HAL
open science

Le rôle pivot des variables latentes pour le raisonnement statistique conditionnel. Exemples d'applications environnementales des processus ponctuels marqués

Eric Parent, Jacques Bernier, Jean-Jacques Boreux

To cite this version:

Eric Parent, Jacques Bernier, Jean-Jacques Boreux. Le rôle pivot des variables latentes pour le raisonnement statistique conditionnel. Exemples d'applications environnementales des processus ponctuels marqués. Revue de Statistique Appliquée, 2006, 54 (4), pp.85-111. hal-02657973

HAL Id: hal-02657973

https://hal.inrae.fr/hal-02657973v1

Submitted on 30 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Rev. Statistique Appliquée, 2006, **LIV** (4), 85-111

LE RÔLE PIVOT DES VARIABLES LATENTES POUR LE RAISONNEMENT STATISTIQUE CONDITIONNEL. EXEMPLES D'APPLICATIONS ENVIRONNEMENTALES DES PROCESSUS PONCTUELS MARQUÉS

Eric PARENT[1], Jacques BERNIER[1,2] et Jean-Jacques BOREUX[2]

[1]*Laboratoire de MOdélisation, Risque, Statistique, Environnement, ENGREF, 19, Avenue du Maine, F-75732 Paris Cedex 15,*
[2]*Département des Sciences et Gestion de l'Environnement, ULg, Avenue de Longwy, 185, B-6700 Arlon,*
Parent@engref.fr, Jacques.Bernier2@wanadoo.fr, jj.Boreux@ulg.ac.be

RÉSUMÉ

En statistique bayésienne, le raisonnement conditionnel probabiliste fonde la solidarité étroite entre modélisation et inférence. Dans le cadre de ce raisonnement, un rôle pivot est joué par les variables latentes. Ce mode de raisonnement peut s'avérer utile aussi en statistique fréquentiste. On présente trois séries hydrométéorologiques décrites par des modèles de processus ponctuels marqués de complexité croissante. Sur ces exemples, on montre comment récolter les fruits de la souplesse de modélisation et des facilités de calcul apportées par les variables latentes. Ces variables améliorent la conceptualisation de variables de structure dans les modèles statistiques et permettent aux calculs d'inférence de bénéficier des algorithmes MCMC en relation avec les techniques « d'augmentation de données ».

Mots-clés : *Modélisation bayésienne en environnement, méthodes de Monte Carlo par chaînes de Markov, variables latentes, modélisation graphique.*

ABSTRACT

Conditional reasoning is the cornerstone of both steps of Bayesian analysis : stochastic models can be naturally specified by assembling conditionnal layers of variables and inference - via MCMC algorithms- also relies on probabilistic conditionning. Conditional reasoning casts the stage ligths onto latent variables. It may be useful, even in a frequentist perspective. We emphasize the use of latent variables in the analysis of three different marked points models applied to meteorogical data. The explicit introduction of such hidden variables helps for a better conceptualisation of a realistic model. Together with the « augmentation data algorithms » it also provides a straightforward treatment of the inferential problems.

Keywords : *Bayesian models for environmental data, Monte Carlo Markov Chain methods, latent variables, graphical modelling.*

1. Introduction

1.1. Le conditionnement probabiliste, direct pour la modélisation, inverse pour l'inférence bayésienne

L'approche bayésienne (Gelman *et al.*, 1995) a surmonté les difficultés d'inférence statistique par l'emploi de techniques de simulation des lois *a posteriori*, en particulier grâce aux algorithmes Monte Carlo par chaînes de Markov (MCMC). L'évaluation des lois *a posteriori* décrivant les valeurs probables des paramètres d'un modèle s'effectue aujourd'hui sans difficulté même pour des modèles possédant un grand nombre de paramètres ou pour des structures complexes comme les modèles hiérarchiques échangeables (Tanner, 1996; Robert et Casella, 1999). Une conséquence majeure, quoique peu soulignée, de cette révolution statistique déclenchée par l'emploi des MCMC dans le cadre bayésien (Brooks, 2003), c'est une surprenante connivence entre démarche de construction du modèle et calcul d'inférence : les deux étapes s'effectuent dans le même cadre du calcul des probabilités et l'inférence bayésienne n'est finalement qu'un « simple » renversement du conditionnement des assemblages probabilistes que constituent les équations du modèle.

Si dans cet article, l'exposé suit souvent la piste bayésienne, un peu de recul montre que ces avancées ne s'inscrivent pas dans le cadre d'une querelle d'école fréquentiste ou bayésienne de plus, mais qu'il s'agit du plein emploi de la théorie du calcul conditionnel des probabilités : bref, le statisticien quelle que soit sa culture, doit tirer plein parti du raisonnement statistique conditionnel.

1.2. Notations

On notera $[A]$ la probabilité de l'évènement aléatoire A et on utilisera indifféremment le signe d'intégration $\int$ pour le calcul d'intégration avec des mesures discrètes ou diffuses. Ainsi la loi des probabilités conditionnelles s'écrira $[A, B] = [A\,|B]\,[B]$ et le calcul des probabilités totales s'exprime par $[A] = \int_z [A\,|z]\,[z]\,dz$ que z soit une grandeur conditionnante à support discret partitionnant l'espace de définition ou une grandeur aléatoire à support réel.

1.3. Le rôle pivot des variables latentes

La modélisation graphique met en lumière la conception conditionnelle de la structure d'un modèle statistique. Les graphes acycliques orientés (ou DAG pour *direct acyclic graph* selon la terminologie de Spiegelhalter *et al.*, 1996) offrent une représentation sagittale qui relient des grandeurs stochastiques par des arcs orientés, traduisant le conditionnement probabiliste : les sommets d'entrée du DAG sont les paramètres (noeuds sans parents, notés θ dans cet article), tandis que les noeuds de sortie sont les variables observées (noeuds sans enfants, désignés génériquement par la lettre y). Quel est donc le statut des autres grandeurs (notées ci-après z) en *sandwich* entre les deux premières? Ce sont les variables latentes.

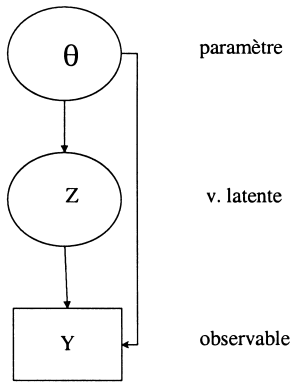

FIGURE 1

Graphe acyclique direct avec variable latente. Les grandeurs aléatoires sont représentées par des cercles, les grandeurs observées par des rectangles

L'objectif de cet article est d'illustrer sur des applications environnementales l'importance de ces variables latentes. De fait, elles jouent un double rôle que ne laisse pas présager la simplicité du DAG représenté sur la figure 1 :

– elles révèlent d'abord la structure profonde des modèles et facilitent la compréhension du phénomène sous-jacent (elles facilitent le dialogue même avec le non-statisticien),
– dans le contexte bayésien, elles favorisent de plus la mise en oeuvre de techniques puissantes d'inférence.

Adoptant un point de vue plus statistique que probabiliste, Gelman, 2003, propose d'interpréter le graphe de conditionnement probabiliste de la figure 1 en généralisant la notion de modèle et de vraisemblance. Son objectif de statisticien est de mettre en évidence l'information *réelle* résultant d'une politique de collecte des données particulière. En d'autre termes, partant des paramètres θ, le *modèle étendu* fournit d'abord des observables potentielles z. Puis une procédure d'observation $[y\,|z,\theta]$ particulière au cas spécifié donne des grandeurs observées y, qui finalement ne sont que la manifestation partielle mais tangible des grandeurs cachées que sont les variables latentes z ou les paramètres θ. Gelman appelle *vraisemblance étendue* la probabilité $[y,z\,|\theta]$. D'une certaine façon la statistique classique fréquentiste visera à se débarrasser des variables latentes, en retournant trop rapidement à la vraisemblance des paramètres θ vis-à-vis des seules observées y, exprimée par intégration de la *vraisemblance étendue* :

$$[y\,|\theta] = \int_z [y,z\,|\theta]\,dz = \int_z [y\,|z,\theta]\,[z\,|\theta]\,dz \tag{1}$$

A contrario, la démarche bayésienne, non seulement explicite les variables latentes z mais, de plus, les intègre au mode de calcul en chaîne des distributions *a posteriori* des paramètres. D'un point de vue strictement technique, cette intégration

est permise grâce aux techniques *d'augmentation des données* de Tanner, 1996, qui exploitent les constructions conditionnelles parallèles du modèle et de l'algorithme MCMC. D'un point de vue conceptuel, l'introduction de variables latentes enrichit extraordinairement la description phénoménologique. Ainsi quoique par définition, la variable latente n'existe tout bonnement que par la structure du modèle, le modélisateur a pris l'habitude de procéder à des raffinements et de distinguer :

– *les variables latentes structurelles,* proprement dites, décrivant une part des conditionnements internes des modèles,
– *les* « *données manquantes* », qui phénomènologiquement s'interprètent comme des observables non observées,
– *les variables prédictives* qui interviennent par leurs distributions prédictives liées aux données futures dans une optique de validation ou pour l'expression de coûts de décisions éventuelles.

Dans cet article, nous adoptons tour à tour les points de vue bayésien et fréquentiste afin d'illustrer les avantages de l'explicitation des variables latentes en relation avec le raisonnement conditionnel sur l'analyse d'exemples issus d'une même classe de structure de *modèle étendu* : les processus ponctuels marqués. Une telle structure contient en germes de nombreuses variables potentiellement observables et des mécanismes variés de production de variables effectivement observées (Parent et Bernier, 2001). Sur un plan phénoménologique, on va, pour la suite de cet article, imaginer qu'il s'agit de la représentation des processus de précipitations atmosphériques observés sur différentes échelles de temps ou d'espace. Mais l'intérêt de ces modèles et de leur traitement bayésien dépasse largement le domaine de l'environnement.

1.4. Le processus ponctuel de Poisson à marques exponentielles forme la base du **modèle étendu**

La figure 2 présente la trajectoire théorique d'un processus ponctuel marqué (Snyder, 1975) où des événements supposés ponctuels (sans durée) surviennent aléatoirement dans le temps. À chaque occurrence d'un événement (la date d'une averse pour fixer les idées) est associée une marque X_i, variable aléatoire représentant ici pour développer cet exemple environnemental plus avant, la quantité d'eau générée au cours de cette averse.

Le modèle le plus simple de ce type suppose des événements survenant dans le temps selon un processus de Poisson tel que :

– N est le nombre d'événements sur un intervalle fixé $[0, T]$. Cette variable aléatoire est distribué selon la « loi de Poisson » :

$$[N = n \,|\mu, T] = e^{-\mu T} \frac{(\mu T)^n}{n!} \quad N = 0, 1, 2, \cdots, n, \cdots \tag{2}$$

– les marques X_i successives sont indépendantes et sont distribuées selon une même loi $F(x)$. La loi « exponentielle » est un archétype parcimonieux (et généralement

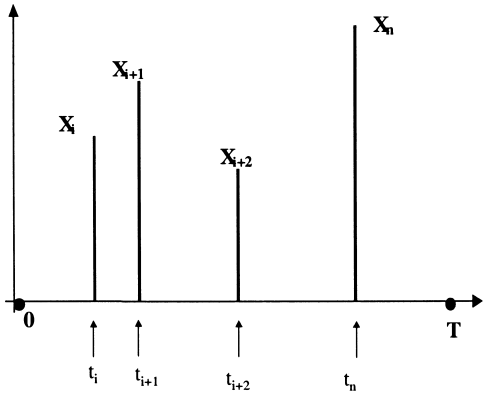

FIGURE 2

Un exemple de trajectoire de processus ponctuel marqué

préconisé par les hydrologues pour représenter par exemple la quantité d'eau générée par une averse) :

$$[X_i = x|\rho] = \rho e^{-\rho x} \tag{3}$$

Nous allons voir que cette structure étendue avec seulement deux paramètres ($\theta = (\mu, \rho)$), issue des formules (2)+(3) possède de multiples applications dans le domaine de l'environnement. Dans toute la suite, on appellera $Z = (N, X_1, ...X_N)$.

Dans la section 2, on cherche à caractériser l'intensité des précipitations et leur nombre à partir d'un jeu de données de pluies (à Bar sur Seine) sur une période donnée. Dans ce premier cas élémentaire, toutes les variables potentiellement observables du *modèle étendu* sont directement observées :

$$Y = Z = (N, X_1, ...X_N) \tag{4}$$

Dans la section 3, on associe à cette même structure de *modèle étendu* (2)+(3), un autre type de collecte d'observations constituée des valeurs « extrêmes ». Par exemple, si les X_i représentent les crues de la Seine, l'hydrologue peut n'avoir à sa disposition que le maximum sur la période de durée T, généralement l'année. Dans ce second cas, la variable observée se réduit à :

$$Y = \max_i(X_i) \tag{5}$$

La section 4 reprend encore le même *modèle étendu* (2)+(3), mais s'intéresse cette fois à un troisième type de collecte caractérisée par la variable formée par les sommes :

$$Y = \sum_{i=1}^{N} X_i \tag{6}$$

Ce modèle, appelé *loi des fuites* est employé pour les pluies mensuelles du barrage de Ghezala en Tunisie. Dans certains pays en effet, on ne dispose que des seuls relevés des cumuls mensuels de précipitations. À cette échelle de temps, il est souvent acceptable de décrire la précipitation totale mensuelle par une somme de pluies ponctuelles instantanées (et ici inconnues, faute de mesures).

La section 5 compare les angles d'attaque bayésien et classique pour comprendre les structures des modèles à variables latentes et en réaliser l'inférence. Elle conclut sur les perspectives multiples et intéressantes de généralisation de ce type de modèle.

2. Inférence bayésienne de la structure de Poisson marquée à marques exponentielles

2.1. Matériel et modèle

Le tableau 1 décrit la pluviométrie de la période du 19 juillet au 18 août à la station de Bar sur Seine : on a enregistré, pour ces 31 jours de l'année, le nombre d'averses et la quantité d'eau (en mm) de chacune de ces averses, pour 27 années allant de 1975 à 2001.

La figure 3 montre que, sur la période choisie, les hypothèses du modèle à la base des équations (2)+(3)+(4) semblent réalistes : l'histogramme du nombre d'averses est proche d'une répartition de Poisson, celui des dates de début d'évènement est assez uniforme, et on peut donc imaginer que les dates de précipitations forment un processus de Poisson; enfin l'histogramme des pluies ne s'oppose pas à l'idée que les averses puissent être des variables aléatoires exponentielles iid. Notons $j = 1..m$ les différentes répétititions ($m = 27$ années de mesures) de l'expérience aléatoire, n_j le nombre d'averses de l'année j et X_{ij} la $i^{ème}$ averse de la $j^{ème}$ année au cours de la période (à comportement pluvieux supposé homogène) choisie pour l'étude (du 19 juillet au 18 août). Dans ce qui suit, on a changé d'échelle en prenant $T = 1$, de telle sorte que le paramètre μ s'interprète comme un nombre mensuel d'événements. La vraisemblance du modèle fait apparaître les statistiques exhaustives, nombre moyen d'événements $\bar{n} = \frac{1}{m}\sum_{j=1}^{m} n_j$ et précipitation moyenne $\bar{x} = \frac{\sum\sum_{i=1,j=1}^{i=n_j,j=m} x_{ij}}{\sum_{j=1}^{m} n_j}$

$$[n_1, x_{11}, ...n_j, ...x_{ij}, ...n_m, ...x_{mn_m} \,|\mu, \rho] = [z_1, ..z_m \,|\mu, \rho]$$

$$[z_1, ..z_m \,|\mu, \rho] = \frac{\mu^{\bar{n}m}\rho^{\bar{n}m}\exp\left(-\mu m - \rho\bar{n}m\bar{x}\right)}{n_1!n_2!...n_j!..n_m!}$$

TABLEAU 1
Quantité de pluie tombée au cours d'averses à la station de Bar sur Seine entre le 19 juillet et le 18 août pour les années 1975 à 2001

date	pluie tombée (en mm) par averse						
1975	20.2	1.0	12.0	34.5	1.0	25.1	
1976	77.1						
1977	2.4	24.4	17.3	89.4			
1978	1.2	5.0	2.7	1.4	16.6	0.9	
1979	14.7	14.65	39.8				
1980	4.5	0.9	4.6	38.5			
1981	2.2	19.3	6.4	0.9	6.6	53.9	
1982	1.0	4.8	12.9	0.8			
1983	4.2	5.3	13.5				
1984	13.1	5.4	6.2	4.8			
1985	2.2	30.4	2.0	6.6	19.0		
1986	2.5	12.0	14.2				
1987	0.6	8.7	1.0	8.1	3.5		
1988	31.4	0.8	0.5				
1989	16.0	1.1	11.8	9.3			
1990	7.4	4.2					
1991	0.9	24.7	9.2	14.0			
1992	1.4	1.0	47.0	5.1	16.1		
1993	20.0	0.7	20.0	7.8			
1994	11.3	15.0	4.7	1.5	14.3	34.0	33.8
1995	13.3	17.0	23.0				
1996	1.2	55.0					
1997	12.3	6.1	1.4	6.9	79.0		
1998	9.4	9.1	0.5	4.8	2.1		
1999	1.0	16.5	21.1	5.3	8.0	6.0	
2000	23.3	67.3	13.0	7.0			
2001	25.2	3.4	6.4	39.5	21.0		

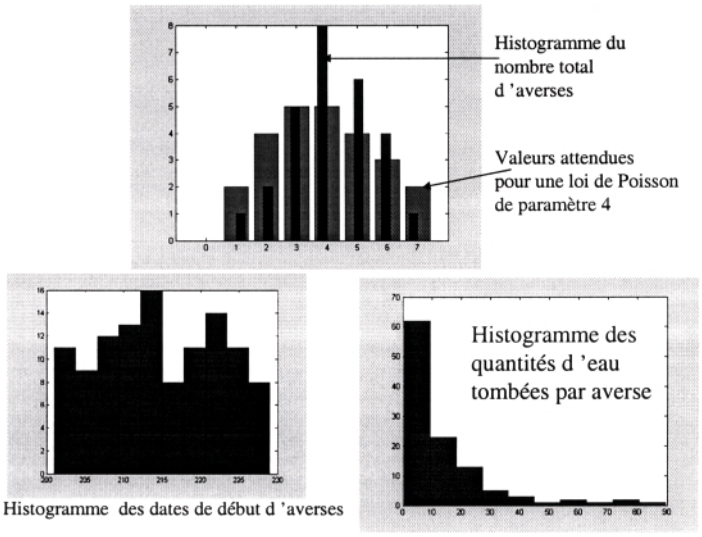

FIGURE 3

Vérification empirique des hypothèses du modèle (2)+(3)+(4)

2.2. Inférence bayésienne explicite par conjugaison Gamma-exponentiel et Gamma-Poisson

L'inférence bayésienne est particulièrement simple si l'on choisit des distributions *a priori* conjuguées gamma à la fois pour μ, paramètre caractérisant le nombre moyen d'événements et pour ρ, le paramètre réglant l'inverse de leur intensité moyenne (avec indépendance des *a priori*). Appelons $gampdf\,(\theta; a, b)$, la distribution gamma d'hyperparamètres (a, b) pour la variable aléatoire θ

$$gampdf\,(\theta; a, b) = \frac{b^a}{\Gamma\,(a)}\theta^{a-1}e^{-b\theta}$$

Dans ce cadre de conjugaison, la formule de Bayes fait passer de la distribution *a priori*

$$[\mu, \rho\,|a_\mu, b_\mu, a_\rho, b_\rho] = gampdf\,(\mu; a_\mu, b_\mu) \times gampdf\,(\rho; a_\rho, b_\rho)$$

à la distribution *a posteriori,* encore explicite :

$$\begin{aligned}[\mu, \rho\,|a_\mu, b_\mu, a_\rho, b_\rho, z_1, ..z_m] &= gampdf\,(\mu; \bar{n}m + a_\mu, m + b_\mu) \\ &\times gampdf\,(\rho; \bar{n}m + a_\rho, \bar{n}m\bar{x} + b_\rho)\end{aligned} \tag{7}$$

On remarque qu'*a posteriori* μ et ρ demeurent indépendants.

Le tableau 2 extrait quelques statistiques à partir de la loi *a posteriori* : espérances, écarts-type et intervalles de crédibilité pour les paramètres μ et ρ peuvent être obtenus grâce à la forme «gamma» de l'équation 7, tandis que les mêmes caractéristiques pour une variable prédictive, l'écart entre la plus grande et la plus petite pluie sur la période allant du 19 Juillet au 18 août, ont été évaluées par simulation.

TABLEAU 2
Résultats de l'inférence du modèle d'averses à la station de Bar sur Seine entre le 19 juillet et le 18 août pour les années 1975 à 2001

statistiques *a posteriori*	moyenne	écart-type	quantile 10%	médiane	quantile 90%
ρ	0.07	0.01	0.06	0.07	0.08
μ	4.19	0.40	3.69	4.17	4.69
écart	25.92	17.69	6.52	22.66	49.31

3. Modèle d'extrême du domaine de Gumbel

3.1. Matériel et génèse du modèle étendu

TABLEAU 3
Débit Maximum annuel à la station de Bar sur Seine pour les années 1950 à 1980 (en dixièmes de mm sur le bassin-versant)

1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960
360	384	509	251	240	1163	269	550	579	642	369
1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971
498	553	159	147	513	561	330	697	305	768	162
1972	1973	1974	1975	1976	1977	1978	1979	1980		
184	446	347	432	413	635	675	553	587		

Le tableau 3 donne pour les 31 années de la période 1950-1980 le maximum annuel des débits journaliers de la Seine, ramené en dixièmes de mm sur le bassin-versant. Même si on ne fournit pas ici la série complète de tous les débits, rien ne nous empêche de définir une crue comme un maximum de débit journalier (à condition que celui-ci dépasse 147 dizièmes de mm, la plus faible valeur de notre série, atteinte par le maximum des débits pour l'année 1964); les hydrologues considèrent de plus que, compte tenu de la vitesse de réaction du bassin-versant, un épisode de crue doit être séparé de son voisin d'au moins dix jours, sinon ils constituent le même évènement. Un modèle réaliste de crues, celui des dépassements, encore appelé *POT* pour *Peaks*

over Threshold (Wang, 1991) peut se construire sur la base des équations (2)+(3). Considérons un ensemble de variables aléatoires indépendantes $Q_1, Q_2 ... Q_n$ ayant en commun la même fonction de répartition G et supposons que les Q_i se produisent à intervalle de temps régulier. Appelons u un niveau seuil et étudions la loi des dépassements au-delà de ce seuil. Sous des conditions techniques de régularité de G, le théorème de Pickands (Coles, 2001) stipule que lorsque le seuil u croît, on peut caractériser à la fois l'intensité et la fréquence des dépassements :

– le nombre de dépassements par unité de temps T suit une loi de Poisson de paramètre μT (puisque c'est ici un cas limite de tirage binomial de paramètre $\mu = 1 - G(u)$)
– l'intensité du dépassement quand il se produit (sachant $Q > u$) obéit à la loi de *Pareto généralisée* à deux paramètres ξ_u et σ_u de fonction de répartition :

$$[Q < q \,|Q > u] = 1 - \left(1 + \xi_u \left(\frac{q-u}{\sigma_u}\right)\right)^{-\frac{1}{\xi_u}} \tag{8}$$

On utilise souvent ce résultat théorique comme modèle pour décrire les températures d'une saison supérieures à un seuil ou les débits d'une rivière dépassant un niveau de référence (Parent et Bernier, 2003). On peut fortement justifier les hypothèses du modèle de façon pragmatique : pour peu que l'on travaille avec un seuil suffisamment élevé et que l'hypothèse d'indépendance soit acceptable pour les crues de ce niveau, les conditions asymptotiques s'appliquent et entraînent la validité progressive de la représentation mathématique (8). Dans l'exemple de la Seine que nous traitons ici, il a été simplifié pour les besoins du calcul (tout en restant réaliste pour nombre de cas) en posant $\xi = 0$ auquel cas l'équation (8) devient par continuité la loi exponentielle (en passant à la variable intensité du dépassement en posant $X = Q - u$ avec $\rho = \frac{1}{\sigma}$ et $u = 147$ dixièmes de mm). Dans cette section, ρ s'interprétera comme l'inverse de l'espérance du dépassement.

Si on disposait de tous les dépassements (événements de crues), l'inférence se conduirait comme à la section précédente (sur la base des équations (2)+(3)+(4)). Il est pourtant courant de ne disposer pour toute information que du maximum annuel $Y = \max_i(X_i)$ et nous partons sur ces bases pour poursuivre l'étude des crues de la Seine.

En conditionnant sur le nombre d'événements qui se produisent sur une année, puis en intégrant sur cette variable aléatoire de Poisson, il est facile de montrer que la loi du maximum annuel issu des marques Pareto- Généralisé (eq (8)) d'un processus de Poisson est une loi généralisée des extrêmes, et qu'en particulier pour une marque exponentielle, la loi de Y est une loi de Gumbel. Néanmoins, nous n'effectuons pas cette opération mentale : la section qui suit laisse encore un peu survivre les variables latentes et montre leur utilité pour l'inférence bayésienne.

3.2. Inférence bayésienne

L'inférence bayésienne consiste à reconstruire les données manquantes N_j et les X_{ij} à partir des conditionnements probabilistes sur les données Y. Travaillons

pour une année donnée j. Y_j étant le maximum des n_j v.a X_{ij}, toutes de même fonction de répartition $F(x) = 1 - e^{-\rho x}$, on peut écrire

$$[Y_j = y \,|n_j, \rho, \mu] = n_j \rho e^{-\rho y} \left(1 - e^{-\rho y}\right)^{n_j - 1} \quad \text{avec } \; n_j > 0$$

Or $[N_j = n \,|\rho, \mu] = \frac{\mu^n e^{-\mu}}{n!}$ donc la loi conjointe de N_j, Y_j s'écrit, en densité :

$$[Y_j = y, N_j = n \,|\mu, \rho] = \frac{1}{(n-1)!} \left(\mu \left(1 - e^{-\rho y}\right)\right)^{n-1} \mu \rho e^{-\rho y - \mu}$$

Sachant Y_j, N_j ne peut pas prendre la valeur 0 (il s'est produit au moins un événement X_{ij} puisqu'on donne Y_j). Si on pose $N'_j = N_j - 1$, on *lit* sur l'équation conjointe (en focalisant son attention sur les seuls termes fonction de $n - 1$) que la loi conditionnelle de N'_j est proportionnelle à $\frac{\left(\mu\left(1-e^{-\rho y}\right)\right)^{n-1}}{(n-1)!}$, c'est à dire que l'on reconnait une loi de Poisson de paramètre $\mu\left(1 - e^{-\rho y}\right)$:

$$\left[N'_j = n - 1 \,|\mu, \rho, y_j\right] = \frac{e^{-\mu\left(1-e^{-\rho y_j}\right)}}{(n-1)!} \left(\mu \left(1 - e^{-\rho y_j}\right)\right)^{n-1} \tag{9}$$

Sachant Y_j et n_j, les $(n_j - 1)$ grandeurs X_{ij} différentes de Y_j sont des variables aléatoires iid exponentielles tronquées par y_j :

$$\begin{aligned}&\left[X_{(1)j}, X_{(2)j}, ...X_{(n_j-1)j} \,|\mu, \rho, y_j, n_j\right] \\ &\qquad = \left(\frac{\rho}{\left(1 - e^{-\rho y_j}\right)}\right)^{n_j - 1} e^{-\rho\left(\sum_{i=1}^{n_j} x_{ij}\right)} \prod_{ij} 1_{x_{ij} < y_j}\end{aligned} \tag{10}$$

L'algorithme d'inférence bayésienne le plus immédiat est un algorithme de Gibbs : on part de valeurs ρ et μ quelconques (mais dans leur support) puis on effectue un nombre important d'itérations de tirages dans les conditionnelles complètes, chaque itération comprenant trois étapes :

1. générer les N_j dans les lois de Poisson tronquées en 0 données par l'equation (9);

2. générer chacune des séries formées par les $n_j - 1$ v.a X_{ij} dans les lois exponentielles tronquées à droite par y_j données par l'eq (10);

3. générer ρ et μ dans les lois gamma données par les eq (7).

De façon asymptotique avec le nombre d'itérations, les tirages (n_j, x_{ij}, μ, ρ) se répartissent suivant la loi a posteriori $[N_j, X_{ij}, \mu, \rho \,|y]$. En pratique, on réalise une période de chauffe permettant d'atteindre ce régime de comportement asymptotique

et on ne conserve ensuite que la dernière partie des tirages générés par l'algorithme de Gibbs.

La figure 4 montre les lois *a posteriori* de μ et ρ , et la forme de la covariation entre ces grandeurs, car, contrairement au modèle précédent fondé sur les équations (2)+(3)+(4), ρ et μ ne sont plus ici indépendants. La figure 5 montre la loi prédictive de la crue centennale (la valeur de la crue max annuelle associée à une probabilité de 10^{-2} de dépassement) : l'année 1955 où le débit a atteint $1163mm/10$, la plus forte valeur de notre série, ne semble donc pas véritablement exceptionnelle (sous nos hypothèses, une crue centennale possède plus de 9 chances sur 10 de dépasser cette valeur), et la même figure visualise, sachant que le max annuel annuel a atteint $1163mm/10$, les probabilités associées aux données manquantes suivantes :

– la variable latente N (le nombre de crues de l'année 1955),

– l'éventuelle valeur de la seconde crue de l'année.

Le tableau 4 précise les intervalles de crédibilité pour les paramètres, pour le débit centennal et pour l'éventuelle seconde crue de l'année 1955.

TABLEAU 4
Moyennes, écart-types et intervalles de crédibilité pour les paramètres, le débit centennal et pour l'éventuelle seconde crue de l'année 1955

statistiques *a posteriori*	moyenne	écart-type	10%	25%	médiane	75%	90%
ρ	0.0040	0.0006	0.0032	0.0035	0.0040	0.0044	0.0048
μ	2.60	0.56	1.91	2.19	2.55	2.95	3.34
q_{100}	1418	214	1164	1266	1394	1545	1701
crue n°2 1955	425	238	132	238	395	588	779

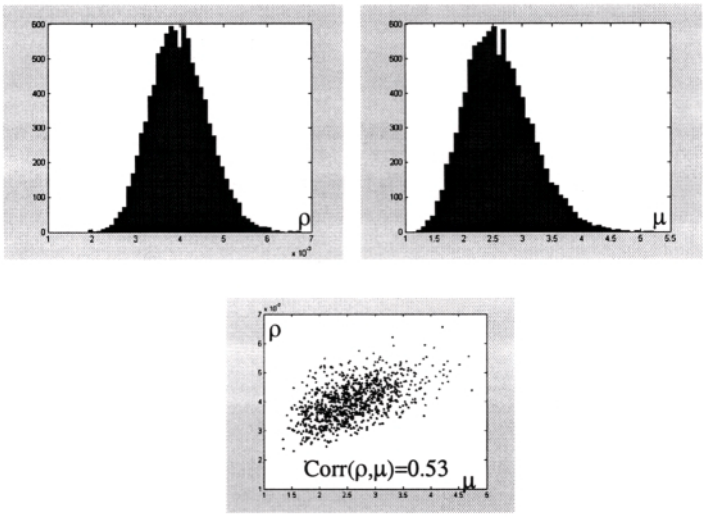

FIGURE 4

Lois a posteriori de μ et ρ , marginales et covariation

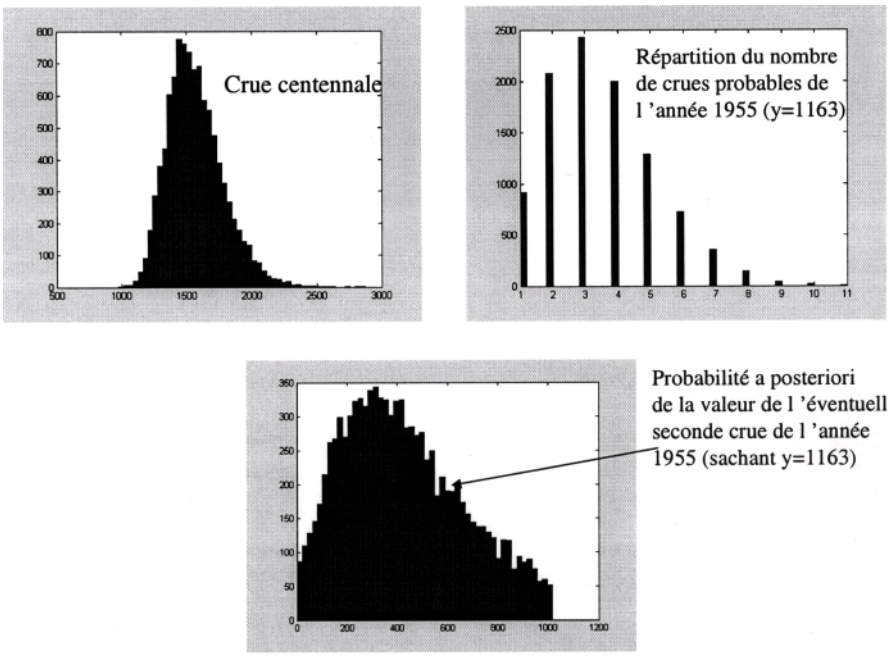

FIGURE 5

Lois a posteriori de la crue centennale, du nombre de crues de l'année 1955, de la valeur de l'éventuelle seconde crue de l'année 1955 (sachant $y = 1163$)

4. Le modèle de la distribution des *fuites*

G. Morlat a baptisé *loi des fuites* le *modèle étendu* des marques exponentielles d'un processus Poisson associé à l'observation spécifique de la somme des marques sur une période donnée. Ce nom vient d'une étude du volume de fuites d'une canalisation de transport de gaz : un processus de Poisson localisait des trous sur le tuyau et chaque fuite locale donnait une quantité aléatoire exponentielle de gaz. Elle est également connue des assureurs puisque c'est la loi du total des dommages à indemniser sur une période quand l'occurrence des dommages est supposée Poissonnienne et leur gravité exponentielle. On transpose directement ce modèle vers le monde de l'hydrologie en faisant l'hypothèse d'un nombre Poissonnien d'averses exponentielles *iid*, quant la collecte de données est limitée à la seule observation des «cumuls de pluies mensuels» (Tapsoba, 1997). Ce modèle est particulièrement utile pour générer des mois sans aucune pluie, ce qui est indispensable pour représenter les pluies cumulées à l'échelle mensuelle de stations météorologiques situées dans des régions méditerranéennes ou semi-arides. Nous utilisons ici l'exemple Tunisien des pluies mensuelles au barrage de Ghezala pour lequel ce modèle s'avère réaliste. La station météorologique tunisienne de Ghezara-barrage est située à proximité d'un périmètre d'irrigation dont il importe de connaître la pluviosié naturelle pour estimer les apports d'eaux complémentaires nécessaires. Le tableau 5 donne, en millimètres les totaux mensuels de précipitations pour février et août observés de 1968 à 2001 à cette station. La figure 6 donne l'histogramme des pluies mensuelles des cumuls des mois de février et d'août à la station du barrage de Ghezala en Tunisie pour les années de 1968 à 2001. Le gestionnaire du barrage (et du périmètre d'irrigation adjacent) souhaite s'appuyer sur un modèle, notamment pour connaître la probabilité que le cumul des pluies en août demeure en-dessous d'un seuil critique pour les cultures (10 mm).

TABLEAU 5
Précipitations mensuelles de Février et d'Août au barrage de Ghezala

	1968	1969	1970	1971	1972	1973	1974	1975	1976	1977	1978
Février	55.7	99.9	105.9	186.9	75.6	150.0	145.1	156.6	80.3	23.8	102.2
Août	0	0.6	0	0	1.6	10.2	0	23.6	10.2	29.7	13.2
	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
Février	123.8	65.0	65.5	73.8	26.0	102.6	91	106.1	164.9	82.2	65.8
Août	0	20.5	0	14.1	0	2.4	0	0	0.6	0	0
	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
Février	34.2	136.4	126.0	71.2	79.6	0.7	291.3	87.2	61.3	76.8	50.4	79.0
Août	33.3	7.5	3.4	0	0.1	31.5	8.6	4.3	52.2	1.4	0	0.2

Les X_{ij} apparaissent comme des variables latentes conditionnant les observations $\{Y_1, Y_2, \ldots, Y_j, ...Y_m\}$: Y_j est la somme des n_j v.a X_{ij} , toutes de même fonction de répartition $F(x) = 1 - e^{-\rho x}$. Distinguons le cas $n_j = 0$ qui fournira nécessairement $y_j = 0$ du cas $n_j \neq 0$ pour lequel y_j, somme de n_j exponentielles

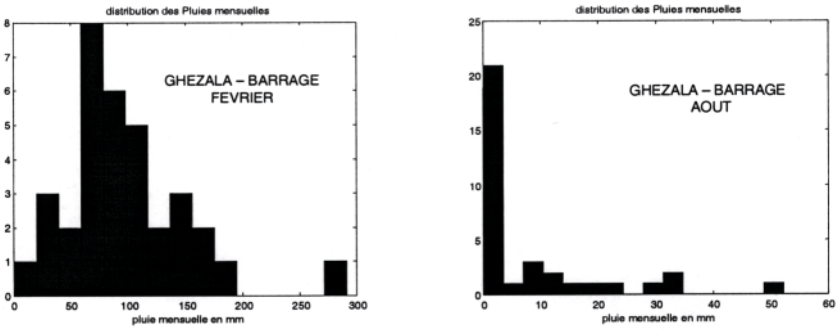

FIGURE 6

Histogrammes des 34 années d'apports mensuels (février et août) à la station météorologique tunisienne du barrage de Ghezala de la valeur de l'éventuelle seconde crue de l'année 1955 (sachant $y = 1163$)

iid est une loi gamma :

$$[Y_j = y \,|n_j = 0, \rho, \mu] = \delta_0(y)$$
$$[Y_j = y \,|n_j \neq 0, \rho, \mu] = \frac{\rho^{n_j}}{\Gamma(n_j)} y^{n_j - 1} e^{-\rho y}$$

où $(y \longmapsto \delta_0(y))$ est la fonction de Dirac : $\delta_0(y)$ égale à 1 si $y = 0$ et à 0 sinon.

Ainsi, compte-tenu de ces propriétés, les seules variables latentes qui importent pour ce modèle, sont les nombres n_j d'événements exponentiels pour l'année j. Avec les notations de l'introduction $Z = \{N_1, ..N_j, ...N_m\}$. Or $[N_j = n \,|\rho, \mu] = \frac{\mu^n e^{-\mu}}{n!}$ et la loi conjointe de N_j, Y_j s'écrit :

$$\begin{aligned}&[Y_j = y, N_j = n \,|\mu, \rho] \\ &\quad = \left(\delta_0(y)\delta_0(n) + (1 - \delta_0(y))\,(1 - \delta_0(n))\,\frac{\rho^n}{\Gamma(n)} y^{n-1} e^{-\rho y}\right) \frac{\mu^n e^{-\mu}}{n!} \qquad (11)\end{aligned}$$

La *vraisemblance étendue* sera le produit de ces lois conjointes pour j parcourant 1 à m.

4.1. Inférence bayésienne

On *lit* sur l'équation conjointe (12) , en focalisant l'attention sur les seuls termes fonction de $n - 1$, la loi conditionnelle de N_j :

$$[N_j = n \,|\mu, \rho, y_j = 0] = \delta_0(n)$$
$$[N_j = n \,|\mu, \rho, Y_j = y\,(\neq 0)] \propto \frac{(\rho\mu y)^n}{n!\,(n-1)!} \qquad (12)$$

L'inférence bayésienne s'appuie encore ici sur un algorithme de Gibbs, itérant, à partir de valeurs ρ et μ quelconques (mais dans leur support), les tirages dans les conditionnelles complètes, faciles à simuler :

1. générer les N_j dans une loi donnée par l'équation (12) si le y_j correspondant est non nul (sinon poser $n_j = 0$ sinon). L'équation (12) est celle d'une distribution à valeurs discrètes (son support est l'ensemble des entiers naturels privé de 0). Elle n'est pas répertoriée dans les distributions «standard» mais elle est aisément simulable (par la méthode d'inversion par exemple).

2. générer ρ et μ dans les lois gamma données par les eq (7).

Si besoin était, on peut facilement inclure dans l'algorithme la regénération des X_{ij} . C'est en effet un résultat classique de la théorie des processus de Poisson que, conditionnellement à n_j et y_j , les variables $U_{1j} = X_{1j}/y_j, U_{2j} = (X_{1j} + X_{2j})/y_j, ...U_{ij} = \left(\sum_{k=1}^{i} X_{kj}\right)/y_j, ...U_{n_{j-1},j}$ sont distribuées comme les $(n_j - 1)$ statistiques d'ordre d'une loi uniforme.

Ces équations ajoutées aux formules (4) permettent la mise en oeuvre de l'algorithme de Gibbs. Les résultats présentés aux figures 7 et 8 ont été obtenus en supposant un prior non informatif sur les paramètres c'est-à-dire : $a_\mu = a_\rho = 0$ et $b_\mu = b_\rho$ très petits.

4.2. Résultats du modèle des «fuites»

L'algorithme de Gibbs a été appliqué avec 4000 itérations dont 2000 de chauffe. La longueur de ces séries de calcul peut paraître courte mais les résultats se sont avérés stables pour cette longueur. En particulier pour le mois d'août (*cf.* figure 7), on trouve que le paramètre μ se situe aux alentours de 1 : cette valeur correspond à 70% de chances environs qu'il ne pleuve pas ou qu'il ne pleuve qu'une fois seulement au cours du mois ($[N = 0\,|\mu] = e^{-\mu} = [N = 1\,|\mu]$), ce qui est conforme à nos attentes compte-tenu du climat semi-aride de la région.

Il est facile de construire une validation non paramètrique bayésienne du modèle des fuites (en créant une structure de probabilités sur l'espace des fonctions de répartitions candidates à représenter le jeu de données, à l'aide du modèle de processus de Dirichlet, initialement formulé par Ferguson, 1973, et appliqué à l'hydrologie par Fortin *et al.*, 1997). Fondée sur la méthode décrite dans (Parent et Bernier, 2004), la figure 8 compare, en situation prédictive, la répartition modélisée par la loi des fuites (en traits pointillés) et la gamme non paramétrique des répartitions de probabilités possibles (en traits pleins). Les ordonnées sont graduées en logarithme décimal de la probabilité au dépassement, et à chaque probabilité le statisticien peut associer un intervalle (horizontal) d'incertitudes :

– L'approche non paramétrique donne pour espérance des modèles possibles la répartition empirique (c'est la ligne centrale en traits continus de la figure 8).

– une fourchette autour de cette valeur moyenne se calcule grâce aux propriétés des processus de Dirichlet. On a ici représenté également en traits pleins les courbes inférieure et supérieure caractérisant les intervalles de crédibilités à 80%.

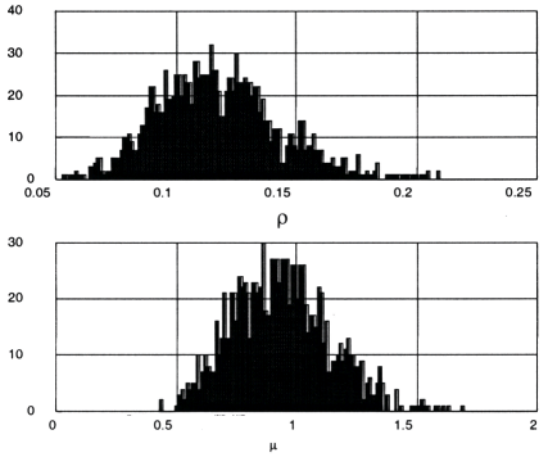

FIGURE 7

Densités a posteriori *marginales des paramètres ρ et μ du modèle de loi des fuites conditionnée sur 34 pluies mensuelles d'Août à la station de Ghezala*

– La courbe en traits pointillés représente la loi *prédictive a posteriori* du modèle à valider (la loi des fuites), c'est-à-dire la probabilité qu'un bayésien accorderait à une nouvelle donnée y_{rep} connaissant le modèle et un échantillon d'apprentissage y :

$$[y_{rep}\,|y] = \int_{\theta} [y_{rep}\,|\theta]\,[\theta\,|y]\,d\theta$$

Elle est calculée en chaque abscisse y_{rep} de la figure 8 en éliminant les incertitudes des paramètres $\theta = (\rho, \mu)$ par intégration de la vraisemblance $[y_{rep}\,|y]$ calculée en y_{rep} et pondérée par la loi *a posteriori* $[\theta\,|y]$. Le jugement à partir du graphe doit se faire conditionnellement aux données y (que l'on a observées et qu'on doit considérer ici comme fixées), c'est-à-dire en prédictif et non en rétrospectif comme en statistique classique.

Retournant au cas d'application, la structure du modèle des fuites offre une bonne réprésentation prédictive des pluies du mois d'août comme de celles du mois de février, alors que les deux répartitions de pluies ont un comportement saisonnier très marqué vis-à-vis de l'occurrence des valeurs nulles. On notera la probabilité prédictive des pluies nulles en août, bien représentée par ce modèle des « fuites » qui ne possède pourtant que deux paramètres d'ajustement. En février, si la pluie de 1996 s'écarte du modèle, l'écart est toutefois dans les limites de crédibilité, étendues à ce niveau de valeurs.

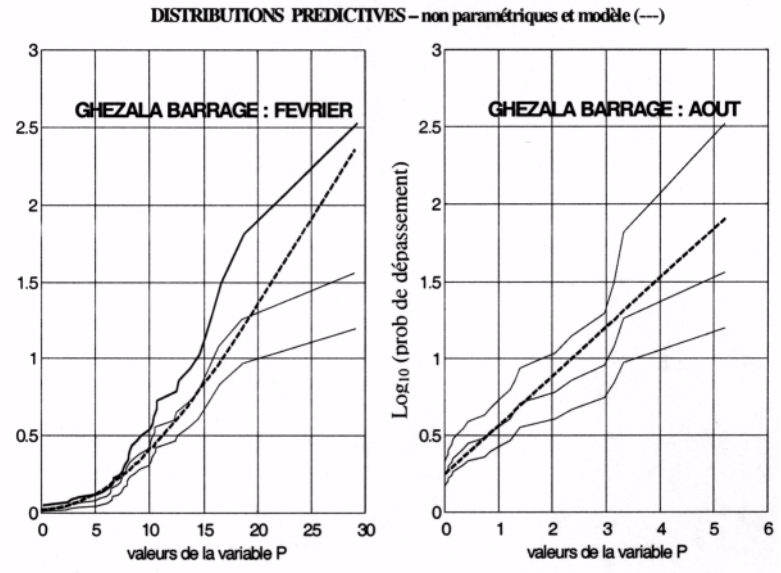

FIGURE 8

Probabilité prédictive modélisée avec intervalle de crédibilité non paramétrique pour les pluies du mois d'août et de février (pluies mensuelles à Ghezala)

5. Un regard sceptique de statisticien

5.1. Le modèle étendu *complètement observé*

En choisissant des lois *a priori* non informatives obtenues par passage à la limite des hyperparamètres $a_\mu, b_\mu, a_\rho, b_\rho$ vers zéro, le bayésien retrouve des résultats proches de ceux de l'estimateur classique du maximum de vraisemblance :

$$\begin{aligned}
E(\mu\,|z_1,..z_m) &= \frac{\bar{n}m+0}{m+0} = \bar{n} \\
Var(\mu\,|z_1,..z_m) &= \frac{\bar{n}m}{m^2} = \frac{\bar{n}}{m} \\
E(\rho\,|z_1,..z_m) &= \left(\frac{\bar{n}m\bar{x}}{\bar{n}m}\right)^{-1} = \frac{1}{\bar{x}} \\
Var(\rho\,|z_1,..z_m) &= \frac{\bar{n}m}{(\bar{n}m\bar{x})^2} = \left(\frac{1}{\bar{x}}\right)^2 \frac{1}{\bar{n}m}
\end{aligned}$$

et il est facile de montrer que les lois *a posteriori* (quoique de type gamma) effacent leur dissymétrie pour tendre vers la loi normale quand n tend vers l'infini. Ce comportement est général (voir par exemple le théorème d'approximation du posterior par une loi normale, Berger, 1985, page 224) : du point de vue asymptotique, bayésiens et fréquentistes marchent ici main dans la main.

5.2. Le modèle de Gumbel

La fonction de répartition du modèle classique de Gumbel, résumant les équations (3)+(5), s'écrit ici :

$$[Y_j < y \,|\mu, \rho] = \exp(-\mu \exp(-\rho y)) \tag{13}$$

D'où s'obtient la vraisemblance pour $j = 1...m$ années d'enregistrements $(y_1, ...y_j, ...y_m)$:

$$[(y_1, ...y_j, ...y_m) \,|\mu, \rho] = (\mu\rho)^n \exp\left(-\rho \sum_{j=1}^{n} y_j\right) \exp\left(-\mu \sum_{j=1}^{n} \exp(-\rho y_j)\right)$$

La vraisemblance n'est pas dans la famille exponentielle, et aucune méthode d'inférence ne s'impose d'emblée. Quoique que les équations du maximum de vraisemblance ne soient pas rebutantes à écrire, compte tenu de la forme explicite de la fonction de répartitition, l'analyste de culture fréquentiste se tournera plutôt vers une technique d'estimation fondée sur la répartition empirique : en effet, après avoir effectué la transformation $y \longmapsto Log(-Log(y))$, les quantiles et les probabilités associées s'alignent sur une droite de pente ρ et d'ordonnée à l'origine $Log(\mu)$. Ici, bayésiens et fréquentistes se renvoient la balle :

- le fréquentiste dispose d'un modèle simple et explicite avec une inférence grâce à la fonction de répartion empirique plus simple que par la technique du max de vraisemblance,
- le bayésien critique ce choix pragmatique au nom d'une certaine rupture de mode de pensée. Le cadre bayésien définit l'inférence sans référence à une méthode, mais par la recherche de la loi *a posteriori*. D'autre part, la simplification donnée par l'intégration explicite de (1) sous la forme (13) lui semble un avatar voire une trahison vis-à-vis *du modèle étendu* : dans certains pays, comme la Suisse, il arrive que les services de navigation ne conservent que les 2 plus grandes valeurs de débits, soit en notant $X_{(i)}$ la $i^{ème}$ statistique d'ordre de $(X_1, ...X_N)$:

$$Y = (X_{(n)}, X_{(n-1)})$$

Si les k plus grandes statistiques d'ordre étaient enregistrées ou, autre possibilité de type d'enregistrements, l'appareil de mesure ne conservait que les événements dépassant un seuil, le fréquentiste changerait sans état d'âme son fusil d'épaule (retour vers le max de vraisemblance) tandis que le bayésien appliquerait la même démarche, à partir des mêmes équations (2)+(3), mais en prenant en compte la nouvelle procédure d'observation (Coles et Powell, 1996).

5.3. Le modèle des fuites

Imaginons un instant que n'ait pas été donné le mode de construction conditionnelle eqs (2)+(3)+(6), mais que passant directement à la vraisemblance par intégration sur les variables latentes (eq (1)), le statisticien ait demandé directement au lecteur de s'intéresser à la variable aléatoire Y dont la distribution de probabilité s'écrit avec deux paramètres (μ et ρ) sous la forme :

$$\begin{cases} [Y = y|\mu, \rho] = \mu\rho\exp(-\mu - \rho y)\frac{I_1(2\sqrt{\mu\rho y})}{\sqrt{\mu\rho y}} \quad \text{si } y > 0 \\ [Y = 0|\mu, \rho] = \exp(-\mu) \ \text{ si } y = 0 \end{cases} \tag{14}$$

Cette expression où apparaît la fonction de Bessel modifiée de seconde espèce I_1 n'est pas standard et entraînera chez tout individu (hormis les spécialistes des fonctions de Bessel) un mouvement immédiat de recul ! Certains lecteurs intrépides et courageux se seraient efforcés, le temps d'un paragraphe ou deux, de surmonter cette légitime répulsion : auraient ils considéré le modèle (14) comme un bon modèle ? On peut toujours plaider que nombre d'arguments vont dans ce sens :

- Le modèle est parcimonieux (2 paramètres μ et ρ seulement). Il s'adapte à la fois à la situation unimodale de gauche de la figure 6 et à la situation de droite (pour créer une masse de probabilité importante en zéro, prendre μ grand)
- Le modèle des fuites est connu, sa fonction caractéristique est explicite et simple, et des applications existent (voir par exemple, Bernier et Fandeux, 1970) pour des applications à l'hydrologie avec une estimation réalisée dans un cadre fréquentiste classique).
- Sans difficulté calculatoire insurmontable, les paramètres μ et ρ peuvent être estimés en maximisant la vraisemblance à partir de l'expression (14). Le logiciel (INRS-Eau, 2000) réalise par exemple cette estimation, mais d'autres approches sont possibles. Les deux premiers moments fournissent un système d'équations facile à exploiter : $E(Y) = \mu/\rho$; $Var(Y) = 2\mu/\rho^2$, mais s'il existe des valeurs nulles, des estimateurs intéressants peuvent également être construits à partir de la fréquence de ces valeurs et de la moyenne des valeurs non nulles $[Y = 0|\mu, \rho] = \exp(-\mu)$; $E(Y|Y \neq 0) = \mu/\left((1 - \exp(-\mu))\rho\right)$
- Il répond à la question d'ingéniérie posée (il évalue à 70 % de (mal)chance environ qu'il pleuve moins de 10 mm en août)... et il s'adapte remarquablement aux données, notamment à la possibilité de pluie nulle en août, spécifique de cette région semi-aride.

Pourtant loin s'en faut que ce modèle (14) séduise au premier abord l'hydrologue normalement constitué. L'apparente complication de sa densité de forme inhabituelle est un repoussoir. La parcimonie du modèle (14) et les autres qualités énumérées précedemment ne suffisent pas : malgré le faible nombre de données du tableau 5, il y a fort à parier que l'analyste préferrera un modèle de mélange à trois paramètres (π, α, β) d'une masse de Dirac et d'une loi gamma :

$$[X = x|\alpha, \beta, \pi] = (1 - \pi)\delta_0(x) + \pi gampdf(x; \alpha, \beta)$$

car chacun de ces coefficients lui semble plus *directement interprétable* : π caractérise la probabilité de l'occurrence de pluies (non nulles) dont l'intensité dans ce cas suit alors une loi gamma de moyenne $\alpha\beta$ a et de variance $\alpha\beta^2$. Bref, pour être satisfait d'une représentation formelle, un modélisateur doit pouvoir comprendre et donner à interpréter son modèle. La meilleure preuve est que, si on lui avait tenu le discours précédent sans fournir la clé de construction conditionnelle, nombre de lecteurs se seraient vraiment sentis pris à rebrousse-poil. La parcimonie, la commodité et la faculté de simuler des séries synthétiques statistiquement semblables aux données initiales seraient-elles des propriétés insuffisantes pour convaincre un analyste que tel ou tel modèle est un « bon » modèle ?

La cohérence probabiliste du cadre bayésien « récupère » les avantages procurés par les variables latentes, ici pivot du raisonnement conceptuel. Une conception fréquentiste trop étroite, qui bondirait trop rapidement sur la vraisemblance (1), peut brider la portée d'une structure de modèle comme l'illustre la figure 9, qui met côte à côte le DAG du modèle (14) et celui du même modèle où se révéle la structure interne (2)+(3)+(6).

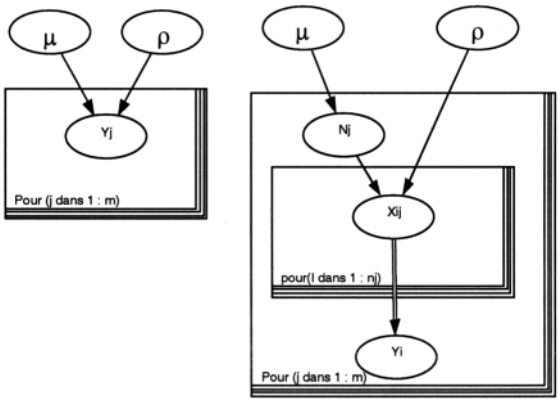

FIGURE 9

Modèle bayésien versus *modèle classique*

5.4. Quel sens conceptuel pour les variables latentes ?

Encore ne faut-il pas aller trop loin dans la confiance que l'on accorde aux variables latentes : ce sont des outils de soutien au raisonnement et un modèle conceptuel ne doit pas s'interpréter au pied de la lettre. Il décrit le fonctionnement d'un système à l'aide d'opérateurs stochastiques, en simplifiant tout en gardant les traits essentiels. À titre de mise en garde, nous avons repris l'étude du modèle (2)+(3), avec toutes les données de crues de la Seine dépassant 147, pas seulement les maximum : la figure 10 indique que les estimations sont plus précises car les écart-types des paramètres μ et ρ sont plus petits. Certains résultats conforteraient « l'existence » de

la variable latente : ainsi, en 1955 sous le modèle de Gumbel, sachant que le maximum avait atteint 1163, le nombre le plus probable de crues était de 3 (figure 5), avec 50% de chance que la seconde crue soit entre 238 et 588 dixièmes de mm (figure 5 et table 4). Il y a eu trois crues en 1955 et la seconde d'importance, le 11 février 1955, a atteint 583 dixièmes de mm ! D'autres résultats incitent à la prudence, à juste titre car les variables latentes ne sont que des constructions intellectuelles, dépendantes des paramètres inconnus du modèle : la figure 10 montre que le modèle de Gumbel a tendance à fortement sous-estimer les paramètres μ et ρ : autrement dit, il n'a pas généré suffisamment de crues, et il a tendance à accorder à ces variables latentes une bien trop forte intensité...

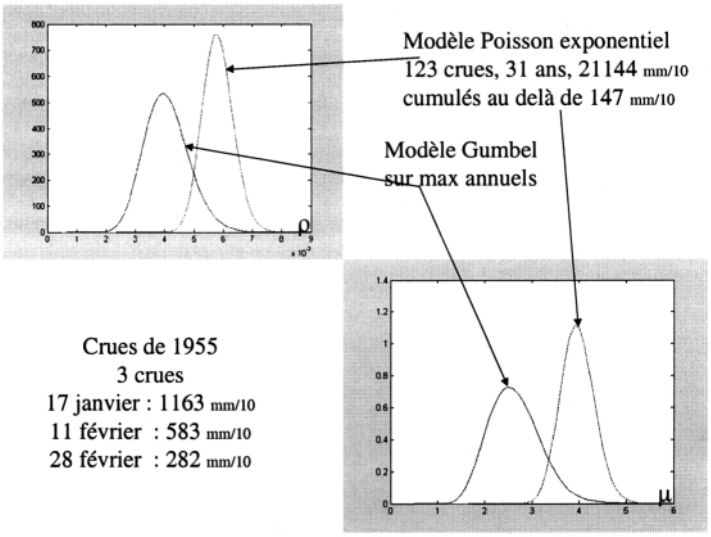

FIGURE 10

Comparaison des estimations d'un modèle de Gumbel sur les maxima annuels avec celles d'un modèle Poisson à marques exponentielles sur les dépassements pour les données de crues de la Seine à Bar/Seine (1950-1980)

Quoique séduitantes, de telles interprétations sont à prendre avec beaucoup de précautions. En toute rigueur, les variables latentes n'ont pas la valeur conceptuelle des grandeurs d'un modèle général. Bien que cohérentes avec le modèle et souvent parées de la beauté du diable car c'est l'analyste lui-même qui les fait exister en les baptisant d'un nom qu'il peut choisir évocateur, elles n'ont de sens que par rapport à ce modèle et sa finalité : par exemple servir à représenter les seuls maxima annuels. Il est donc délicat d'extrapoler leur interprétation vers une autre finalité : par exemple pour tenter d'interpréter des observations de répétition de crues dans l'année ou des fréquences de seconds maxima de chaque année, toutes observations non prises en compte dans l'estimation.

Enfin, on aura également noté les écarts entre les résultats du modèle Poisson-exponentiel et ceux du modèle Gumbel. Il n'est pas douteux que le choix du seuil u, de même que le respect des hypothèses d'indépendance, jouent ici un rôle essentiel. À cet égard la valeur u =147mm/10 est peut-être trop basse pour la validité du comportement asymptotique des dépassements (indépendance et distribution exponentielle des intensités).

5.5. Vers d'autres modèles

Le raisonnement conditionnel avec explicitation de variables latentes permet d'adapter de nombreuses généralisations construites à partir d'occurrences Poissonniennes de marques exponentielles.

5.5.1. Le modèle des fuites

On peut vouloir éviter les désaisonnalisations empiriques approximatives par découpage du temps. Une solution est d'imaginer que le processus des perturbations génératrices devienne un processus de Poisson non homogène avec une intensité développée en série de Fourrier, comme, par exemple, un modèle à 3 paramètres (a, b, c) :

$$\mu(t) = c\left(\exp\left(a\cos(\varpi t) + b\sin(\varpi t)\right)\right)$$

où ω est calculé sur la période annuelle

L'hypothèse de marque exponentielle est souvent réaliste et parcimonieuse, mais on peut complexifier le modèle en adoptant une autre loi. On pense en premier lieu à la loi gamma qui étend naturellement le modèle exponentiel, mais également à la Pareto généralisée (8) justifiée par des considérations asymptotiques quand on travaille dans le domaine des dépassements.

On peut enfin mettre en doute l'indépendance entre l'intensité des marques successives. Il faut alors imaginer des modèles plus complexes, par exemple de type bivarié exponentiel entre marques successives.

5.5.2. Neyman – Scott

Plusieurs modèles ont été proposés dans la litérature hydrologique pour décrire le processus des précipitations à échelle fine (Bernier, 1970; Maidment, 1993). Nous présentons un de ceux-ci, le modèle de Neyman Scott en Grappes (Favre, 2001), en faisant ressortir les variables latentes.

Le modèle de Neyman Scott en Grappes fait intervenir une hiérarchie de plusieurs processus :

- un processus de Poisson censé représenter les origines τ_i de perturbations atmosphériques (taux constant μ),
- chaque perturbation donne naissance à une «grappe» en nombre aléatoire K d'averses successives dans le temps survenant à des dates aléatoires :

– le nombre K est distribué selon la loi géométrique :

$$[k\,|\alpha] = \alpha\,(1-\alpha)^{k-1} \quad \text{pour} \;\; k = 1, 2, \ldots$$

les dates u_j de ces averses, comptées depuis «l'origine τ_i de chaque perturbation» sont indépendantes et distribuées exponentiellement :

$$[u\,|\beta] = \beta \exp(-\beta u) \quad \text{pour } u \geqslant 0$$

– chaque averse a une durée aléatoire v , indépendante des autres, également exponentielle :

$$[v\,|\eta] = \eta \exp(-\eta v) \quad \text{pour } v \geqslant 0$$

– chaque averse a une intensité X, constante sur toute sa durée et distribuée selon une exponentielle.

$$[x\,|\rho] = \rho \exp(-\rho x) \quad \text{pour } x \geqslant 0$$

Ce modèle se distingue par une certaine complexité qui en a rendu assez délicates les applications par les méthodes de statistique classique mises en oeuvre par les hydrologues. Si les étapes de conditionnement sont nombreuses, elles sont généralement simples. À ce jour pourtant, les chercheurs du domaine ne semblent pas avoir reconnu que sa structure hiérarchique puisse faire une utilisation fructueuse du raisonnement conditionnel pour en réaliser l'inférence.

5.5.3. Algorithmes d'inférence

Les exemples présentés dans cet article bénéficiaient des commodités du calcul conditionnel (algorithme de Gibbs). Mais cette facilité n'est pas essentielle. L'homme d'études peut vouloir adopter d'autres loi *a priori* que des gammas indépendantes sur μ et ρ. La boite à outils de calcul bayésien n'est pas dégarnie de techniques d'inférence alternative (approximations, échantillonnage pondéré, méthodes MCMC, ...). Les nombreuses généralisations du modèle initial entraîneront certes le recours à d'autres techniques d'inférence que l'algorithme de Gibbs (Kuczera et Parent, 1998). Néanmoins ces algorithmes de calcul bayésien reconstruiront tous, à un moment ou à un autre des étapes de l'inférence, les variables latentes.

6. Conclusions

Être bayésien ou ne pas être bayésien, finalement telle n'est plus la question, même pour un environnementaliste (Clark, 2005; Bernier *et al*, 2000). Il faut plutôt insister sur les fruits du raisonnement probabiliste conditionnel : conditionnement direct utile pour la modélisation, conditionnement inverse lors des calculs d'inférence par simulation Monte Carlo utile au bayésien (tirage MCMC de la loi *a posteriori*) ou au classique (algorithme EM).

6.1. Sens direct : la modélisation

Le paradigme bayésien favorise naturellement la construction d'un modèle par étages successifs de conditionnements probabilistes. Cette approche abolit la différence de nature mathématique entre paramètre et variable observable, et ainsi, la variable latente trouve mieux sa place duale dans l'échafaudage intellectuel d'un modèle statistique : tantôt elle joue le rôle de paramètre quand elle conditionne la naissance d'observations, tantôt elle joue le rôle d'un résultat potentiellement observable quand elle est générée dans les couches internes de la structure du modèle. Pour le modélisateur classique, payer le prix d'une fonction de vraisemblance non explicite peut être largement compensé par la richesse de construction que procure l'introduction de variables latentes. Exprimer ces variables cachées est de plus bien souvent pertinent pour renforcer le sens conceptuel d'un modèle. Les trois exemples pris à la météorologie montrent la puissance du concept de variables latentes sous ses divers aspects : structurel, données manquantes et variables prédictives. Ce concept est ici directement opérationnel : il éclaire la modélisation des phénomènes et en permet le calcul complètement cohérent en limitant les hypothèses de commodité qui président à l'adoption d'un modèle *ad hoc*, par empirisme ou par souci de parcimonie. Les variables latentes sont également sources d'inspiration pour construire des extensions intéressantes du modèle de départ. La même démarche s'applique ainsi à des modélisations complexes (modèle de Neyman - Scott en grappes non stationnaire par exemple, modèles exposés par Cox et Isham, 1994, pour la modélisation de la stochasticité des pluies) et aux diverses formes que peuvent prendre les données disponibles pour l'estimation et la validation.

6.2. Sens inverse : l'inférence

L'inférence bayésienne repose sur un conditionnement probabiliste inverse de celui qui préside à l'élaboration d'un modèle. Là encore le rôle pivot de la variable latente peut grandement faciliter la tâche de l'analyste. Ce conditionnement simplificateur est pleinement exploité pour la construction d'algorithmes MCMC. C'est sans doute la justification de «l'abus de langage» si répandu qui présente les calculs de simulations MCMC (si on en voit que l'aspect calculatoire) comme de véritables méthodes d'estimations statistiques. Quant au point de vue classique, qui prend aujourd'hui fréquemment un chemin très similaire de conditionnement au travers les algorithmes EM stochastiques (Celeux *et al.*, 1996), il bénéficie, grâce aux variables latentes-pivots, d'avantages équivalents à ceux illustrés par cet article.

Remerciements

Les auteurs remercient Vazken Andreassian de la division hydrologie du CEMAGREF pour avoir mis à leur disposition les données de la Seine ainsi que Fethi Lebdi de l'Institut Agronomique de Tunis pour les données du barrage de Ghezala.

Références

BERGER J.O. (1985), *Statistical Decision Theory and Bayesian Analysis*, New-York : Springer-Verlag.

BERNIER J. (1970), Inventaire des modèles de processus stochastiques applicables à la description des débits journaliers des rivières, *Revue de l'institut International de Statistique*, **38**, 49-61.

BERNIER J. et FANDEUX D. (1970), Théorie du renouvellement – Application à l'étude statistique des précipitations mensuelles, *Revue de Statistique Appliquée*, **XVIII**(2), 75-87.

BERNIER J., PARENT E. et BOREUX J.J. (2000), *Statistique pour l'environnement. Traitement bayésien de l'incertitude*, Tec et Doc, Lavoisier.

BROOKS S.P. (2003), Bayesian Computation : A Statistical Revolution, *Transactions of the Royal statistical society, series A*, **15**, 2681-2697.

CELEUX G., CHAUVEAU D. et DIEBOLT J. (1996), Some Stochastic Versions of the EM Algorithm, *Journal of Statistical Computation and Simulation*, **55**, 287-314.

CLARK J. S. (2005), Why environmmental scientists are becoming Bayesians, *Ecology Letters*, **8**(1), 2-14.

COLES J.H. et POWELL E.A. (1996), Bayesian Methods in Extreme Value Modelling : A Review and New Developments, *International Statist. Review*, **U 64**, 119-136.

COLES S.G. (2001), *An Introduction to Statistical Modeling of Extreme Values*, New-York : Springer-Verlag.

COX D.R. et ISHAM V. (1994), Stochastic Models of Precipitation. Pages 3-18 of : V. Barnett et K. Turkman (eds), *Statistics for the Environment 2 : Water Related Issues*, Chichester : Wiley.

FAVRE A.C. (2001), *Single and Multi-Site Modelling of Rainfall Based on the Neyman-Scott Model*, Ph.D. thesis, École Polytechnique Federale de Lausanne.

FERGUSON T. (1973), A Bayesian Analysis of some Non Parametrics Problems. *Annals Statist.*, **1**, 209-230.

FORTIN V., BERNIER J. et BOBÉE J. (1997), Simulation, Bayes and Bootstrap in Statistical Hydrology, *Water Resources Research*, **33**(3), 439-448.

GELMAN A. (2003), A Bayesian Formulation of Exploratory Data Analysis and Goodness of Fit Testing, *International Statistical Review*, **71**, 369-382.

GELMAN A., CARLIN J.B., STERN H.S. et RUBIN D.B. (1995), *Bayesian Data Analysis*, Londres : Chapman and Hall.

INRS-Eau, Chaire Hydrologie Statistique (2000), *Hydrological Frequency Analysis Software Package*. INRS Eau Québec, programme (http ://www.inrs-eau.uquebec.ca/activites/groupes/chaire$_hydrol$).

KUCZERA G. et PARENT E. (1998), Monte Carlo Assessment of Parameter Uncertainty in Conceptual Catchment Models : The Metropolis Algorithm, *Journal of Hydrology*, **211**, 69-85.

MAIDMENT D.R. (1993), *Handbook of Hydrology*, MacGraw-Hill.

PARENT E. et BERNIER J. (2001), Méthodes bayésiennes et modélisation des risques géophysiques extrêmes, *La revue Modulad*, **N° 28**, 1-26.

PARENT E. et BERNIER J. (2003), Bayesian POT Modeling for Historical data, *Journal of hydrology*, **274**, 95-108.

PARENT E. et BERNIER J. (2004), Une procédure bayésienne de sélection/ validation différentielle pour déterminer le domaine d'attraction des valeurs extrêmes, *Revue de Statistique Appliquée*, **LII**(4), 5-31.

ROBERT C.P. et CASELLA G. (1999), *Monte-Carlo Statistical Methods*, Springer.

SNYDER D.L. (1975), *Random Point Processes*, New-York : Wiley.

SPIEGELHALTER D.J., THOMAS A. et BEST N.G. (1996), Computation on Bayesian Graphical Models, *pages 407-425 of :* Bernado J.M., Berger J.O., David A.P. et Smith A.F.M. (eds), *Bayesian Statistics*, Oxford University Press.

TANNER M.A. (1996), *Tools for Statistical Inference : Methods for the Exploration of Posterior Distribution and Likelihood Functions*, New-York : Springer-Verlag.

TAPSOBA D. (1997), *Caractérisation évènementielle des régimes pluviométriques ouest africains et de leur récent changement*, Ph.D. thesis, Université Paris-Sud, Orsay.

WANG Q.G. (1991), The POT Model Described by the Generalized Pareto Distribution with Poisson Arrival Rate, *Journal of Hydrology*, **129**, 263-280.