Rebastaba : Construction d'un paquet R pour la manipulation de réseaux bayésiens en vue d'une inférence par statistique bayésienne

Jean-Baptiste Denis, Isabelle Albert

Rebastaba

Construction d'un paquet R pour la manipulation de réseaux bayésiens en vue d'une inférence par statistique bayésienne

Jean-Baptiste Denis* — **Isabelle Albert****

* *MiaJ - INRA - domaine de Vilvert - 78352 Jouy-en-Josas Cédex - France*
Jean-Baptiste.Denis@Jouy.Inra.Fr

** *Mét@Risk - INRA - 16 rue Claude Bernard - 75321 Paris Cédex05 - France*
albert@agroparistech.fr

RÉSUMÉ. *Les réseaux bayésiens sont surtout développés et promus par les spécialistes de l'intelligence artificielle ; on peut aussi remarquer qu'ils sont généralement basés sur l'emploi de variables aléatoires discrètes. Nous tentons dans cette communication d'introduire le point de vue de statisticiens confrontés à la modélisation de systèmes et enclins à l'inférence statistique bayésienne. Les logiciels utilisant le langage BUGS de description des réseaux bayésiens permettent cette démarche. Néanmoins, ils représentent une boîte noire et ouvrir le code pour des interventions spécifiques n'est guère possible. À partir de situations concrètes et de la programmation de rebastaba, paquet R pour manipuler les réseaux bayésiens, nous développons nos idées.*

ABSTRACT. *Bayesian networks are mainly supported and promoted by the artificial intelligence community; also most of the developments are made in the framework of discrete variates. Here we attempt to give a statistician viewpoint when modelling systems and/or doing Bayesian inference. Softwares of the BUGS family performs such approaches based on Bayesian networks with a known graph. If their potential is tremendous, for the standard user they represent a black box very difficult to access for specific interventions. Based on several cases that we have tackled, and on the construction of the rebastaba package to manipulate Bayesian networks, we exemplify these opinions.*

MOTS-CLÉS : *réseaux bayésiens, inférence bayésienne, paquet R.*

KEYWORDS: *Bayesian networks, Bayesian inference, R package.*

1. Introduction

L'usage des réseaux bayésiens [RB] se répand rapidement dans de nombreux domaines, y compris en biologie, champ d'application dans lequel se situe l'INRA. Les raisons sont multiples : (i) ils se basent sur une modélisation graphique ne nécessitant que peu de formalisme mathématique, (ii) ils permettent d'aborder des situations complexes, difficiles avec des outils méthodologiques plus traditionnels, (iii) de puissants et conviviaux logiciels sont maintenant disponibles. Mais le terme de RB recouvre différents types d'approches méthogologiques et une certaine confusion règne au moins dans l'esprit des experts des domaines de l'application. Après quelques années d'investissement autour d'un certain type de RB, il nous est apparu intéressant de tenter d'apporter la vision de statisticiens modélisateurs sur leur emploi. Notre réflexion s'appuie sur des cas d'espèces précis et la mise en chantier d'un paquet R (R, 2008) nommé rebastaba (Rebastaba, 2008).

S'il compulse les publications autour des réseaux bayésiens, par exemple à partir de l'ouvrage de Naïm (2004) (Naïm *et al.*, 2004), le statisticien est frappé par un certain nombre de constatations :

– Dans les logiciels disponibles sur le marché, les RB sont toujours (?) associés à des variables discrètes catégorielles. Même si la plupart proposent des utilitaires pour discrétiser automatiquement et de diverses manières les variables continues, les algorithmes au coeur des applications sont fondés sur un schéma implicite de distributions multinomiales. Ceci a pour conséquences :

- la difficulté et l'arbitraire du choix des limites de classes lorsque la vraie variable est de type continu.

- la multiplication du nombre de paramètres nécessaires pour définir les lois conditionnelles, entraînant la réduction du nombre de classes et du nombre de parents.

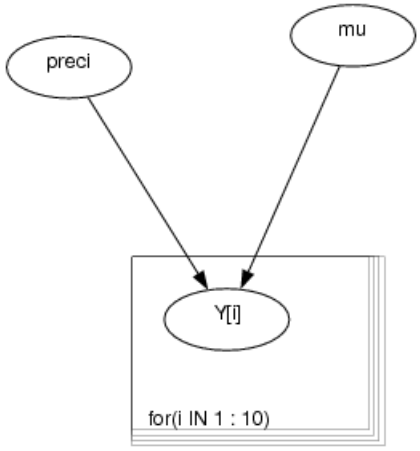
– Dans ces mêmes logiciels, ce sont les distributions marginales des variables du système qui sont principalement représentées alors que les RB permettant la définition parcimonieuse (en nombre de paramètres) d'une distribution de probabilité conjointe sur une grand nombre de variables aléatoires, autorisant l'exploitation du caractère conjoint de la modélisation.

– La mise en avant, car c'est ludique et intuitivement abordable, du graphe de la structure du réseau au risque d'oublier que toutes les liaisons ne sont pas de même importance, en fonction des distributions de probabilité qui leur sont associées.

– Un certain flou (malgré les travaux de Pearl (Pearl, 2000)) sur l'association que l'on peut faire entre les arcs aboutissant à un noeud et les causalités qui gouvernent la distribution de la variable aléatoire liée au noeud.

– L'élaboration et la promotion des RB sont principalement réalisées par les chercheurs liés à l'intelligence artificielle. Sont mises en avant les problématiques liées à cette discipline comme l'apprentissage alors que les aspects statistiques (inférence et qualité des estimations ; choix de modèles) sont moins apparents. On peut aussi relever une certaine confusion entre la statistique bayésienne et les réseaux bayésiens.

Figure 1. *Doodle produit avec l'interface graphique d'OpenBUGS. A chaque noeud est associée sa distribution de probabilité (marginale ou conditionnelle), consultable interactivement. Classiquement, le vecteur Y est associé aux observations.*



Il existe cependant une exception aux généralités précédentes : celle des logiciels de la famille BUGS (OpenBUGS, 2007; JAGS, 2008). Ils sont fondés sur une démarche statistique et utilisent sans le dire les RB, le plus souvent avec des variables aléatoires continues ! De fait, c'est à partir de leur usage que nous sommes venus aux RB.

Nous montrerons successivement comment la statistique bayésienne s'exprime assez directement au moyen de RB (§2), pourquoi il nous paraît logique que la distribution conjointe priore des paramètres devienne l'objectif central d'une démarche de modélisation et comment elle se situe par rapport à la technique de simulation de Monte-Carlo (§3) pour finalement évoquer quelques traits de la programmation que nous poursuivons en parallèle de nos travaux de modélisation (§4).

Signalons que nous n'abordons pas ici la question importante de l'apprentissage (ou estimation) de la stucture du réseau. Également, nous n'évoquerons pas l'utilisation de la statistique fréquentiste pour l'inférence d'un RB ; elle est possible si les données sont assez fournies pour qu'aucun problème d'identifiabilité ne surgisse, ce que favorisent par exemple les modélisations hiérarchiques (par exemple les modèles mixtes).

2. Réseaux Bayésiens et Statistique Bayésienne

Les logiciels de la famille BUGS permettent le tirage dans les distributions conditionnelles aux observations par la construction et l'usage d'algorithmes MCMC (Gilks

Tableau 1. *Code OpenBUGS résultant de la figure 1 et analyse qu'en fait Jags. Les 16 noeuds recensés sont Y, mu, preci, les 2 paramètres de la normale (`dnorm`) et les 2 paramètres de l'uniforme (`dunif`).*

```
model {
    mu ~ dnorm(50,1.0E-6);
    preci ~ dunif(0.01,10);
    for( i in 1 : 10 ) { Y[i] ~ dnorm(mu,preci);}
}
############
. model in doodle.jam
. compile
Compiling model graph
   Resolving undeclared variables
   Allocating nodes
   Graph Size : 16
```

et al., 1996). Leur affichage est celle d'une inférence statistique bayésienne, néanmoins la modélisation qu'ils permettent des données est basée sur des RB comme le montrent la figure 1 d'un *doodle* généré avec OpenBUGS (OpenBUGS, 2007) et le tableau 1 où se trouvent le code résultant produit par OpenBUGS et l'analyse syntaxique qu'en fait Jags (JAGS, 2008). La construction du *doodle* à l'aide de la souris et de menus contextuels est bien celle d'un RB ; la compilation qu'en fait Jags passe bien par le comptage du nombre de noeuds du RB.

Du point de vue de la statistique bayésienne, on définit deux composantes principales dans la modélisation probabiliste :

1) la vraisemblance qui n'est autre que la distribution conditionnelle des données sachant les paramètres : $[Y \mid \theta]$,

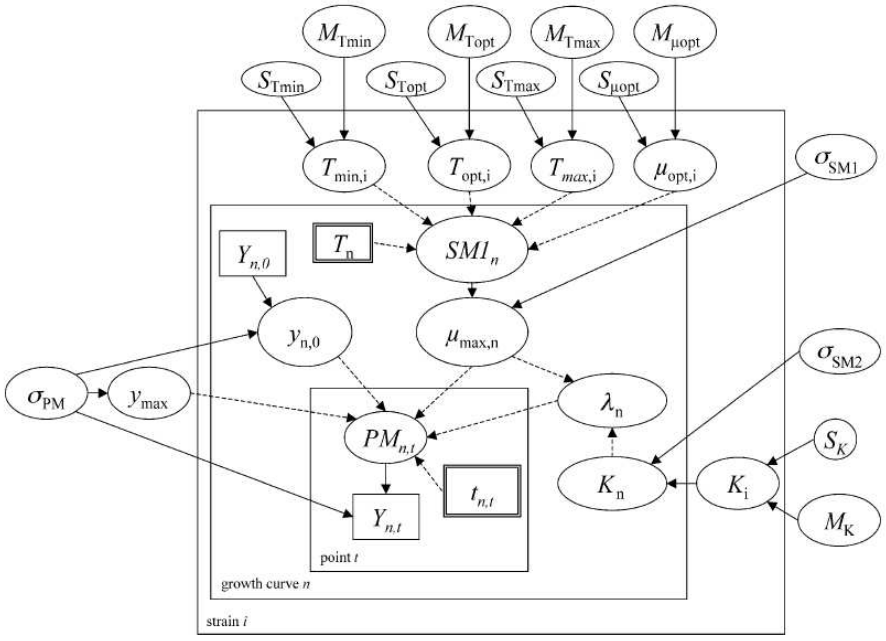
2) la priore qui n'est autre que la distribution conjointe marginale des paramètres : $[\theta]$.

Elles permettent, grâce au théorème de Bayes, de trouver :

– la postériore qui n'est autre que la distribution conditionnelle des paramètres par rapport aux données : $[\theta \mid Y] \propto [\theta] \,.\, [Y \mid \theta]$.

Dans une perspective de RB, on définit le réseau à deux noeuds : $\theta \rightarrow Y$ puis on renseigne Y par les valeurs observées, ce qui met à jour la distribution de θ (passage de la priore à la postériore). En fait, les deux noeuds de ce réseau conceptuel sont des *super*-noeuds puisque le premier contient l'ensemble des paramètres du modèle (qui sont en général assez nombreux) et le second l'ensemble des données (qui est - classiquement - encore plus fourni). Ceci explique qu'en général, chacun des deux noeuds est en fait lui-même un réseau bayésien plus ou moins complexe. A titre d'exemple traité, la figure 2 reproduit la figure 1 de Pouillot et al. (Pouillot *et al.*, 2003) où est présenté le

Figure 2. *Graphe du réseau bayésien utilisé pour modéliser une croissance bactérienne (Pouillot* et al.*, 2003). Les données sont $Y_{n,0}$ et $Y_{n,t}$. Les T_n et $t_{n,t}$ jouent un rôle de covariables : toute la modélisation leur est conditionnelle.*

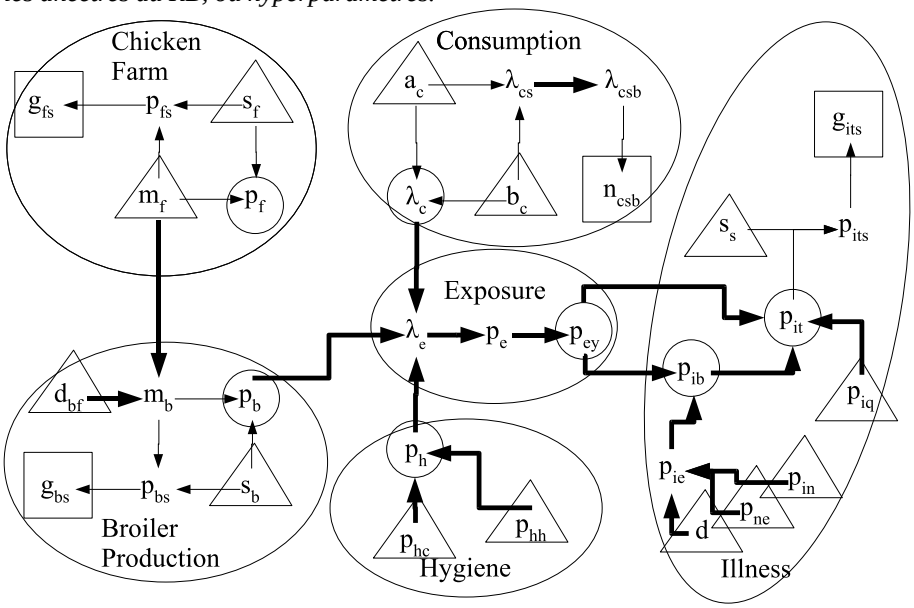


RB d'un modèle assez imbriqué de la croissance d'une population bactérienne.

Dans un contexte de méta-analyse statistique où différents ensembles de données hétérogènes doivent être analysés conjointement, nous avons poursuivi cette démarche en identifiant clairement ce que nous avons appelé le *core model*, distribution priore conjointe des paramètres servant à modéliser l'ensemble des données disponibles (c'est à dire définir la vraisemblance). Dans la présentation en *super*-noeuds, nous aurions, si les ensembles de données étaient au nombre de deux, un réseau de la forme : $Y_1 \leftarrow \theta \rightarrow Y_2$ correspondant à la définition d'une priore ($[\theta]$) et de deux vraisemblances ($[Y_1 \mid \theta]$ et $[Y_2 \mid \theta]$). Bien entendu, le résultat obtenu est cohérent avec une introduction séquentielle des données quelque soit l'ordre utilisé :

$$\begin{aligned}
[\theta \mid Y_1, Y_2] &\propto [\theta]\,[\theta \mid Y_1]\,[\theta \mid Y_2] \\
&\propto [\theta]\,[\theta \mid Y_1]\,[(\theta \mid Y_1) \mid Y_2] \\
&\propto [\theta]\,[\theta \mid Y_2]\,[(\theta \mid Y_2) \mid Y_1]\,.
\end{aligned}$$

Figure 3. *Modélisation de la chaîne alimentaire pour décrire le processus conduisant à des campylobactérioses dans une population au travers de poulets (Isabelle Albert* et al.*, in press). Les données correspondent aux noeuds rectangles. Les triangles sont les ancêtres du RB, ou hyperparamètres.*

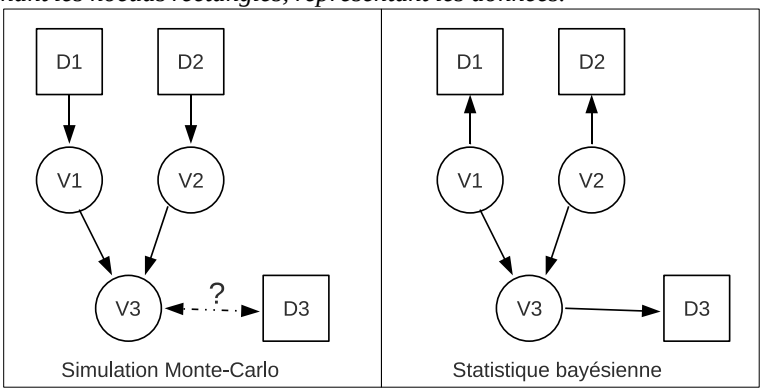


Cette façon de procéder a été appliquée pour modéliser l'appréciation du risque cambylobacter par le poulet au travers de la description de la chaîne alimentaire comme l'illustre la figure 3 issue d'Albert et al. (Isabelle Albert *et al.*, in press).

3. Réseaux Bayésiens et Simulateur Stochastique

Il existe deux manières opposées pour déterminer la loi priore en statistique bayésienne. Soit on tente de minimiser son influence sur les résultats de l'inférence ; cela conduit entre autres à proposer des distributions impropres, comme la mesure de Lebesgue qui rend la postériore proportionnelle à la vraisemblance puisque $[\theta] = 1$; il faut alors vérifier analytiquement que la postériore est bien une probabilité. Soit, et c'est plutôt notre point de vue, on cherche à définir sans l'aide des données une vision autonome du phénomène étudié, c'est à dire de rassembler dans la priore toutes les informations disponibles, représentant au mieux les connaissances des experts. Ce type de démarche, qui s'apparente à la construction de systèmes experts, est connue sous la dénomination d'*élicitation* (P. H. Garthwaite *et al.*, 2005). Cette posture nous a conduit tout naturellement à nous pencher sur la construction de priores mettant en jeu un nombre important de variables aléatoires. Ce faisant, nous rejoignons une pratique

Figure 4. *RB associés respectivement à une situation de simulation de Monte-Carlo et à une situation d'inférence statistique bayésienne. On notera la différence des arcs concernant les noeuds rectangles, représentant les données.*



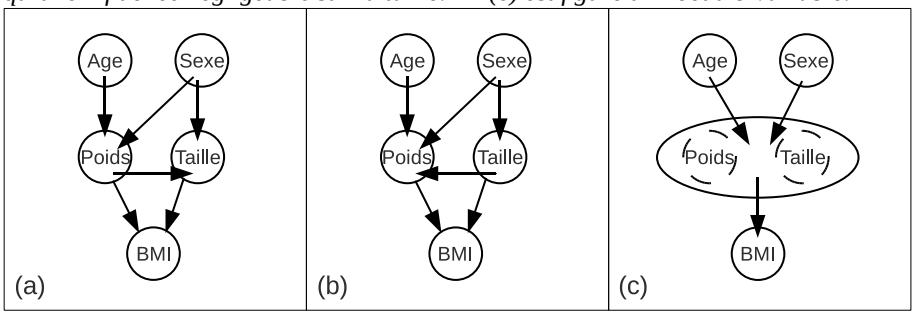
connue dans le domaine de l'appréciation des risques sous le terme de simulation de Monte-Carlo. Placé dans le cadre de §2, selon notre terminologie, c'est simplement la construction du *core model*. Cependant notre démarche est distincte de la simulation de Monte-Carlo classiquement utilisée, même si en pratique cette position de principe n'est pas toujours tenue car certains experts préfèrent s'appuyer sur des données pour faire des propositions quantitatives. Nous ne voulons pas faire intervenir les données disponibles dans la définition du *core model* et tentons de nous baser uniquement sur les idées des experts, réservant les données pour le calcul de la postériore. Un avantage important de cette position est que nous pouvons utiliser des données qui informent des noeuds non ancêtres du RB sous-jacent comme le montre la figure 4. Lors d'une simulation Monte-Carlo, les données servent à renseigner les distributions de probabilité des variables ancêtres du RB, pour les autres (cas de la variable $V3$) elles ne peuvent servir qu'à valider par confrontation la construction basée sur les ancêtres. S'il y a conflit, le modélisateur est embarrassé pour modifier le RB en conséquence. Au contraire, dans une démarche de statistique bayésienne, ce sont les données qui sont les enfants des variables du système, la distribution conjointe est définie de manière unique et le conditionnement ne pose aucun problème théorique. Construire un *core model* se fait assez naturellement en trois étapes principales :

1) choix des variables d'intérêt (y compris leurs unités et leurs natures),

2) définition du graphe des relations qui expriment leurs covariations, soient directes ou indirectes (par l'ajout de variables intermédiaires),

3) précision des distributions de probabilités sur l'ensemble des variables recensées (distributions marginales pour les ancêtres, conditionnelles pour les autres).

Figure 5. *Différents types de RB pour une même situation. Chez les adultes, l'âge n'a qu'une influence négligeable sur la taille. En (c) est figuré un noeud bivariable.*

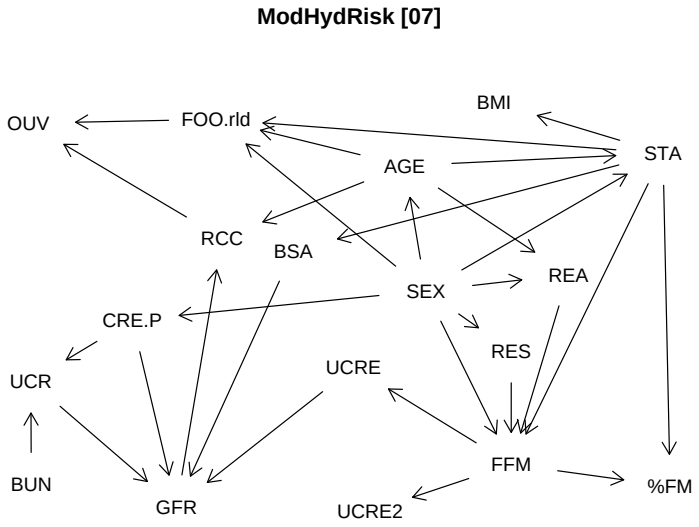


Chaque cas de figure est un cas d'espèce ; nous voudrions cependant attirer l'attention sur la manière de définir la structure du graphe du RB. Il ne doit pas comporter de cycles, mais la difficulté n'est pas là. Beaucoup de structures sont équivalentes, c'est à dire que l'on peut établir les mêmes distributions de probabilité conjointe à partir de chacune ; un certain nombre de choix doivent donc être opérés. Considérons le cas très simple de deux noeuds P et T, représentant le poids et la taille d'individus d'une population définie : vaut-il mieux retenir $P \rightarrow T$ ou $P \leftarrow T$ pour modéliser la corrélation forte qui existe entre ces deux variables ? Pour illuster le propos considérons les trois possibilités représentées en figure 5 . Y interviennent assez naturellement les variables Age et Sexe comme parents, et l'indice de masse corporelle (BMI) comme enfant direct. Plutôt que de choisir le sens entre P et T pour représenter la corrélation non expliquée par l'âge et le sexe entre les deux variables, il peut être préférable d'introduire un noeud bivariable **sta**ture composé de la taille et du poids des individus de la population modélisée : c'est la troisième possibilité. C'est la solution qui a été adoptée dans le cadre d'un travail mené avec L. Mioche (INRA-Clermont) dont le graphe du RB utilisé pour le projet ModHydRisk est reproduit en figure 6 où STA est ce noeud bivariable.

Le principe que nous utilisons pour choisir le sens des arcs, est de coller au mieux aux relations supposées de causalité car toute distribution de probabilité est contingente d'un certain contexte. Si on souhaite que la modélisation soit pertinente dans un maximum de situations possibles mieux vaut emprunter les relations de causalité qui sont les plus stables. Une autre bonne raison pour procéder ainsi est que les experts des applications raisonnent naturellement de cette manière. Il est parfois cependant difficile de toujours trancher, des noeuds multivariables peuvent permettre d'éviter le dilemne.

La détermination des distributions de probabilité n'est pas non plus toujours évidente (prise en compte de disymétrie, de queues lourdes, de mélanges de populations...). Par exemple, contraitrement à ce que l'on pourrait imaginer les relations

Figure 6. *Graphe du RB associé à une modélisation du risque de déshydratation chez les seniors.*



conditionnelles entre (Poids, Taille) pour le Sexe et l'Age ne sont pas du tout simple à établir. Un recours, si les données sont disponibles, est ce que nous avons dénommé les distributions empiriques. La technique est simple, il s'agit de remplacer un tirage pseudo-aléatoire par un tirage dans une base de données réelles. Celle-ci doit être adaptée (représentative) de la situation que l'on souhaite modéliser. Si le noeud n'a pas de parent dans le RB, on tire - de manière uniforme - une valeur de la variable parmi celles que comprend la base. Si le tirage est conditionné par un ou plusieurs parents, le tirage est restreint au sous-ensemble d'individus de la base qui leur correspond. Les tirages conditionnels posent deux problèmes : la définition de la restriction (qui peut être améliorée par des tirages non uniformes) et la taille du sous-ensemble qui pour les événements rares (ceux qui en général intéressent) peut devenir critique. Une conséquence défavorable des distributions empirique est que la démarche statistique bayésienne devient quasi impossible à partir du *core model* résultant.

4. Rebastaba

Même si l'usage de procédures MCMC requiert prudence, les logiciels de la famille BUGS nous ont donné satisfaction. Cependant, l'impossibilité de manipuler les RB implicitement utilisés devient vite une petite frustration et c'est la principale raison pour laquelle le projet *rebastaba* (**ré**seaux **ba**yésiens traités par **sta**tistique **ba**yésienne) a été lancé. Les sources et le manuel d'utilisation sont disponibles à partir de (Rebastaba, 2008). Dans le cadre d'une programmation **R** (R, 2008), les principaux objectifs sont :

– créer facilement des RB comprenant une grande variété de distributions de probabilité, en particulier associant variables discrètes et continues,

– permettre leur manipulation et leur utilisation par programmation,

– ne pas reproduire ce qui existe déjà mais y donner accès par des interfaces adaptées. En particulier, la génération de code de modèle BUGS qui sont particulièrement peu adaptés pour le codage à la main de variables discrètes multinomiales. Pour disposer de l'équivalent des boucles simples, une forme spécifique de noeud multivariable a été mise en place : leurs variables sont indépendantes et suivent des distributions identiques ou similaires.

Tous ces objectifs sont plus ou moins en cours de réalisation ; au fur et à mesure qu'avance le projet des difficultés non imaginées surgissent, d'autres sous-objectifs intéressants se présentent. De plus, l'avancement n'est pas théorique mais suit d'assez près les besoins ressentis pour l'avancement de quelques cas concrets comme le projet ModHydRisk évoqué ci-dessus. La programmation orientée objet suit les classes S4 de S (Venables *et al.*, 2000) qui facilitent par leurs contrôles internes les constructions périlleuses de codes prototypaux. A terme, il est envisagé d'en faire un paquet d'extension contributive de **R**(R, 2008).

Parmi beaucoup d'autres, trois principaux types d'objets (informatiques mais surtout conceptuels) se sont peu à peu mis en place : les RB au sens strict (objets `bn`), les graphes de parenté (objets `gn`) et les tableaux de données (objet `dn`).

– Les objets `bn` représentent les RB dans toute leur généralité, c'est à leur seul niveau que sont spécifiées les distributions de probabilité et que la structure noeud/variable est exploitée. Leur spécification est double : description fournie par l'usager et fonctions générées à partir de celle-ci pour la standardisation du calcul numérique. A titre d'exemples, des `bn` peuvent être générés pseudo-aléatoirement.

– Les objets `gn` représentent des graphes dirigés ; ils permettent de manipuler les graphes associés aux RB mais acceptent aussi les graphes comportant des cycles. Les distributions de probabilités y sont donc ignorées et les noeuds multivariables ne sont pas distingués. Plusieurs algorithmes d'exploration des propriétés des parentés sont ou seront implémentés.

– Les objets `dn` représentent des tables de données associables à des RB, soit générées à partir d'eux, soit utilisables pour les estimer. Un certain nombre de traitements statistiques élémentaires (comme des représentations graphiques) leur sont associés,

avec le souci de bien prendre en compte la nature (continue / entière / catégorielle ordonnée / catégorielle non ordonnée) des variables aléatoires présentes.

5. Conclusion

Les RB représentent un formidable outil pour les modélisateurs et les statisticiens : ils permettent d'aborder des situations complexes de manière progressive, réaliste et accessible aux non spécialistes. Ces qualités vont de pair avec l'explosion de leur utilisation et la disponibilité de logiciels nombreux et variés. La possibilité de pouvoir les programmer pour s'adapter à des cas spécifiques nous semble essentielle pour certaines applications. La réalisation d'un tel outil, lors de la spécification d'objets informatiques, est un activateur bénéfique de la réflexion. Nous avons pleine conscience du caractère limité de notre approche dans la mesure où nous traitons de RB de taille relativement modeste dont la structure n'est pas estimée automatiquement. Néanmoins, nous restons persuadés que le point de vue statistique peut être de quelque utilité ; c'est dans cette conviction que ce papier a été rédigé. En particulier, nous espérons avoir montré que l'usage des RB est beaucoup plus large que celui qui en est fait par la prise en compte de variables continues et l'introduction de la statistique bayésienne.

6. Bibliographie

Gilks W. R., Richardson S., Spiegelhalter D. J. (eds), *Markov chain Monte Carlo in practice*, Chapman and Hall, London, 1996.

Isabelle Albert E. Grenier J.-B. D., Rousseau J., « Quantitative risk assessment from farm to fork and beyond : a global Bayesian approach concerning food-borne diseases », *Risk Analysis*, in press.

JAGS, http ://www-fis.iarc.f/ martyn/software/jags/, 2008.

Naïm P., Wuillemin P.-H., Leray P., Pourret O., Becker A., *Réseaux Bayésiens*, Editions Eyrolles, Paris, 2004.

OpenBUGS, http ://mathstat.helsinki.fi/openbugs/, 2007.

P. H. Garthwaite J. B. K., O'HaganRégis A., « Statistical methods for eliciting probability distributions », *J. American Statistical Association*, vol. 100, p. 680-701, 2005.

Pearl J., *Causality : Models, Reasoning and Inference*, Cambridge University Press, Cambridge, 2000.

Pouillot R., Albert I., Cornu M., Denis J.-B., « Estimation of uncertainty and variability in bacterial growth using Bayesian inference », *Int. J. Food Microbiology*, vol. 81, n° 2, p. 87-104, 2003.

R, http ://www.r-project.org/, 2008.

Rebastaba, http ://w3.jouy.inra.fr/unites/miaj/public/matrisq/jbdenis/outils/welcome.html, 2008.

Venables W., Ripley B., *S programming*, Springer, New York, 2000.