	0,024
0,280	0,300
0,292	0,312
0,523	0,543
0,570	0,590
0,696	0,716
0,699	0,719
0,751	0,771
0,765	0,785
0,804	0,824

Et la densité de probabilité que nous recherchons est la moyenne des fonctions caractéristiques de chaque intervalle, normalisées :

$$f = \frac{1}{10}\sum_{i=1}^{10} f_i \tag{5}$$

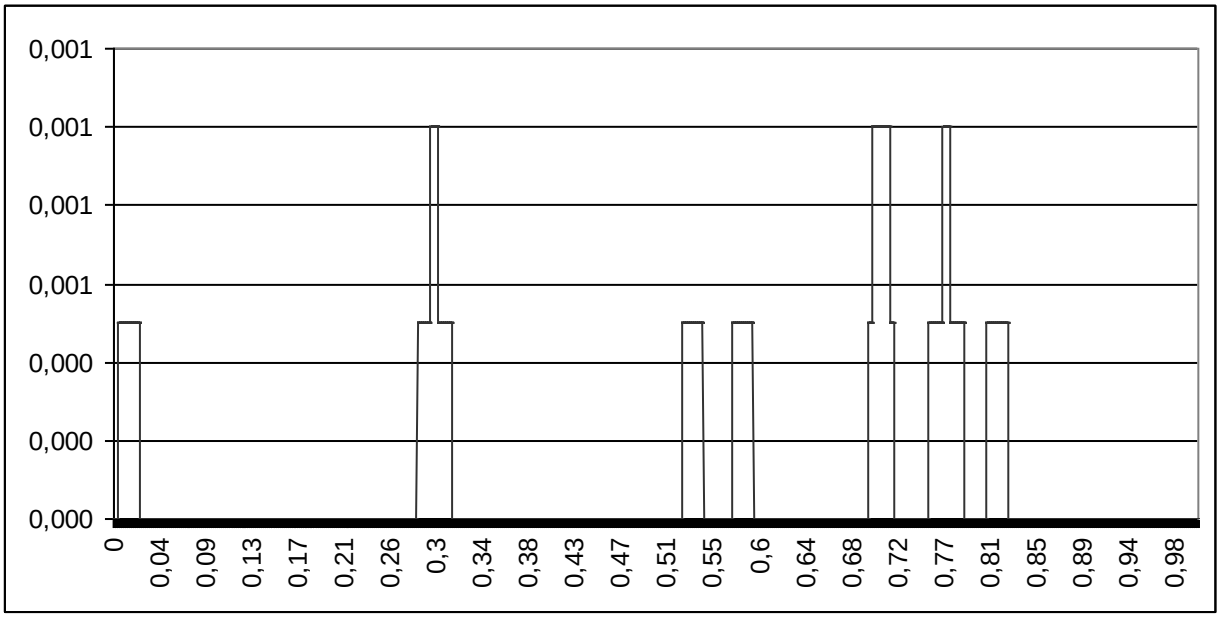
avec :

$$f_i = \frac{1}{0.02} 1_{I_i} \tag{6}$$

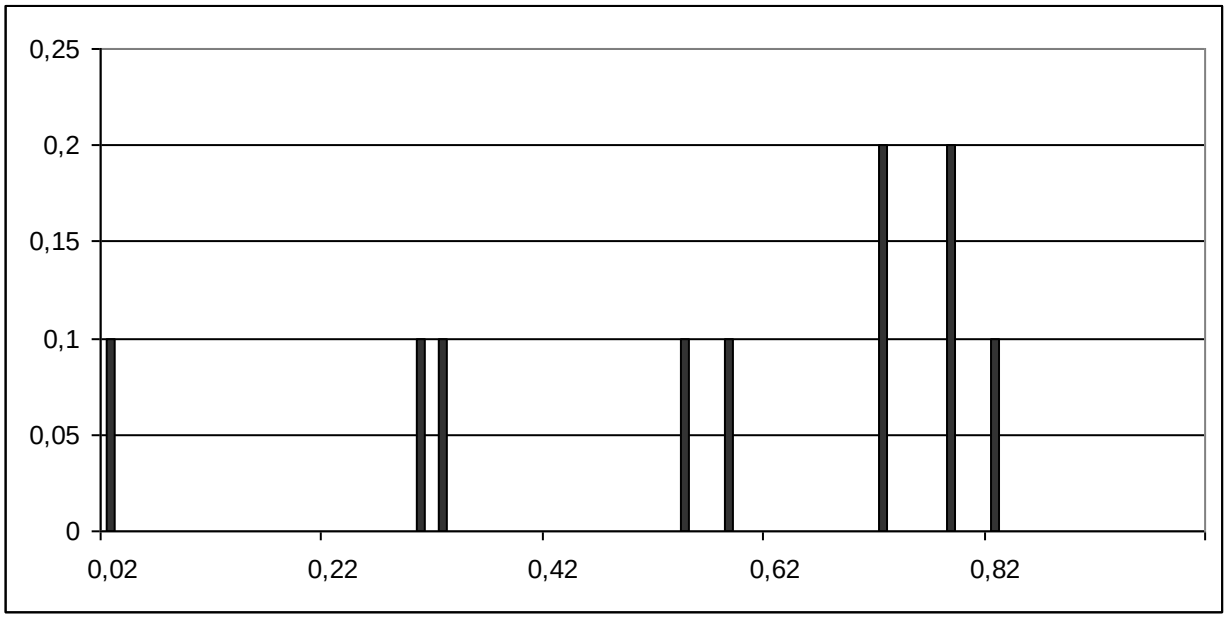
où 1_A désigne la fonction caractéristique de l'ensemble A : $1_A(x) = 1$ si $x \in A$, 0 sinon. Le facteur $\frac{1}{0.02}$ correspond à une normalisation : chaque f_i est une densité de probabilité, et la largeur de chaque intervalle I_i est 0.02.

Il est intéressant de remarquer que ces ensembles I_i ne sont pas disjoints.

Voici le graphe de la fonction f ainsi obtenue :

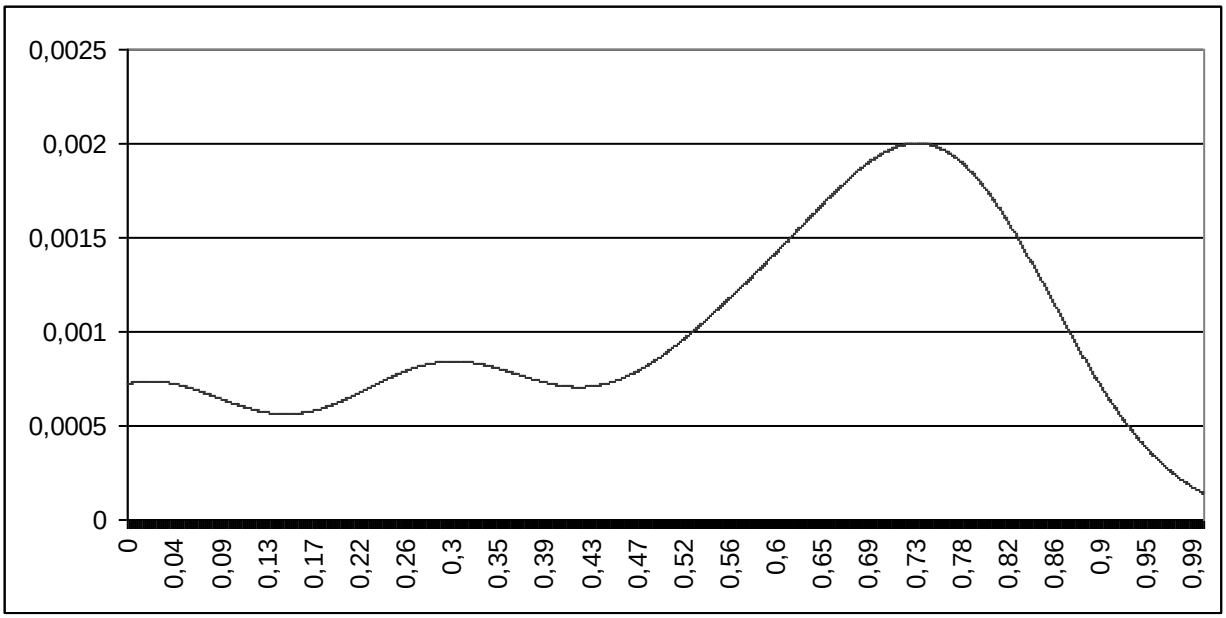


Comparons maintenant avec la construction usuelle d'un histogramme :

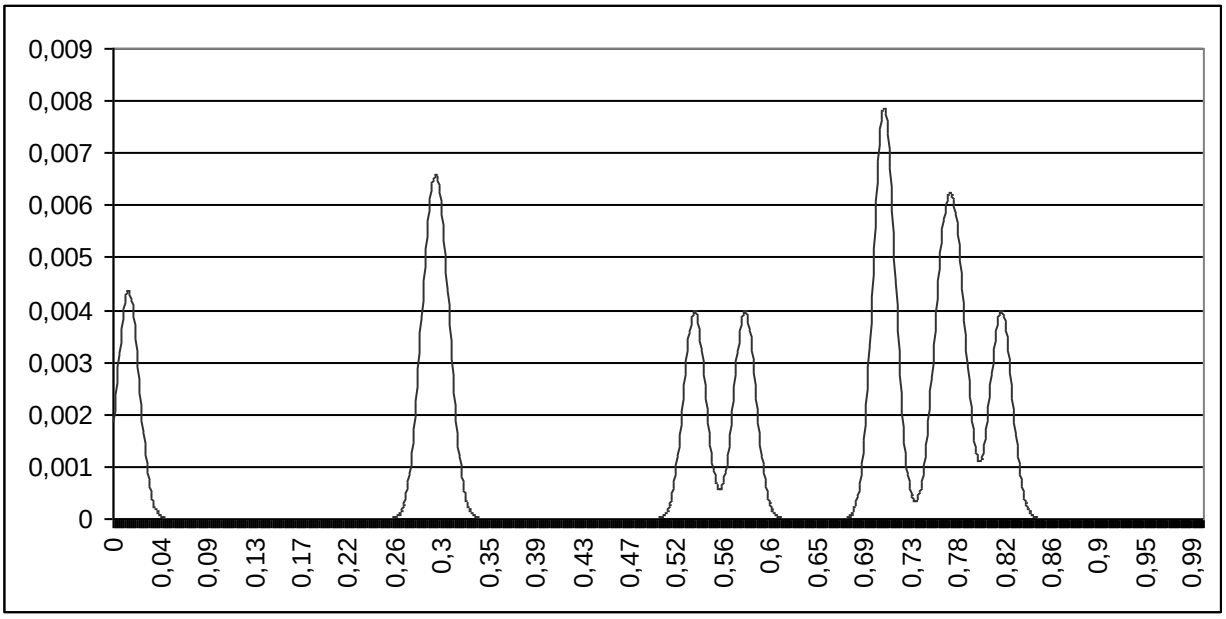


b) Cas de lois d'erreurs gaussiennes

Voici maintenant le graphe, supposant que les erreurs sont des gaussiennes, de moyenne égale aux points de mesure, de variance $\sigma = 0.1$, tronquées entre 0 et 1, et renormalisées :



Et avec $\sigma = 0.01$:



On constate ainsi que ce procédé de construction est à la fois plus souple et plus réaliste que l'histogramme : on peut incorporer la loi d'erreur que l'on veut ; elle peut être différente d'une mesure à l'autre.

4. Prise en compte des incertitudes dans des probabilités conditionnelles

L'étude, en particulier, des lois de propagation fait apparaître l'usage de probabilités conditionnelles : on recherche des énoncés du type « sachant qu'il y a eu un séisme de magnitude tant dans telle zone source, quelle est la probabilité d'observer telle PGA dans telle zone cible ?». Comme nous l'avons montré dans notre rapport 3, pour cela on sélectionne les événements « source » (telle magnitude) et on construit la loi de probabilité « cible » à partir de ces événements : quelle est la répartition des PGA possibles pour une magnitude source donnée ?

Voyons comment incorporer l'information d'incertitude dans cette construction.

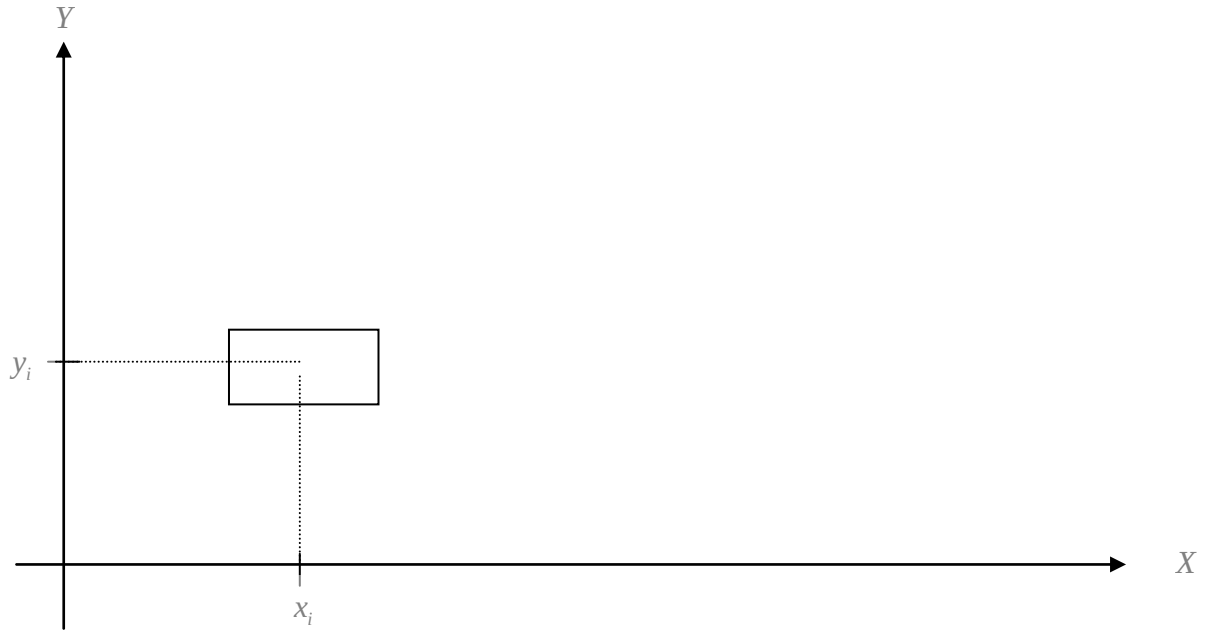
Nous avons deux variables aléatoires X et Y ; dans notre exemple $X =$ PGA cible et $Y =$ magnitude source. Nous nous intéressons à la loi de X sachant Y. Elle est donnée en pratique sous forme de loi conjointe : on dispose d'enregistrements de séismes pour lesquels on a noté simultanément X et Y (ou, plus exactement, pour chaque événement on a enregistré la date, et on peut apparier l'événement source et l'événement cible).

On dispose donc d'un tableau, avec X en ligne et Y en colonne ; ce tableau contient un nuage de points (x_i, y_i), $i = 1, ..., N$, où N est le nombre total d'enregistrements.

X\Y	1	2	3	4
1			(x_2, y_2)	(x_5, y_5)
2		(x_1, y_1)		
3			(x_4, y_4)	
4	(x_3, y_3)			

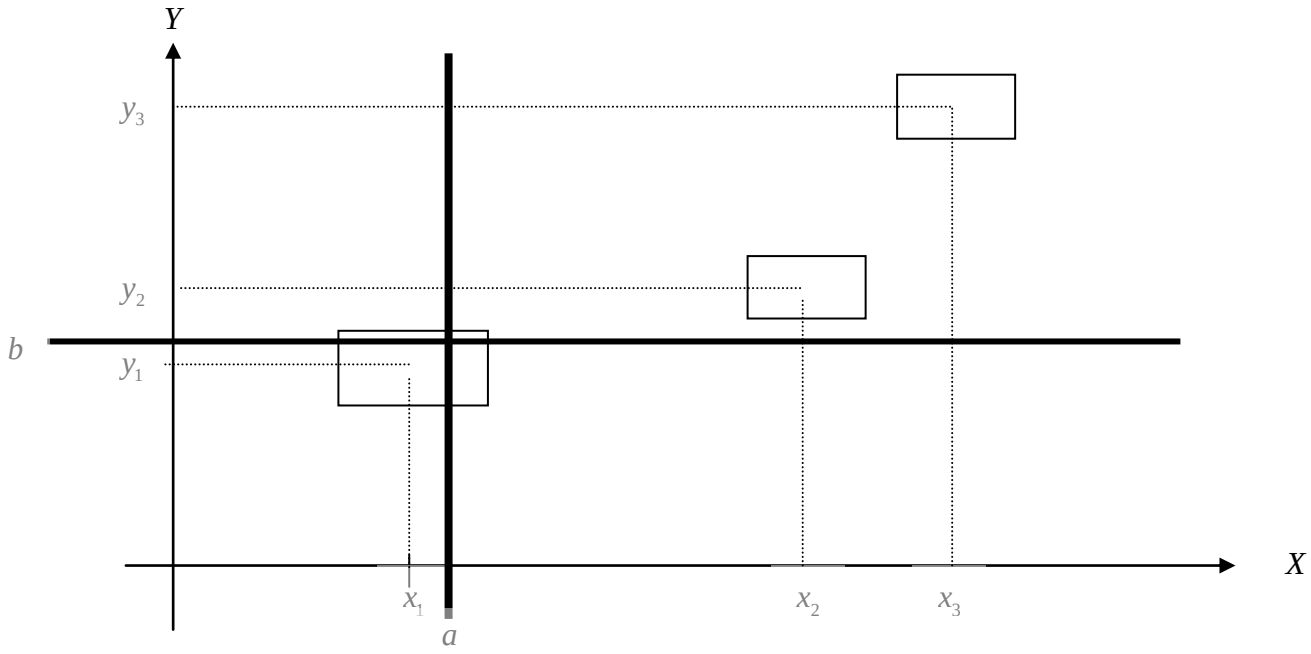
Les colonnes représentent les valeurs de magnitudes possibles pour Y (découpées en intervalles de largeur appropriée) et les lignes les valeurs de PGA possibles pour X (également découpées en intervalles ; dans le tableau ci-dessus, $x_2 \approx x_5$). La loi de X sachant Y est obtenue en sélectionnant une colonne et en fabriquant la loi de probabilité de X à partir de cette colonne.

Voyons maintenant comment introduire les incertitudes. La i-ème observation (x_i, y_i) est entachée d'une incertitude ; nous notons f_i la loi de probabilité de l'erreur pour X et g_i la loi de probabilité de l'erreur pour Y. Supposons, pour simplifier la présentation, que ce soient des lois uniformes (l'intervalle n'étant pas le même d'une observation à l'autre, et n'étant pas le même pour X et pour Y). Nous avons alors un petit rectangle d'incertitude autour du point d'observation (x_i, y_i) :



Voyons comment calculer, par exemple, $P\{X \geq a \,|\, Y \geq b\}$: probabilité d'enregistrer dans la zone cible une PGA supérieure ou égale à un seuil a sachant que l'on a enregistré dans la zone source un séisme de magnitude supérieure ou égale à un seuil b. Par définition des probabilités conditionnelles,

$$P\{X \geq a \,|\, Y \geq b\} = \frac{P\{X \geq a \cap Y \geq b\}}{P\{Y \geq b\}} \tag{7}$$



Sur le dessin ci-dessus, nous avons représenté trois petits rectangles, correspondant à trois observations, et les seuils a et b.

La probabilité $P\{Y \geq b\}$ (loi marginale) se calcule en regardant uniquement l'axe des y : c'est :

$$P\{Y \geq b\} = \sum_{i=1}^{N} \int_{b}^{+\infty} g_i(y)dy \tag{8}$$

Cela revient à faire la somme de tous les intervalles en y qui sont au-dessus de b ; si un intervalle est partiellement au dessus de b, il est compté au prorata de sa largeur (par exemple, s'il est aux 2/3 au dessus de b, on le compte pour 2/3).

Le calcul du numérateur $P\{X \geq a \cap Y \geq b\}$ se fait exactement selon le même principe : on compte tous les rectangles situés à droite de a et au dessus de b ; si un rectangle est à cheval sur l'une des lignes ou à cheval sur les deux, il est compté au prorata de la surface au dessus et à droite. Cela donne la formule :

$$P\{X \geq a \cap Y \geq b\} = \sum_{i=1}^{N} \int_{a}^{+\infty} f_i(x)dx \int_{b}^{+\infty} g_i(y)dy \tag{9}$$

Le calcul se réduit donc à une somme d'aires de rectangles, dans le cas de lois uniformes. Dans le cas d'une loi quelconque, on emploie directement la formule :

$$P\{X \geq a \mid Y \geq b\} = \frac{\sum_{i=1}^{N} \int_{a}^{+\infty} f_i(x)dx \int_{b}^{+\infty} g_i(y)dy}{\sum_{i=1}^{N} \int_{b}^{+\infty} g_i(y)dy} \tag{10}$$

Le calcul usuel de la loi conditionnelle, à savoir compter des points dans les cases d'un tableau, est donc remplacé, lorsqu'on prend en compte les incertitudes, par des calculs d'aires de rectangles (cas d'une loi uniforme) ou d'intégrales (cas de lois quelconques).

5. Travail proposé sur un exemple concret en sismologie

Nous proposons la mise en œuvre des méthodes ci-dessus sur un exemple concret. Il nous faudrait :

- Des enregistrements en zone source ;
- Des enregistrements en zone cible ;
- Vos estimations quant aux incertitudes sur chacun de ces enregistrements (par exemple : précision dans le passé, dans le présent, tout ceci de manière grossière).

Le résultat fourni serait une loi de probabilité, tenant compte des incertitudes sur les données. Cette loi serait du type suivant : probabilité d'avoir une PGA supérieure à telle valeur dans la zone cible sachant qu'on a eu un séisme de magnitude supérieure à tant dans la zone source.