Statistique Bayésienne des processus ponctuels marqués. Le rôle des variables latentes dans la démarche de modélisation

Jacques BERNIER*, Eric PARENT* et Jean Jacques BOREUX**

* *ENGREF, Laboratoire de Gestion des Risques En Sciences de l'Environnement, 19, Avenue du Maine, F-75732 Paris. Cedex 15,*
Parent@engref.fr, Jacques.Bernier2@wanadoo.fr

** *Fondation Universitaire Luxembougeoise (FUL) 185 Avenue de Longwy , B-6700 Arlon .*

RÉSUMÉ. En statistique bayesienne le raisonnement conditionnel est à la base de la solidarité étroite entre modélisation et calcul par algorithmes MCMC. Un des aspects les plkus fructueux de son utilisation est l'utilisation explicite des variables latentes ou cachées. On présente trois exemples de séries météorologiques représentées par des modèles de processus ponctuels marqués de complexité croissante. Sur ces exemples on montre la souplesse de modélisation et les facilités de calcul apportées par les variables latentes en relation avec les techniques « d'augmentation de données de Tanner »..

ABSTRACT. In this paper we are concerned with the use of bayesian conditional reasoning in the two basic processes of stochastic modelling and computation via MCMC algorithms. We emphasize the use of latent or hidden variables in the analysis of three different marked points models applied to meteorological data. The explicit introduction of such hidden variables together with their use in the « augmentation data algorithms » allows realistic and easy treatment of the problems.

mots clés : Modélisation bayésienne en Environnement, méthodes Markov Chain Monte Carlo ; variables latentes ; modélisation graphique.

Introduction

Les applications statistiques en Environnement utilisent souvent des processus ponctuels marqués: lois de type Neymann Scott pour la modélisation d'évènements orageux, « lois des fuites » pour la représentation des pluies mensuelles par exemple. En pratique, l'approche bayésienne a résolu les difficultés d'inférence par l'emploi des techniques Monte Carlo par Chaîne de Markov (MCMC). L'évaluation des lois a posteriori décrivant les valeurs probables des paramètres s'effectue aujourd'hui sans difficulté même pour des modèles ayant un grand nombre de paramètres ou pour des structures complexes comme les modèles hiérarchiques échangeables. Une conséquence majeure , quoique peu soulignée de l'emploi des MCMC dans un cadre bayésien, c'est que la démarche de construction de modèle et celle du calcul d'inférence s'effectuent dans le même contexte probabiliste : il s'agit d'un simple renversement du conditionnement que décrit le modèle. Cette structure est bien mise en lumière par la présentation sous forme de modèles graphiques : les graphes acycliques orientés ou DAG selon la terminologie de Spiegelhater et al. (1996).
Dans un cadre bayésien, l'introduction explicite de variables latentes favorise à la fois la mise en oeuvre de techniques puissantes d'inférence et révèle aussi la structure profonde des modèles.
Ces variables latentes (ou cachées) sont des variables structurelles inobservables dont l'explicitation facilite ainsi la compréhension du modèle. Mais plus encore elle en permet la mise en œuvre concrète selon le paradigme bayésien. La démarche statistique classique cherche à se débarrasser de ces variables en exprimant au plus

vite la vraisemblance des paramètres vis à vis des seuls observables. Au contraire la démarche bayésienne non seulement les explicite mais de plus les intègre au mode de calcul en chaîne des distributions de probabilité a posteriori des paramètres. Cette intégration est permise grâce aux techniques « d'augmentation des données » de Tanner qui exploitent les constructions conditionnelles parallèles du modèle et de l'algorithme MCMC.

On distingue généralement :

- *les variables latentes structurelles*, proprement dites, décrivant une part des conditionnements internes des modèles tels ceux qui décrivent les processus de Markoff cachés dont le développement est lié à cette approche.

- *les « données manquantes »* représentées techniquement par des variables qui ne se différencient pas des variables latentes. Comme ces dernières, ce sont des « non-observables » liées au modèle et n'existant que par lui. Comme elles aussi, on les intègre au calcul en s'appuyant sur des conditionnements probabilistes semblables.

Il n'existe pas de différences techniques entre variables latentes et données manquantes si on suit Gelman (2003) en généralisant la vraisemblance

$L(y|\theta)$ des paramètres θ vis à vis des observables y sous la forme $L(y, I|\theta)$ où I représente l'information « réelle » résultant d'une politique de collecte particulière. Ici les « y » ont *le* sens d'observables potentiels, variables d'état liées à une **structure de modèle** bien définie. Ces variables sont latentes dans la mesure où elles n'appartiennent pas à I. Toutefois la modélisation doit spécifier le conditionnement par y des variables définissant I.

- Enfin nous citerons pour mémoire ici *Les variables prédictives*, incluses dans y, produits du modèle mais considérées comme non observées pour l'inférence a posteriori. Elles interviennent par leurs distributions prédictives liées aux données dans les validations et l'expression des coûts de décision éventuels. Même si leur utilisation est extérieure à la procédure d'estimation du modèle, on peut tirer avantage de leur intégration directe dans l'algorithme markovien d'estimation comme variable latentes supplémentaires calculées « on line » dans les séquences MCMC..
Dans cet article nous illustrons 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 : les processus ponctuels marqués. Une telle structure contient en germes de nombreuses variables d'états qui peuvent décrire y. Il importe d'en choisir les plus pertinentes pour le conditionnement de I . Nous insisterons également sur la modélisation de cette interaction $y \leftrightarrow I$.
Les exemples illustratifs appartiennent au domaine de la météorologie. Sur un plan phénoménologique il s'agit de la représentation des processus de précipitations atmosphériques cumulées sur différentes échelles de temps au moyen de processus ponctuels marqués. En environnement l'intérêt de ces modèles et de leur traitement bayésien n'est d'ailleurs pas limité au seul domaine de la pluviométrie.

Les modèles stochastiques ponctuels marqués

Nous utiliserons cette structure de modèles sous sa forme la plus simple mais la démarche bayesienne peut s'appliquer aux structures plus complexes (cf Snyder, 1975).
La figure 1 (page ci après) présente la trajectoire théorique d'un processus ponctuel marqué où des événements supposés ponctuels (sans durée) surviennent aléatoirement dans le temps.
A chaque occurrence d'un événement est associée une marque X_i, variable aléatoire que l'on considère au dessus d'un seuil u_0.

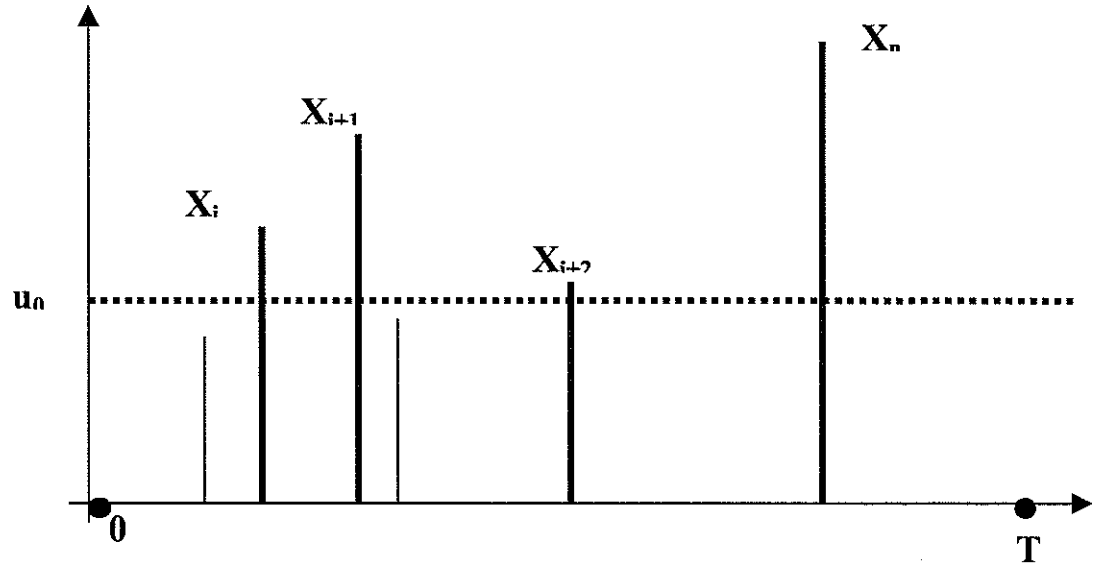


Figure 1 : Un exemple de trajectoire de processus ponctuel marqué

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

- si N_T est le nombre d'événements sur un intervalle fixé *[0,T]* , ce nombre est distribué selon la « loi de Poisson » :

$$[n|\mu,T] = e^{-\mu T}\frac{(\mu T)^n}{n!} \qquad \mathrm{N_T} = 0,1,2,....,\mathrm{n},...$$

- les X_i successifs sont indépendants et leurs dépassements d'un seuil u_0 sont distribués selon la « loi exponentielle » :

$$[x|\rho, X_i \geq u_0] = \rho e^{-\rho(x-u_0)} \quad [x|\rho, X_i \geq u_0] = \rho e^{-\rho(x-u_0)}$$

- pour u_0 fixé, les deux paramètres μ et ρ caractérisent entièrement ce modèle.

Cette structure a de multiples applications dans le domaine de l'environnement concernant notamment les valeurs extrèmes (modèle POT en hydrométéorologie). Dans ce cas on associe à cette structure une collecte ***I*** constituées des valeurs extrèmes dépassant un niveau u_0:

$$Y_T = \underset{[0,T]}{Max}(X_i)$$

Dans la même conférence on trouvera, pour valider ce type de structure et de collecte, une présentation de l'utilisation de diverses variantes de ***I***.que l'on pourrait appeler « information de calage du modèle » (Parent et al., 2003).

On peut s'intéresser ainsi à d'autres types de collecte caractérisées par des variables comme les sommes $P_T = \underset{[0,T]}{Somme}(X_i - u_0) = \sum_{i=1}^{N_T}(X_i - u_0)$

Dans cette communication nous nous intéresserons aux précipitations totales mensuelles considérées comme sommes de pluies ponctuelles instantanées X_i décrites par le processus ci dessus auec T=1 mois (u_0 =0). Si aux très courtes échelles de temps, ce modèle n'est pas réaliste. C'est surtout l'hypothèse de pluie instantanée et non pas la distribution des intensités qui est en cause. Cependant à une échelle de temps comme le mois, la chronologie des pluies aggrégées importe moins. Ceci suggère l'emploi de ce modèle pour les pluies mensuelles.

Inference bayesienne de la structure de poisson marquée

La vraisemblance complète du modèle incorporant toutes les variables « latentes » X_{ij} s'écrit :

$$L(y,\mu,\rho) = \mu^n \rho^n \exp(-\mu N - \rho SS) \qquad (3)$$

avec $y = \{\bigcup X_{ij}, \bigcup n_j\}$ (j est l'indice des années) $\quad n = \sum_j n_j, SS = \sum_{ij} X_{ij} = \sum_j P_j$

Avec cette structure la mise en œuvre de l'algorithme de Gibbs est particulièrement simple :

En utilisant des distributions a priori conjuguées $gamma(a,b) = \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}$ d'hyperparamètres (a,b) différents selon que θ est μ ou ρ, les distributions a posteriori conditionnelles complètes sont explicites :

$$[\mu|\rho,y] = gamma(n+a_\mu, N+b_\mu) \text{ indépendante de } \rho$$
$$[\rho|\mu,y] = gamma(n+a_\rho, SS+b_\rho) \text{ indépendante de } \mu \qquad (4)$$

Du fait de cette indépendance et si les *y* étaient complètement données les méthodes MCMC seraient inutiles. Mais la considération de variables latentes change la donne.

Le modèle de la distribution des « fuites »

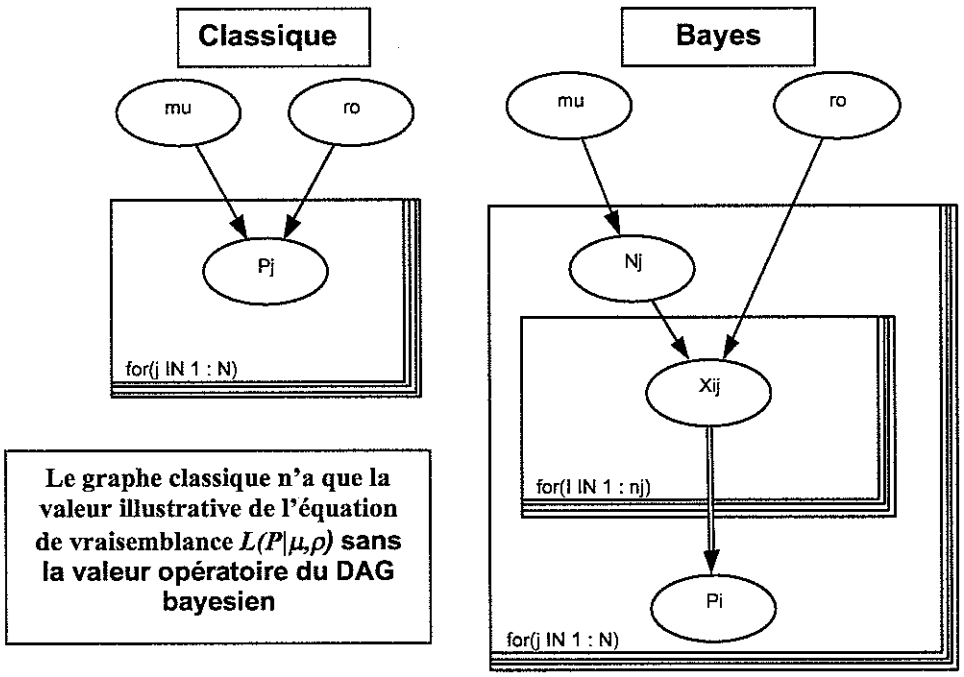
Le premier exemple concerne « la loi des fuites » selon la terminologie de G. Morlat qui a appliqué le modèle à la représentation du volume de fuites sur les canalisations de transport de gaz et que nous conservons ici. Il s'agit de la structure précédente où la collecte de données ***I*** est limitée à l'observation des « pluies mensuelles » c'est à dire les sommes sur un mois calendaire fixé T : $P_j = \sum_{\tau_i \in T^*} X_{ij}$. La distribution des somme s'obtien aisément :

$$[P|\mu,\rho] = \begin{cases} \exp(-\mu-\rho p)\frac{I_1(2P)}{\sqrt{\mu\rho P}} \text{ si } P>0 \\ \exp(-\mu)\delta_0(P) \quad \text{où} \quad \delta_0(P) \text{ est un Dirac en 0. si } P=0 \end{cases} \qquad (5)$$

$I_1(.)$ est la fonction de Bessel modifiée de 1ère espèce.

C'est une densité mixte avec masse finie en P=0 qui est capable de représenter les pluies cumulées à l'échelle mensuelle de stations météorologiques situées dans des régions méditerranéennes ou semi-arides notamment. Nous montrerons des exemples africains où ce modèle s'avère réaliste.

En statistique classique Il y a une difficulté importante issue de la forme non-standard de cette distribution lorsqu'on veut l'estimer directement sur des échantillons de sommes mensuelles $\{P_1, P_2, ...P_n\}$, séries habituellement répertoriées.



Le graphe classique n'a que la valeur illustrative de l'équation de vraisemblance *L(P|μ,ρ)* sans la valeur opératoire du DAG bayesien

Figure 2

En Bayes la structure du modèle complet est éclairée par son graphe acyclique orienté (DAG, figure2) construit selon la logique du conditionnement markovien local.

Les X_{ij} apparaissent clairement comme des variables latentes conditionnant les observations $\{P_1, P_2, ...P_n\}$. Mais on observe que la vraisemblance complète dépend directement de ces

observations puisque $SS=\Sigma P_j$. De fait les seules variables latentes qui y apparaissent sont les n_j. Conditionnellement aux P_j les n_j sont indépendants et leurs distributions conditionnelles complètes sont :

$$[n_j \mid \mu, \rho, P_j] \propto e^{-\mu-\rho P_j} \frac{(\mu\rho P_j)^{n_j}}{n_j!(n_j-1)!} \quad \text{pour } n_j = 1,2,.... \qquad (6)$$

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

Applications de l'inférence bayesienne du modèle des « fuites »

Nos exemples concernent les précipitations mensuelles de la station météorologique d'Atakpamé au Togo. L'étude statistique a été faite systématiquement sur les 12 mois de l'année mais on en présente les résultats pour 2 mois : décembre (saison sèche) et mars (fin de saison sèche). Ces résultats sont illustrés par les graphiques de la page suivante.
On notera les différences de régimes avec des distributions en J comportant une fréquence significative de pluie nulle en Décembre
Le second ensemble de graphiques concerne les densités de probabilité a posteriori des deux paramètres μ et ρ dans chaque cas. Les histogrammes (à gauche) ont été obtenus par simulation Gibbs. Le nombre d'itérations a été systematiquement de 3000 dont 1000 de « chauffe ». Ce choix a eté conforté par la stabilité des résultats sur plusieurs répétitions. Les figures de droite montrent les lissages efficaces obtenus par la technique dite de « Rao-Blackwell » applicable aisément ici.
Ces deux ensembles de graphiques (figure 3 et figure 4) sont situés sur les pages suivantes :
Les figures suivantes présente une méthode graphique de vérification du modèle qui ressemble à une technique graphique classique mais on doit en faire une interprétation bayesienne. Celle ci est donc conditionnelle aux données que l'on doit considérer comme fixées et le jugement est relatif aux probabilités de non dépassement incertaines que le statisticien peut lui associer.

- la courbe en traits plein est la « valeur moyenne prédictive » des probabilités de non dépassement des pluies mensuelles du niveau observé pour une année future. Cette valeur moyenne intègre complétement les incertitudes du modèle des fuites. Les croix représentent cette valeur moyenne des probabilités en ne faisant que l'hypothèse dite « non paramètrique » sur la distribution des pluies c'est à dire en supposant une distribution quelconque. Le support de ces probabilités est certes la gamme des pluies observées mais celles ci ne sont là que pour la visualisation graphique de valeurs futures possibles. Les courbes doivent donc être interprétées en « prédictif » et non en « retrospectif » comme en statistique classique où un modèle est jugé directement sur les données observées.
- Cette valeur moyenne est assortie d'une « fourchette : courbes inférieure et supérieure » en tiretés caractérisant les intervalles de crédibilités à 80% non paramétrique associés à la moyenne prédictive non paramétrique.
- Les résultats non paramétriques sont les fruits de l'approche bayesienne non paramétrique (modèle de processus Dirichlet de Ferguson, 1973).

Ces graphiques permettent un jugement global heuristique utile qui nous semble même supérieur au jugement retrospectif d'ajustement classique des courbes de fréquence parce qu'on se place d'emblée dans la position prédictive de l'utilisation du modèle.
L'approche bayesienne propose d'ailleurs des méthodes de test et de choix de modèles peut être plus précises et efficaces mais beaucoup plus complexes qu'on ne saurait présenter ici.(voir la présentation de Parent et al., 2003 dans les mêmes journées).
Cette validation graphique fournit donc des indications heuristiques intéressantes sur la valeur du modèle des fuites. Elle permet notamment de s'assurer que ce modèle semble particulièrement réalste pour les mois où la fréquence des pluies nulles est notable. Dans notre cas il semble que pour l'autre mois (mars) des écarts peuvent apparaître pour les fortes précipitations notamment ; l'écart est toutefois le fait des valeurs les plus extrèmes. Nous y reviendrons plus loin.

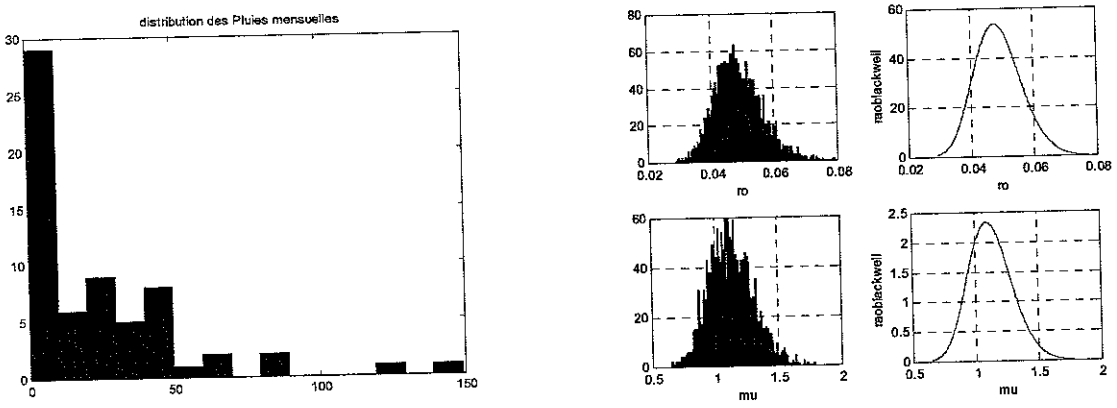
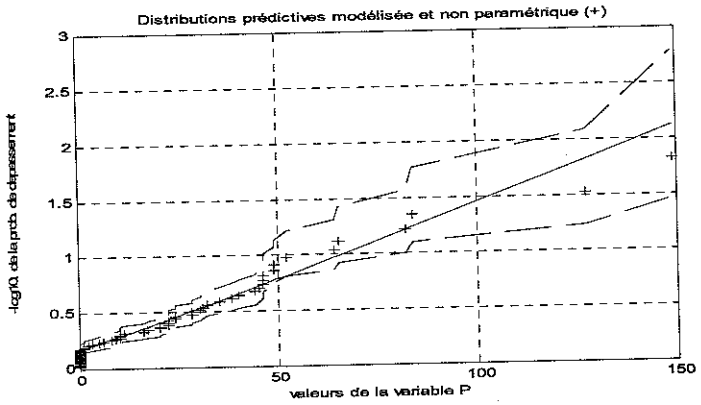
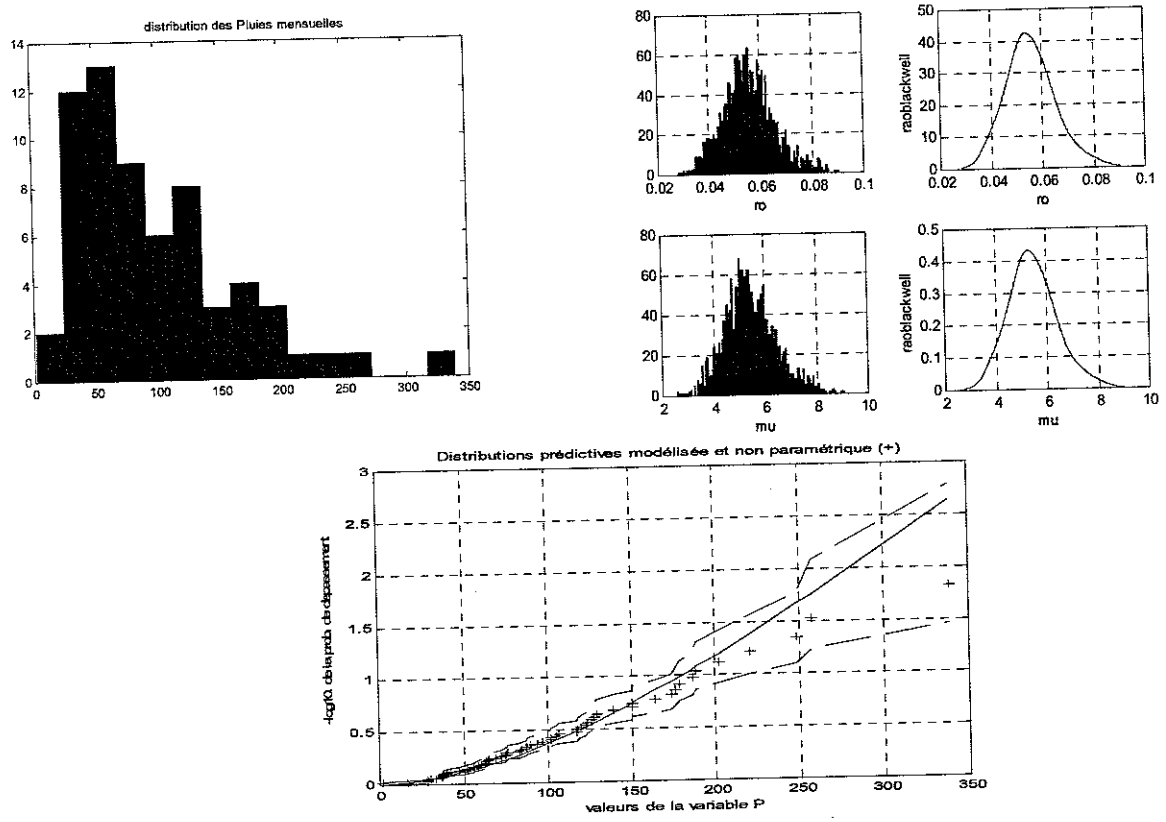


Figure 3 : Atakpamé : Histogramme des pluies de Décembre et Densités a posteriori de μ et ρ



Figue 4 : Atakpamé : Comparaison prédictive des probabilités de Décembre



Figures 5 et 6 : Atakpamé : Histogramme des pluies de Mars, Densités a posteriori de μ et ρ et comparaison prédictive des probabilités

Paramètres et variables supplémentaires

Nous avons déjà souligné la grande souplesse de l'outil combiné « modélisation bayesienne – calcul MCMC » pour l'estimation des distributions de multiples variables prédictives, intégrées comme variables latentes supplémentaires. Notons ici une remarque evidente mais utile : s'il s'agit de paramètres du modèle ou de fonctions logique d'entre eux, leur distribution prédictive n'est autre que leur distribution a posteriori puisque leur construction n'est pas liée à un alea supplémentaire quelconque. A titre d'exemples nous montrons deux figures déduites directement de l'algorithme de Gibbs, et donnant pour une gamme de valeurs de la durée de retour (en log. décimaux) les estimations médianes et les limites des intervalles de crédibilités à 90% des quantiles correspondants.

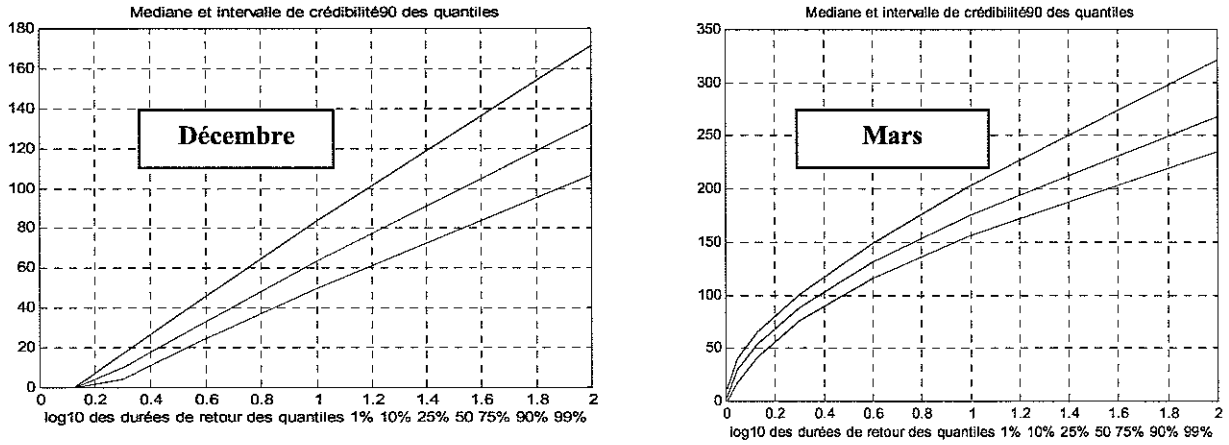


Figure 7 : Medianes et intervalles de crédibilité à 90%
Des quantiles de durées de retour fixées.

Généralisations du modèle des fuites : vers le modèle de Neyman – Scott.

L'approche bayesienne complète « modélisation et calculs conditionnels » est assez souple pour s'adapter à de nombreuses généralisations possibles du modèle.

Dans les applications précédentes à des totaux mensuels on a pu constater certains écarts possibles au modèle des fuites (mars à Atakpamé). D'autre part il est clair que l'hypothèse d'averses instantanées ne permet pas d'appliquer aisément le modèle à des durées plus courtes que le mois. Plusieurs modèles ont été proposés pour décrire le processus des précipitations à échelle fine. Un de ceux ci, le modèle de Neyman Scott en grappes (A.C. Favre, 2003) se distingue par une certaine complexité qui en a rendu assez délicates les applications par les méthodes classiques. A ce jour les chercheurs du domaine ne semblent pas avoir reconnus que sa structure hiérarchique peut en faire une illustration fructueuse des méthodes bayesiennes. Nous présentons ce modèle en suivant A.C Favre avec des notations légèrement différentes.

Il 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 :

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

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

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

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

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

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

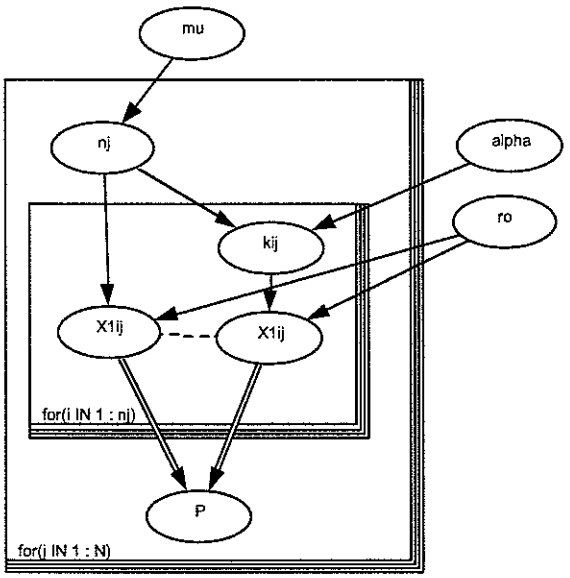
$$[x|\rho] = \rho.\exp(-\rho x) \quad \text{pour x} \geq 0$$

L'inférence statistique de ce modèle n'est pas simple, nous y reviendrons ci après

Dans une première étape où nous continuons à nous intéresser à la pluie mensuelle, nous pouvons négliger la durée des averses et leur succession dans le temps à la suite de l'évenement poissonien générateur. Autrement dit nous ne retenons que la taille des grappes associée à chaque occurrence du processus de Poisson. Bien entendu chaque intensité est distribuée exponentiellement. Nous retenons cette structure de Neyman – Scott simplifiée. Toutefois, pour des raisons de commodité de construction du DAG nous ferons l'hypothèse que K est le nombre d'averses au delà de la première générée par le processus de Poisson des perturbations. Il résulte des propriétés de la loi géométrique que ce nouveau K a la même distribution que l'ancien sur le domaine $k = 0,1,2,....$

Le graphe de la figure 8 présente le DAG de ce modèle de Neyman – Scott simplifié ou modèle des fuites en grappes. Les valeurs de K pour chaque « perturbation » sont maintenant considérées comme variables latentes supplémentaires ainsi que les contributions séparées $X1_{ij}$ et $X2_i$. respectivement de la première averse de la grappe et des suivantes dont la somme reconstitue les P_j. Les conditionnelles complètes des proportions de X1 et X2 par rapport aux P_j sont des distributions Beta . Cet exemple montre bien l'adaptation nécessaire de la part ***I*** du modèle pour tenir compte des mêmes données.

Figure 8 : DAG du modèle des fuites en grappes avec introduction des variables latentes supplémentaires X1 et X2



Nous ne donnons pas ici le détail de la mise en œuvre de l'algoritme de Gibbs et le détail de toutes les conditionnelles complètes nécessaires. Notons que l'introduction de couches supplémentaires de variables latentes nécessite un temps de chauffe plus long de l'algorithme en l'occurrence 2000 itérations supplémentaires.

Toujours pour les deux mois de Décembre et de Mars à Atakpamé, nous donnons ci après d'une part les distributions a posteriori des 3 paramètres puis les graphiques de validation prédictive heuristiques sur les mêmes principes que précédemment.

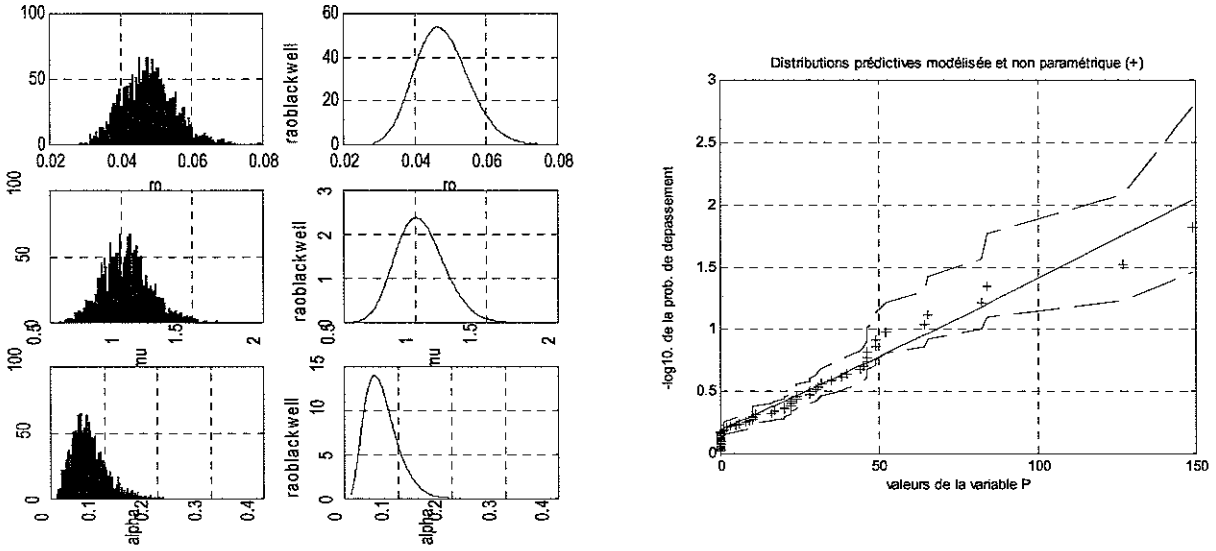


Figure 9 : Résultats du modèle des fuites par grappes
Pour Atakpamé, Décembre

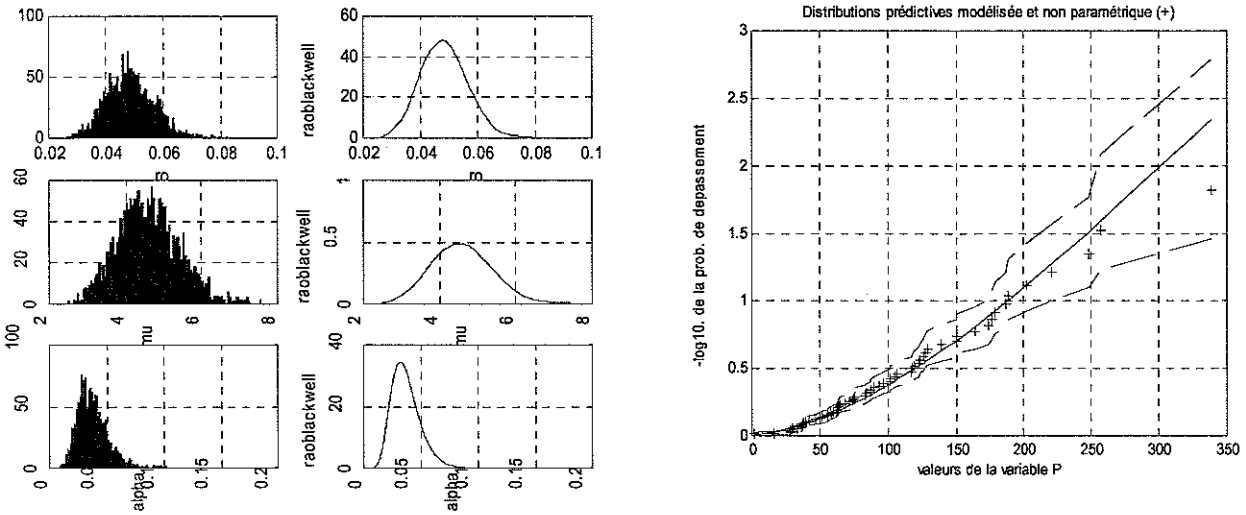


Figure 10 : Résultats du modèle des fuites par grappes
Pour Atakpamé, Mars

Au niveau de la validation prédictive, les résultats sont peu différents de ceux du modèle des fuites simple. On peut noter toutefois une très légère amélioration de la prise en compte des valeurs extrèmes de Mars. Par ailleurs les densités a posteriori des paramètres montrent que ceux ci sont bien identifiés y compris le paramètre de « grappes » α malgré ses faibles valeurs probables. Sur Décembre on note aussi que l'introduction de α modifie légèrement les crédibilités du paramètre de Poisson.

Bien entendu ces résultats, comparables ou non, ne permettent pas un choix entre ces deux modèles. Il y faudrait l'application d'une méthode bayesienne de choix plus rigoureuse (voir Parent et al., 2003).

Premières étapes de la mise en œuvre du modèle de Neyman – Scott

Revenons au modèle de Neyman – Scott, avec cette fois en vue la représentation des pluies à échelle de temps fine, inférieure au mois.

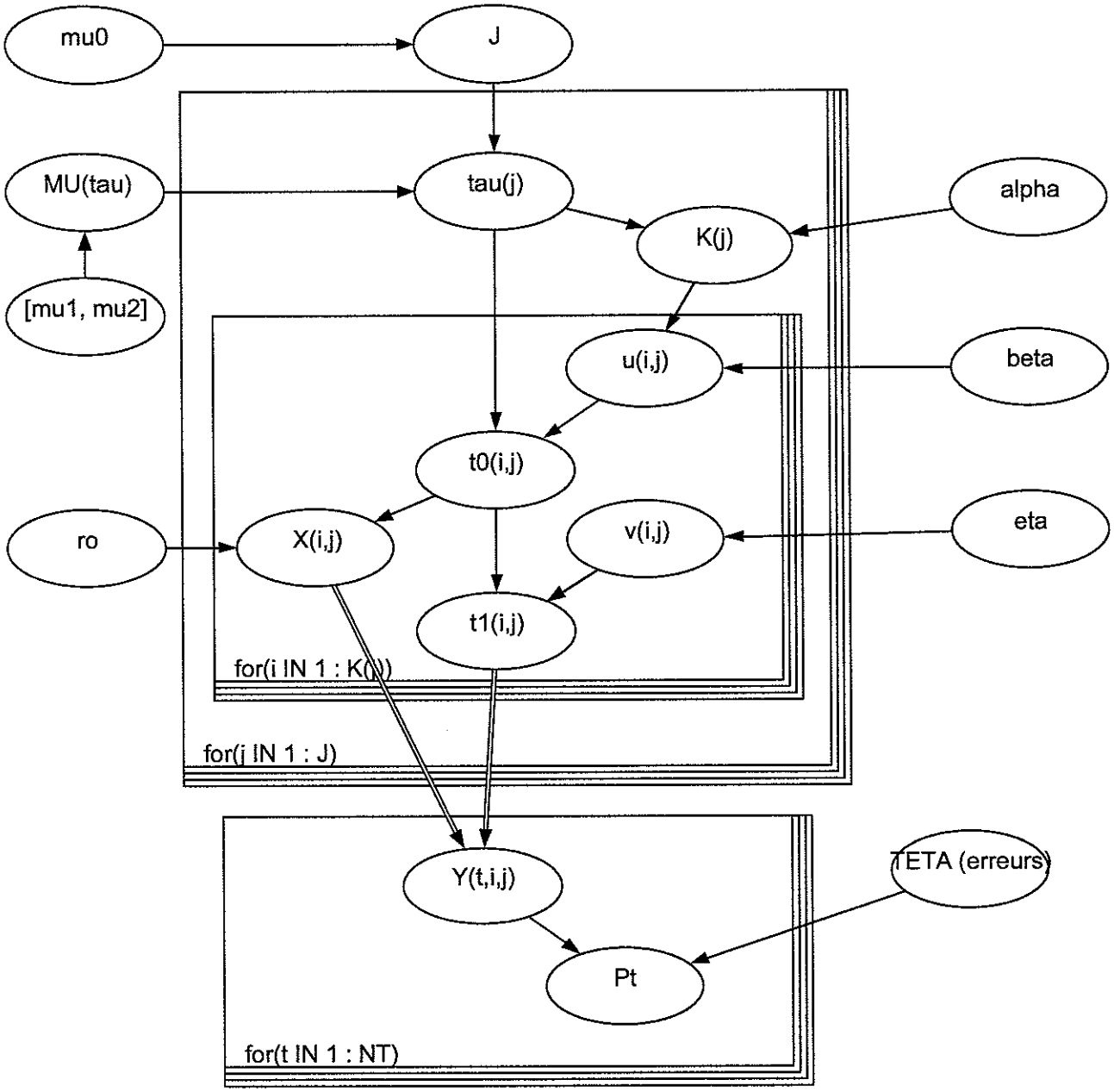


Figure 10 : DAG d'un processus de Neyman – Scott non stationnaire avec « modèle d'erreurs ajouté »

L'estimation classique du modèle de Neyman – Scott pose des problèmes difficiles. La vraisemblance analytique d'un échantillons de « pluies Neyman – Scott » n'est pas connue si bien que les méthodes de calcul d'estimateurs du maximum de vraisemblance par simulation et optimisations ne sont pas disponibles.

Les seules propriétés analytiques utilisables concernent les fonctionnelles caractéristiques du processus des intensités de pluies théoriques dont on peut déduire les moments marginaux et conjoints des sommes sur des intervalles du type P_T. Ceci a conduit les modélisateurs classiques à utiliser des méthodes de moments généralisés en utilisant diverses séquences de moments théoriques et empiriques. A cet égard les méthodes les plus élaborées (cf A.C Favre) utilisent des séquences redondantes de moments dont elles cherchent à minimiser une certaine distance entre moments théoriques et empiriques. Il reste à choisir une distance convenable et on n'a pas de critère de choix vraiment

convaincants. C'est qu'il existe un problème d'estimation non résolu avec ces méthodes. On a vu que l'information utilisée dans les modèles simple et par grappe étaient des sommes P_T assimilées à des sommes mensuelles de pluie. C'est possible parce que l'espace probabilisable de ces sommes peut être assimilé à l'espace des pluies mensuelles. Mais rien n'est moins sûr si des pas de temps journaliers ou horaires par exemple sont choisis. Les chroniques de pluies théoriques générées par Neyman Scott ne ressemblent pas à des chroniques de pluies réelles Nous ne pensons pas que l'aggrégation des pluies à des pas d'une heure ou même d'un jour élimine la difficulté.. Si les écarts entre pluies réelles et pluies théoriques ne sont pas pris en compte , il en résultera un biais non controlé dans la méthode d'estimation.
L'importance de ces écarts apparaît clairement si on tente de construire un DAG logique du modèle de Neyman – Scott tel que celui qui est présenté en figure 10.
Le diagramme est basé sur les principes suivants. Il est conçu à partir de la hiérarchisation de plusieurs processus latents dont le dernier niveau correspond aux données P_T qu'on suppose être des pluies journalières (Ce modèle complexe a peu d'intérêt à un niveau aggrégé). Le principe est de suivre le déroulement dans le temps de la hiérarchie des processus. Pour simplifier l'unité de temps est le jour. Cette discrétisation du temps a peu d'importance pour le processus continu de Poisson considéré comme approchant le processus d'occurrence des précipitations. Par contre les distributions exponentielles des durées doivent être remplaçées par leurs équivalents discrets : les distribution géométriques. L'origine des temps est choisie de telle sorte que les évenements complets soient comptabilisés. Toutes ces hypothèses peuvent être laissées de coté mais elles facilitent notre exposé et ne dénaturent pas le problème statistique vis à vis de la représentation des pluies journalières.
T0 sera la longueur de la séquence des pluies journalières observées sur un site donné.

La figure 10 montre le DAG d'un modèle de « Neyman-Scott » destiné à représenter une chronique de pluies à l'échelle journalière avec toutes ses caractéristiques : intensités moyennes, séquences sans pluies et leurs enchaînements. Remarquons encore qu'il s'agit d'un objectif différent de celui d'une certaine reconstitution des moments simples et croisés des pluies totales sur des périodes fixées. Selon la technique des DAG chaque « flèche » représente un conditionnement probabiliste tel que la « remontée du graphe » permet le calcul des condotionnelles complètes et donc la mise en œuvre dela simulation Gibbs du modèle. Ces conditionnements sont nombreux mais ils sont généralement simples. Nous ne donnerons ici que des informations limitées à certains d'entre eux :
- le processus des perturbations génératrices est supposé être un processus de poisson non homogène, comme, par exemple, un modèle à 3 paramètres :

$$\mu(t) = a_0 T_0 \frac{\exp(a_1 \cos(\varpi t) + a \sin(\varpi t)}{\int \exp(a_1 \cos(\varpi u) + a \sin(\varpi u) du}$$

$$où \quad \omega \text{ est calculé sur la période annuelle}$$

Ce modèle peut éviter les désaisonnalisations empiriques approximatives par découpage du temps.
- On a vu qu'en modélisation bayesienne, il est nécessaire de spécifier les transitions entre les pluies théoriques *Y(t, i, j)* et les pluies « journalières réelles » *P(t)*. Ces transitions étaient relativement simples pour nos modèles des fuites. Ici ils doivent être plus complexes dans la mesure où le processus généré des Y et celui observé des P sont mixtes avec des probabilités non nulles pour les valeurs nulles. Avec cette contrainte , l'hypothèse la plus simple esr d'admettre des transitions indépendantes, sans mémoire, pour chaque jour du type :

$\downarrow Y_t, P_t \rightarrow$	0	$P>0$
0	θ_{00}	$h_0(P\|\theta_0)$
$Y>0$	θ_{00}	$h_1(P\|\theta_1)$

Divers choix sont possibles pour les parties continues h_i de ce modèle de transitions. Par exemples une distribution gamma pour $h_0(P|\theta_0)$ et normale pour $h_1(P|\theta_1)$. Ce modèle plus complexe est en cours de développement et nous ne pouvons en donner d'applications ici.

Conclusions

Avec ces trois exemples pris à la météorologie nous avons voulu montrer la puissance du concept de variables latentes sous ses divers aspects : structurel, données manquantes et variables prédictives. Ce concept est directement opérationnel : il éclaire la modélisation des phénomènes et en permet le calcul **complétement cohérent** sans hypothèses de commodité ad hoc autres que les limitations de temps de calcul. Cette cohérence est le résultat de l'application des mêmes modes de conditionnement : direct dans la modélisation , inverse par Bayes dans le calcul MCMC. C'est d'ailleurs la raison profonde et la justification de « l'abus de langage » si fréquent qui présente les calculs de simulations MCMC (si on en voit que l'aspect calculatoire) comme de véritables méthodes d'inférence et d'estimation statistiques.
Cohérente la démarche est aussi assez souple et efficace pour s'appliquer à des modélisations complexes (modèle de Neyman Scott en grappes non stationnaire par exemple) et aussi aux diverses formes que peuvent prendre les données disponibles pour l'estimation et la validation.

1. Bibliographie restreinte

Coles, J.H. and Powell, E. A. (1996). Bayesian methods in extreme value modelling: a review and new developments. Int.. Statist. Rev., n° 64,119-136.

Favre A.C (2001): Single and multi-site modelling of rainfall based on the Neyman Scott model.
Thèse Ecole Polytechnique Fédérale de Lausanne.

Ferguson T (1973) : A Bayesian analysis of some non parametrics problems . Annals Statist, n°1, pp 209,230.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995) :Bayesian data analysis.} Chapman and Hall, Londres.

Gelman A. (2003) : A Bayesian formulation of exploratory data analysis and goodness of fit testing. To appear in Intern. Statist. Review.

Parent E., Bernier J (2001) : Méthodes bayésiennes et modélisation des risques géophysiques extrèmes. La revue de MODULAB n° 28, pp 1,26.

Parent E., Bernier J (2003) : Bayesian POT modeling for historical data.
Journal of Hydrology , 274, pp 95,108.

Snyder D. L. (1975) : Random Point Processes, Wiley, New York.

Spiegelhalter D. J. , Thomas A. , Best N. G. (1996)} : Computation on Bayesian Graphical Models. Bayesian Statistics n°5 pp 407,425 (Bernardo,Berger, Dawid, Smith editors).

Tanner M. A. (1996) : Tools for statistical Inference. Springer.