De l'interprétation statistique des données à une modélisation probabiliste actualisée par les données : la synthèse bayésienne. Principes et exemples.

Jean-Baptiste Denis*, Isabelle Albert**

*MIAJ - INRA, F78352 Jouy-en-Josas
Jean-Baptiste.Denis@Jouy.Inra.Fr,
http://w3.jouy.inra.fr/unites/miaj/public/matrisq/jbdenis/welcome.html
**Met@risk - INRA, F75231 Paris cedex 05
Isabelle.Albert@Paris.Inra.Fr
http://www.paris.inra.fr/metarisk/members/albert_isabelle

Résumé. Statisticiens plongés dans un domaine d'application où les données sont rares et hétérogènes, nous proposons d'utiliser successivement deux approches bayésiennes complémentaires : les réseaux bayésiens et la statistique bayésienne. Dans une première partie, les deux approches sont brièvement rappelées pour montrer qu'à part le théorème de Bayes, elles n'ont, dans nos acceptions des deux concepts, rien en commun. S'appuyant ensuite sur la modélisation du nombre de campylobactérioses en France liées à la consommation de poulets, un nouveau point de vue est suggéré pour l'interprétation des données. Il s'agit (1) de modéliser en soi le phénomène d'intérêt à l'aide d'un réseau bayésien ; puis (2) de l'étendre pour définir la vraisemblance des données disponibles et extraire l'information qu'elles contiennent par conditionnement, c'est-à-dire en appliquant le principe de la statistique bayésienne. R et les logiciels de la famille BUGS se révèlent bien adaptés pour la réalisation pratique de cette proposition.

1 Introduction

Dans certains domaines, comme celui de la sécurité sanitaire microbiologique liée à l'alimentation humaine dans lequel nous travaillons, les approches statistiques standard ne sont guère aisées à pratiquer. Alors que les questions sont précises et les réponses lourdes de conséquence, les données sont peu nombreuses et celles qui sont disponibles sont très hétérogènes. Par exemple : si on permet à une ville de 10000 habitants de consommer une eau de source dont les deux derniers prélèvements ont révélé la présence d'oocystes de cryptosporidium, quelles peuvent en être les conséquences sur la santé de la population en termes de maladies, voire de décès ?

Traditionnellement, les décisions sont prises par les politiques après consultation d'experts qui synthétisent implicitement et subjectivement un certain nombre de situations similaires soit vécues soit relatées dans la litérature du domaine. Mais cette démarche pragmatique est de moins en moins prisée car d'une part les citoyens exigent de plus en plus les justifications des

décisions prises en santé publique (demande sociétale), et d'autre part les décisions de refus d'importation de denrées alimentaires pour des raisons de sécurité sanitaire doivent désormais être argumentées de manière scientifique (contraintes du commerce international).

L'approche scientifique passe alors par une formalisation quantitative du problème, assez naturellement probabiliste du fait de la méconnaissance de très nombreux facteurs influents et de la variabilité inhérente aux phénomènes impliqués. Pour reprendre l'exemple de la cryptosporidiose, une réponse attendue pourrait être que "sur une année de consommation de ce type d'eau, la probabilité d'avoir plus de n=10 malades dans la population sera comprise entre 0.50 et 0.95 autour d'une valeur centrale de 0.70" ; ceci pour différentes valeurs de n. Pour arriver à une telle réponse, les difficultés sont nombreuses car beaucoup d'éléments sont mal connus : la consommation d'eau du robinet, la fonction dose-réponse des différentes classes d'individus. En effet, les données sont rares, hétérogènes, pas toujours complètement pertinentes pour la situation à traiter. Pour certains pans du problème, on est même réduit à la seule connaissance subjective d'un ou deux experts !

Pour avancer malgré tout dans l'*évaluation quantitative du risque*, nous avons eu recours aux approches bayésiennes. Elles permettent, nous semble-t-il, de relever le défi et de commencer à poser les jalons de démarches destinées à s'enrichir naturellement au fur et à mesure de la disponibilité de nouvelles données. Cette réflexion, élaborée et mise en œuvre pour la modélisation du nombre de campylobactérioses en France provoquées par le poulet [Albert et al. (2008)], a également été éprouvée dans d'autres situations (pour modéliser l'évolution d'une épidémie animale [Courcoul et al. (2009)], pour estimer la prévalence de la campylobactériose en France à partir de données épidémiologiques [Albert et al. (2009)]). Nous sommes persuadés qu'elle est transposable au-delà de notre domaine d'application ; nous en exposons ici les principes et évoquons quelques exemples.

Le cœur de la proposition est assez simple à énoncer : plutôt que de modéliser directement les données disponibles pour en faire l'interprétation statistique comme il est courant, une étape préliminaire de modélisation probabiliste du phénomène étudié est réalisée indépendamment des données. Ce modèle est dénommé modèle central (core model). Puis les données sont rattachées à ce modèle central pour son actualisation. Plus précisément, nous préconisons l'usage des réseaux bayésiens pour la définition du modèle central, et l'usage de la statistique bayésienne pour sa mise à jour par l'information apportée par les données disponibles, d'où le terme de *synthèse bayésienne*.

Dans ce papier, nous commençons par établir la distinction nécessaire entre *modélisation probabiliste* et *interprétation statistique* qui se traduiront respectivement dans notre cas par l'usage de réseaux bayésien et de la statistique bayésienne qui comme on le verra se conjuguent très naturellement dans quelques logiciels.

2 Modélisation probabiliste et Interprétation statistique

Notre proposition consiste à coupler une modélisation probabiliste et une interprétation statistique de données. Plutôt que de rechercher quelle vraisemblance pourrait être associée aux données disponibles, il s'agit d'abord de définir les variables fondamentales qui gouvernent le système objet de l'investigation, décider lesquelles doivent avoir le statut d'aléatoire et leur associer une distribution conjointe. C'est la modélisation probabiliste. Ce n'est qu'ensuite que

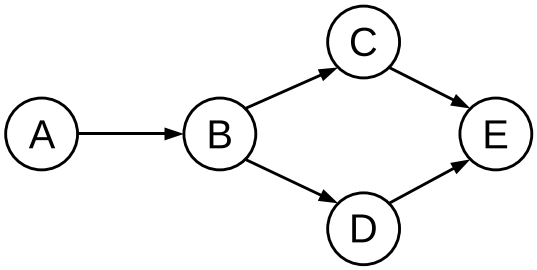

FIG. 1 – *Graphe acyclique dirigé du réseau bayésien de l'équation (1).*

la vraisemblance en est déduite ; ce que nous dénommons l'interprétation statistique. Bien entendu, en pratique la première étape n'est pas complètement indépendante de la seconde.

2.1 Modélisation probabiliste

Il s'agit de définir la distribution de probabilité conjointe d'un ensemble de variables, de manière parcimonieuse en tenant compte des relations plus ou moins directes qu'on leur attribue. Dans cet objectif, les modèles dits graphiques sont très utiles [Edwards (2000)]. Parmi ceux-ci, les réseaux bayésiens [Jensen (2001)] sont particulièrement commodes car les conditionnements successifs sur lesquels ils sont basés, peuvent traduire les relations de causalité que les experts du domaine assument.

Ils définissent en effet la loi de probabilité conjointe d'un ensemble de variables aléatoires au moyen de conditionnements restreints aux relations directes d'un *graphe acyclique dirigé* ; la figure 1 en donne un exemple simple. Ils offrent la possibilité de proposer facilement des sous-modèles par simplification de ces conditionnements. La loi conjointe, $[A, B, C, D, E]$, des cinq variables aléatoires $\{A, B, C, D, E\}$ peut toujours s'écrire comme le produit de $[A]$ par $[B \mid A]$, $[C \mid A, B]$, $[D \mid A, B, C]$ et $[E \mid A, B, C, D]$ où la barre verticale désigne le conditionnement. Si on admet (pour des considérations liées aux variables aléatoires impliquées) de simplifier cette forme générale en

$$[A, B, C, D, E] = [A]\,[B \mid A]\,[C \mid B]\,[D \mid B]\,[E \mid C, D] \tag{1}$$

alors on peut vérifier, quelques soient les distributions en jeu, qu'un certain nombre d'indépendances conditionnelles s'en déduisent, par exemple ici :

$$\begin{aligned} A \mid B \quad &\perp\!\!\!\perp \quad E \mid B \\ C \mid B \quad &\perp\!\!\!\perp \quad D \mid B \end{aligned}$$

En fait, l'attractivité des réseaux bayésiens est de visualiser ce genre de propriétés au moyen d'un graphe acyclique dirigé dont les nœuds sont les variables aléatoires et les variables conditionnantes de chaque variable aléatoire sont les nœuds parents de la variable considérée (voir la figure 1).

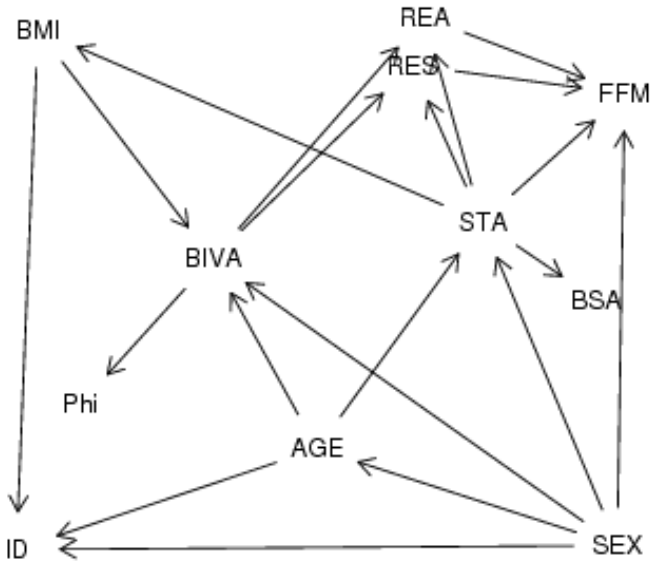

FIG. 2 – *Composition corporelle d'une population donnée d'individus humains (travail mené avec L. Mioche (INRA-Clermont)). Les principales variables impliquées sont le sexe (*SEX*), l'âge (*AGE*), la stature (*STA*) vecteur de la taille et du poids, l'indice de masse corporelle (*BMI*), la réactance (*REA*) et la résistance (*RES*) du corps, la surface corporelle (*BSA*), et la masse de viande maigre (*FFM*) qui est la variable cible du modèle.*

Dans la pratique, on procède à l'inverse : on bâtit le graphe des relations directes entre les variables qu'on veut/doit considérer, le plus souvent en se basant sur les causalités admises ou supposées qui les lient, puis on définit pour chacune une distribution conditionnelle (ou marginale si aucune flèche n'aboutit sur le nœud associé). On obtient ainsi de manière simple car locale une modélisation stochastique complexe, globale et parfaitement cohérente sur l'ensemble des variables. Ajouter ou retrancher une variable d'un réseau bayésien s'opère également de manière locale, tout en conservant la probabilité conjointe des autres variables.

Les réseaux bayésiens sont donc un outil commode pour définir des probabilités conjointes parcimonieuses en paramètres et fidèles à une vision de la réalité. Ils se révèlent également extrêmement efficaces pour envisager une modélisation avec des personnes rebutées par les formalisations mathématiques car la discussion s'opère autour du graphe qui précise les relations directes entre les variables. Ils représentent aussi des supports fort utiles pour la réflexion dans le cas de modèles complexes. Par exemple, les graphes des figures 2 et 3 nous ont servi pour dialoguer avec les experts des domaines concernés.

Les réseaux bayésiens ont beaucoup été étudiés en Intelligence Artificielle dans une optique *apprentissage* [Lauritzen et Spiegelhalter (1988), Pearl (2000), Jensen (2001), Marin et Rossi (2004) et Darviche (2009)], alors que les statisticiens [Whittaker (1990), Lauritzen (1996) et Edwards (2000)] se sont plus attachés à étudier les modèles graphiques basés sur des graphes

non orientés, parfois dénommés réseaux de Markov. Les réseaux bayésiens, souvent dénommés *DAG* pour *directed acyclic graphs*, sont alors mentionnés pour être ramenés par une opération dite de *moralisation* dans le cadre des réseaux de Markov. Mais les deux familles de réseaux (orientés ou pas) ne sont pas équivalentes. Aucune des deux n'englobe l'autre comme il est montré dans Bishop (2006). Une autre différence notable entre ces deux familles est que les réseaux de Markov traités sont principalement multinomiaux pour les variables discrètes et multinormaux pour les variables continues, alors qu'une très grande variété de distributions de probabilité sont utilisables dans les réseaux bayésiens. Une comparaison très intéressante est menée dans Jordan (2004).

2.2 Démarche statistique

Il s'agit d'extraire l'information utile de l'observation d'un sous-ensemble des variables aléatoires du système, en d'autres termes de suivre une démarche statistique. Les deux voies classiques sont envisageables, celle de la statistique fréquentiste ou celle de la statistique bayésienne. Notre parti est de préférer la seconde option pour plusieurs raisons. En premier lieu parce que la complexité des modèles abordables en statistique bayésienne est maintenant plus grande qu'en statistique fréquentiste depuis l'avènement d'algorithmes efficaces mis en œuvre dans des logiciels très généralistes (cf. la section 4). Le nombre d'ouvrages généraux présentant cette démarche en détail ne cesse de croître, citons parmi eux Gelman et al. (2006), Robert (2006) et Carlin et Louis (2009). Et aussi parce que la rareté des données disponibles entraîne la non identifiabilité de certains paramètres du modèle rendant la démarche fréquentiste impraticable.

Le paradigme de la statistique bayésienne est assez naturel, et, si la notion de variable aléatoire est acquise, plus facile à proposer que celui de la statistique fréquentiste. Les données, Y, comme le vecteur des paramètres du modèle, Θ, sont envisagés comme un ensemble de variables aléatoires sur lesquelles une distribution de probabilité conjointe, numériquement déterminée, est posée. La marginale de Θ est dénommée *distribution de probabilité a priori* ; elle est censée synthétiser les connaissances dont on dispose, avant l'utilisation des données, sur les valeurs que peuvent prendre les paramètres du modèle. $\grave{A}$ l'image du terme anglais *prior*, nous la dénommons *priore*. La distribution conjointe sur Θ et Y est déterminée par les spécifications de la priore $[\Theta]$ et de la vraisemblance, distribution conditionnelle de Y relativement à Θ, $[Y \mid \Theta]$, par la simple application du théroème de Bayes :

$$[\Theta, Y] = [\Theta]\,[Y \mid \Theta]\,. \tag{2}$$

Dans ce cadre probabiliste, extraire l'information contenue dans les données se fait naturellement par conditionnement sur Y, et toutes les inférences s'opèrent à partir de la *distribution de probabilité a posteriori*, la *postériore* comme nous la nommerons, formellement obtenue par une seconde application du théorème de Bayes

$$[\Theta \mid Y] = \frac{[\Theta, Y]}{[Y]}. \tag{3}$$

Le passage de la priore à la postériore peut s'interpréter comme une étape d'apprentissage, la postériore d'une étape ayant naturellement vocation à devenir la priore de l'étape suivante...

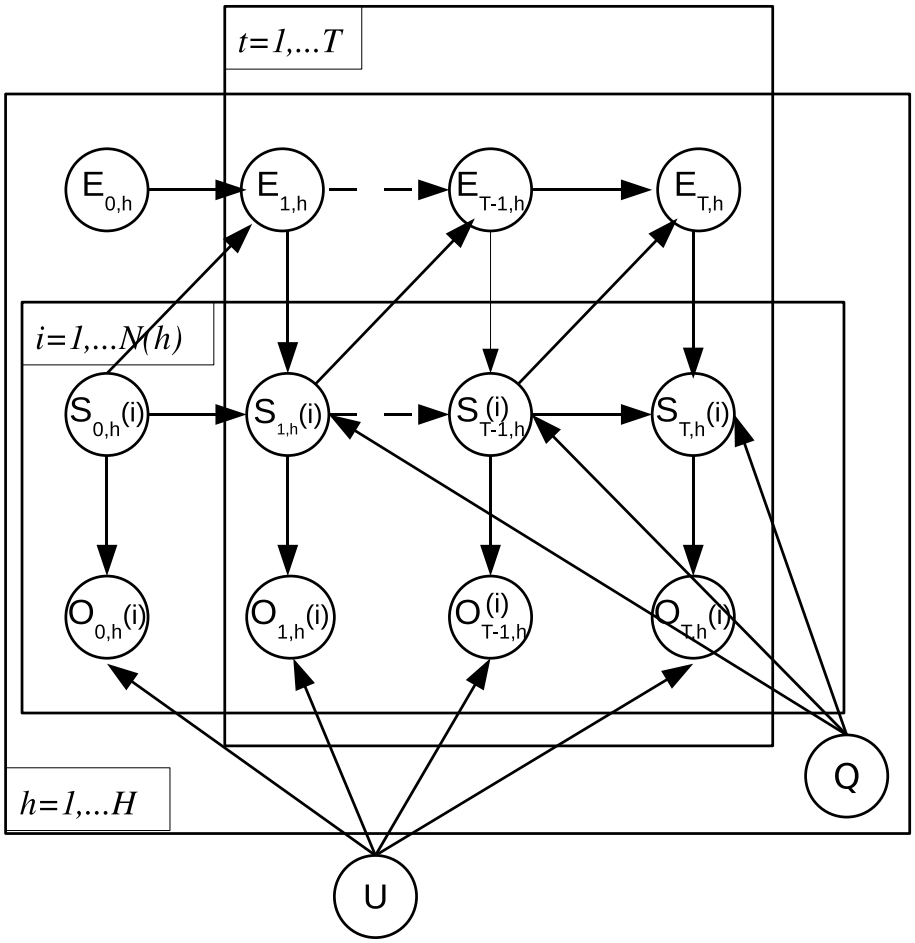

FIG. 3 – *Évolution d'une épidémie [Courcoul et al. (2009)] . Il s'agit d'une modélisation dynamique $t = 1, .., T$ d'épidémie sur les individus de plusieurs troupeaux $h = 1, ..., H$. Les variables impliquées sont les contaminations de l'environnement (*E*), l'état de chaque animal (*S*), l'observation de l'état de chaque animal (*O*), les paramètres de passage d'un état à l'autre pour un animal (*Q*) et une modélisation des erreurs d'observation (*U*).*

FIG. 4 – *Statistique bayésienne sous forme de réseaux bayésiens. Les nœuds des réseaux bayésiens peuvent être vectoriels. On définit la marginale sur* Θ *(priore) et la conditionnelle de* Y *par rapport à* Θ *(vraisemblance), c'est le réseau bayésien (a). Le réseau inversé (b) produit la conditionnelle de* Θ *par rapport à* Y *(postériore).*

Alors qu'en statistique fréquentiste, on considère Θ comme un paramètre prenant une valeur fixe inconnue, le statut de variable aléatoire que lui donne la statistique bayésienne renverse le point de vue. Ce n'est plus celui du statisticien qui accepte de se tromper dans 5% des cas, mais celui de l'obtenteur des données qui met à jour sa connaissance.

Notons enfin que la démarche bayésienne s'illustre parfaitement à l'aide des réseaux bayésiens à deux nœuds de la figure 4 (a). L'inférence n'est autre que le retournement de l'arc du réseau bayésien (ce qui est toujours possible ; figure 4 (b)) produisant la postériore : $[\Theta \mid Y = y]$.

2.3 Réseaux bayésiens et statistique bayésienne

Tels que présentés précédemment, les réseaux bayésiens et la statistique bayésienne répondent à des objectifs différents : définir un modèle probabiliste ou extraire de l'information d'un ensemble de données. Dans les pratiques diverses que l'on peut rencontrer les choses ne sont pas aussi tranchées : la distinction entre paramètres et variables cachées, variables latentes ou observées est souvent difficile. Par exemple, beaucoup d'utilisations des réseaux bayésiens consistent à conditionner les variables ciblées par l'*instanciation* (ou *évidenciation*) d'une ou de plusieurs autres variables. Les variables conditionnantes pourraient alors être considérées comme des données, même si la loi conjointe n'a jamais été formulée comme le produit d'une priore par une vraisemblance [équation (2)]. Mais souvent, il ne s'agit que de raisonnements "*what if ?*" : apprécier quels effets auraient l'occurrence de telle valeur pour telle variable du modèle ?

D'autre part, certains logiciels d'analyse statistitique bayésienne comme ceux de la famille BUGS [Plummer (2009) et Thomas (2009)] se servent - sans les nommer - de réseaux bayésiens pour définir les priores et les vraisemblances des modèles utilisés. C'est aussi le cas en statistique fréquentiste pour la définition de modèles hiérarchiques où les paramètres impliqués dans la vraisemblance sont des enfants d'hyperparamètres plus globaux.

2.4 Synthèse bayésienne

La proposition que nous faisons, détaillée par des exemples dans la section suivante, consiste en l'assemblage d'un réseau bayésien pour définir le modèle probabiliste qui servira de ressource pour définir les priores des paramètres de l'ensemble des données disponibles, et de l'inférence bayésienne qui servira à extraire l'information de l'ensemble des données dis-

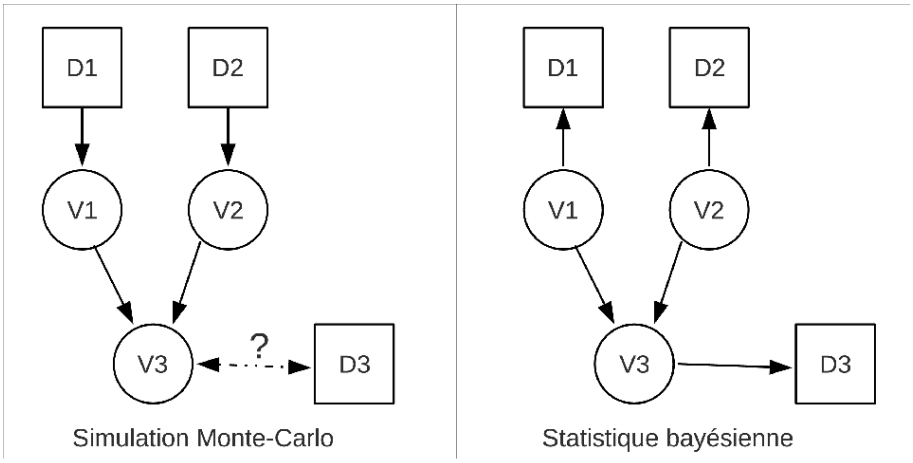

FIG. 5 – *Comparaison des démarches habituelle et proposée. L’optique Monte-Carlo ne fonctionne bien que lorsque les données se trouvent en amont de la modélisation. La statistique bayésienne permet tout positionnement d’un nombre quelconque de jeux de données [D1, D2 et D3].*

ponibles pour réactualiser les priores des paramètres. On pourrait objecter qu’il ne s’agit que d’un simple modèle hiérarchique bayésien. Dans de nombreux cas oui, mais formellement non car cela dépend de la manière dont les données sont rattachées au modèle central. Oui si les paramètres de la vraisemblance ne sont que des fonctions de variables du modèle central, non si des variables intermédiaires doivent être ajoutées pour tenir compte de biais entre le paramètre d’intérêt et le paramètre informé par les données, cf. Turner et al. (2009). De la même manière, on pourrait prétendre que le concept de réseau bayésien n’apporte rien puisqu’il ne s’agit que d’une distribution de probabilité conjointe sur un ensemble de variables aléatoires ! Au niveau des applications, il est cependant très précieux.

En appréciation quantitative des risques alimentaires et environnementaux de nombreuses études [Anonymous (1997)] se basent sur des simulations (dite *de Monte-Carlo*). Un réseau bayésien est défini et des estimations indépendantes de l’ensemble des paramètres à partir de données locales sont alors pratiquées. Elles permettent la génération de simulations des variables d’intérêt.

Notre démarche est aussi une alternative à ce genre de pratique. Elle offre la cohérence complète du cadre probabiliste global dans lequel elle se place. En effet, la pratique simulation de Monte-Carlo présente au moins deux inconvénients : (1) une estimation non globale des paramètres de la modélisation et (2) l’impossibilité de prendre en compte des données en aval de nœuds ancêtres (cf. la figure 5). La synthèse bayésienne proposée résoud naturellement ces deux points par une estimation globale basée sur le conditionnement des données modélisées à partir du modèle central.

3 Modélisations stochastiques appuyées par la statistique bayésienne

Nos idées se sont progressivement élaborées lors d'une longue étude menée pour modéliser l'influence de la consommation de poulets sur le nombre de campylobactérioses en France (Albert et al., 2008). C'est donc sur cet exemple que nous avons choisi d'illustrer notre proposition. Nous évoquerons ensuite plus rapidement le principe d'une modélisation épidémiologique qui fait aussi ressortir le principe de la démarche. Nous avons également utilisé l'approche proposée ici dans d'autres situations, par exemple celles présentées dans les figures 2 et 3.

3.1 Campylobacter et poulet

Les effets délétères sur une population humaine d'aliments contaminés par des bactéries pathogènes sont extrêmement difficiles à quantifier. Ils sont le résultat de l'enchaînement d'un grand nombre d'événements à la fois variables et mal connus. Jusqu'à présent, les mesures de gestion se basent sur des avis d'experts. L'appréciation quantitative des risques microbiologiques est une démarche qui tente de rendre la situation plus explicite et moins subjective.

La problématique générale de cet exemple est celle de l'évaluation du nombre de campylobactérioses en France provoquées, directement et surtout indirectement, par la consommation de poulets. La démarche suivie est de modéliser la chaîne du process comprenant l'élevage des poulets, leur transformation en carcasses, leur préparation dans les cuisines des consommateurs et enfin, les campylobactérioses qui peuvent s'ensuivre.

Les données disponibles sur le sujet se résumaient à peu de choses : des échantillonnages de tests de contamination (présence/absence) dans les élevages et les abattoirs, une enquête de consommation générale des produits alimentaires et une enquête épidémiologique en Grande Bretagne (où les comportements alimentaires sont assez différents de ceux de la France). Des informations supplémentaires nous étaient fournies par des experts vétérinaires et épidémiologistes qui connaissaient (mais de manière plutôt qualitative) la question. Les informations disponibles se résumaient bien à des données rares et hétérogènes, pas forcément appropriées et surtout à des connaissances subjectives, provenant d'experts.

Nous avons choisi de modéliser la chaîne alimentaire (de l'élevage à la maladie humaine) sur les idées *a priori* de nos experts, en orientant le modèle vers les variables d'intérêt et celles associables aux données dont nous disposions. Ceci s'est fait pas la construction d'un réseau bayésien présenté en figure 6 (a). Les liens stochastiques et déterministes qui relient les noeuds du réseau sont résumés dans le tableau 1, extrait de l'article [Albert et al. (2008)], pour permettre une meilleure compréhension du graphe et montrer la compexité de la modélisation globale.

L'approche proposée se décline en un certain nombre d'étapes simples successives, agrémentées de nombreux aller-retours : (1.a) inventaire et définition des variables du système qui seront les nœuds du réseau bayésien, (1.b) établissement des relations directes entre les nœuds, (1.c) spécification des distributions conditionnelles ou marginales des nœuds et (1.d) génération de valeurs de variables connues des experts pour apprécier la cohérence globale de la construction ; (2.a) extension du réseau bayésien pour inclure les données disponibles, (2.b) conditionnement du réseau bayésien par les valeurs observées et (2.c) génération des mêmes variables qu'en (1.d) pour apprécier la cohérence globale de la construction et comparer avec

TAB. 1 – *Extrait de Albert et al. (2008) : Prior distribution of the core model.*
Each line of the table defines a variate. This may involve some parents and can result from either a logical relationship ($=$) or a distribution ($\sim$). $N\left(m,s\right)$ stands for the normal distribution with expectation m and variance s^2. Possible truncation of the distributions are indicated by inequalities. Except for the normal distribution, the parameters defining distributions are those used in WinBUGS and Jags software programs. The variates are classified according to their status : variate of interest (vi), complementary variate (cv) and parameter (pa).

Class	**Variate**	**Parent(s)**	**Distribution / Relationship**
vi	p_f	m_f, s_f	$logit(p_f) \sim N(m_f, s_f)$
vi	p_b	m_b, s_b	$logit(p_b) \sim N(m_b, s_b)$
vi	p_h	p_{hc}, p_{hh}	$= p_{hc}p_{hh}$
vi	λ_c	a_c, b_c	$\sim Gamma(a_c b_c, b_c)$
vi	p_{ey}	p_e	$= 1-(1-p_e)^{13}$
vi	p_{ib}	p_{ey}, p_{ie}	$= p_{ey}p_{ie}$
vi	p_{it}	p_{ey}, p_{ib}, p_{iq}	$= p_{ib}/\left(1-(1-p_{iq})\left(1-p_{ey}\right)\right)$
cv	m_b	m_f, d_{bf}	$= m_f + d_{bf}$
cv	λ_e	p_b, λ_c, p_h	$= p_b\lambda_c p_h$
cv	p_e	λ_e	$= 1-\exp\left(-\lambda_e\right)$
cv	p_{ie}	d, p_{ne}, p_{in}	$= \left(1-(1-p_{ne})^d\right)p_{in}$
pa	m_f	-	$\sim N\left(0, 0.22\right)$
pa	s_f	-	$\sim Uniform\left(0, 0.2\right)$
pa	d_{bf}	-	$\sim N\left(0.1, 0.0696\right) > 0$
pa	s_b	-	$\sim Uniform\left(0, 0.2\right)$
pa	p_{hc}	-	$\sim Beta\left(8, 8\right)$
pa	p_{hh}	-	$\sim Beta\left(8, 28\right)$
pa	a_c	-	$\sim Gamma(4, 4)$
pa	b_c	-	$\sim Gamma(10, 10)$
pa	d	-	$\sim$ (1,2,10,100,300) with p = (0.5,0.163,0.222,0.097,0.018)
pa	p_{ne}	-	$\sim Beta\left(0.024, 0.011\right)$
pa	p_{in}	-	$= 0.33$
pa	p_{iq}	-	$\sim Beta\left(9, 30\right)$

les résultats précédemment obtenus. Plus de détails se trouvent dans Albert et al. (2008). Du point de vue de la statistique bayésienne, les étapes (1) correspondent à la définition des paramètres et de la priore, et les étapes (2) à la définition de la vraisemblance et au calcul de la postériore.

La figure 6 (a) présente la structure du modèle des étapes (1). Sans entrer dans les détails, on constate que le processus est modélisé en grands modules (la contamination par la bactérie est suivie de la ferme à la cuisine en passant par l'abattoir ; l'occurrence de la maladie dépend de la consommation et de la dose-réponse choisie). Cette modélisation est fort ambitieuse, elle rassemble dans une même formalisation des connaissances zootechniques, agro-industrielles, comportementales, économiques, médicales, épidémiologiques... mais elle est aussi rudimentaire puisque chacune est résumée par quelques variables aléatoires. Elle présente cependant l'avantage de proposer un formalisme unique conduisant les spécialistes impliqués vers un référentiel commun. L'élicitation des priores par les experts est à la fois fondamentale et difficile. L'usage du graphe du modèle et de simulations de variables observables sont deux appuis majeurs pour la modélisation.

On comprendra dans cette perpective d'utilisation de la statistique bayésienne, que l'optique n'est pas de rechercher des priores les moins informatives possibles, mais au contraire des priores qui soient les plus fidèles aux idées des experts.

La figure 6 (b) présente la structure du modèle complété pour y intégrer les données. Il apparaît que la définition du modèle initial n'est pas complètement indépendante des données disponibles.

3.2 Modèle épidémiologique

Une difficulté critique de l'appréciation quantitative des risques microbiologiques est l'établissement de la fonction dite "dose-réponse", c'est-à-dire l'expression de la probabilité d'occurrence de la maladie en fonction de la dose de pathogènes ingérée. Pour des raisons éthiques, des expérimentations ne peuvent être entreprises pour des maladies graves et/ou sur des sujets en mauvaise santé. Le recours à l'expérimentation animale présente de forte limite. Une approche alternative consiste à utiliser les données épidémiologiques humaines. Là encore la complexité et la diversité des événements en cause, tout comme la rareté et l'hétérogénéité des données utilisables, rendent l'approche quantitative ardue.

Dans une autre étude en cours d'achèvement [Albert et al. (2009)], nous avons tenté de regrouper l'ensemble des données épidémiologiques relatives à la campylobactériose disponibles sur la France métropolitaine ou de contextes assimilables (selon les experts) pour estimer la prévalence de cette maladie en France. Le DAG du modèle est donné en figure 7.

Les données disponibles présentent la particularité d'être principalement des échantillonnages pour des aspects indirectement liés à la détection de l'événement qui nous intéresse (occurrence de la campylobactériose pour un individu de la population française). Par exemple :

- nombre de personnes ayant eu une campylobactériose et ayant consulté un médecin parmi des personnes ayant été exposées (r_4 et n_4) ;
- nombre de personnes ayant eu une campylobactériose et ayant consulté un médecin parmi des personnes ayant eu une gastro-entérite aïgue (r_{8i} et n_{8i}) ;
- nombre de personnes ayant eu une gastro-entérite aïgue parmi des personnes exposées (r_{3i} et n_{3i}) ;

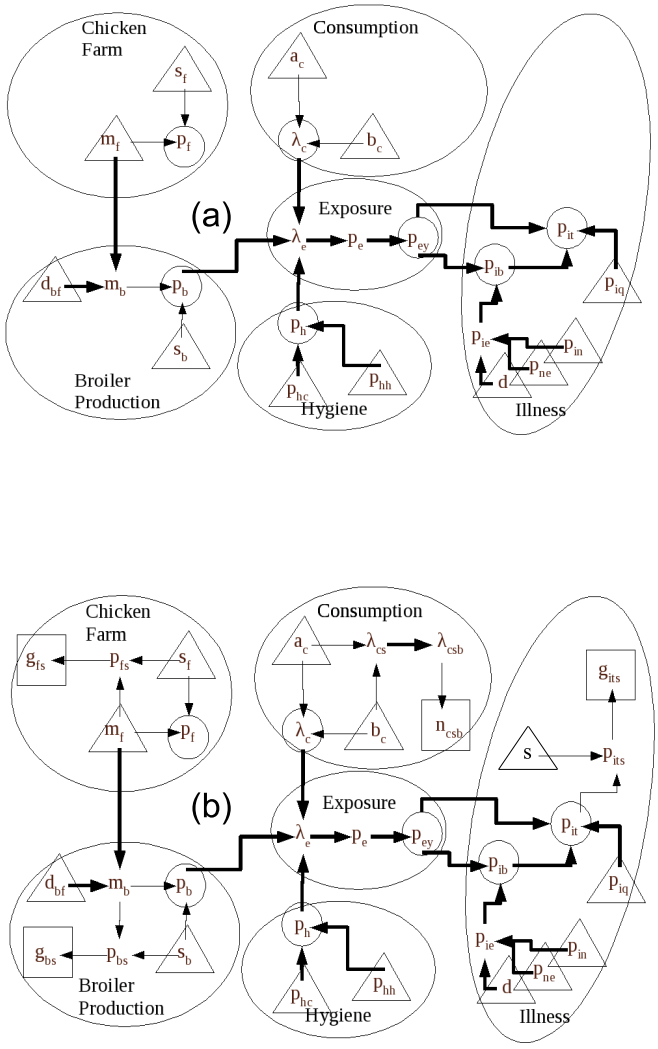

FIG. 6 – *(a) Modèle central de l'occurrence de campylobactérioses dues à la consommation de poulets en France. Il s'agit d'un réseau bayésien structuré selon les six grands secteurs de la chaîne alimentaire. (b) Modèle central complété pour intégrer les données disponibles (nœuds carrés). On observera que deux des six secteurs ne comportent pas de données. Les paramètres clef de la modélisation sont :* p_f*, la probabilité qu'un élevage de poulets soit contaminé ;* p_b*, la probabilité qu'une carcasse de poulet soit contaminée ;* p_h*, la probabilité que l'hygiène de la cuisine d'un ménage soit mauvaise ;* λ_c*, l'intensité de consommation de poulet par un ménage ;* p_e*, la probabilité qu'un membre du ménage ingère une bactérie pathogène (exposition) et* p_{it}*, la probabilité qu'un individu de la population attrape une campybactériose. Les distributions de probabilité associées au modèle central sont données dans le tableau 1.*

– nombre de coprocultures positives à campylobacter parmi des coprocultures analysées (r_6 et n_6) ;
– nombre de personnes ayant eu des coprocultures parmi des personnes ayant eu une gastro-entérite aïgue (r_{1i} et n_{1i}) ;
– et ainsi de suite.

Toutes ces données relèvent, au moins approximativement, d'un schéma binomial ou trinomial. Pour cette étude, le modèle central est basé sur une partition fine (dans le sens où tous les cas de figure recensés par les données sont couverts) de la population générale à laquelle on associe un modèle multinomial pour chaque individu ou sous-ensemble d'individus. Le vecteur, P, de ses probabilités est le point de départ de la modélisation comme le manifeste la figure 7.

Une étude du modèle et des données fait apparaître que la dimension paramétrique du modèle central vaut 11 alors que seules 9 fonctions de ses paramètres sont identifiables. Heureusement la probabilité cible de l'étude est bien identifiable. La statistique bayésienne permet d'échapper à une reparamétrisation délicate, l'inférence pouvant se faire sous le couvert de la priore placée sur l'ensemble des paramètres du modèle.

3.3 Apport de la synthèse bayésienne

Pour les deux exemples présentés, la synthèse bayésienne a permis de conduire une inférence globale cohérente et finalement assez directe. La figure 8 présente un exemple de résultats obtenus dans l'application campylobacter-poulet introduite en section 3.1. On peut y constater (i) que la relation entre les deux variables d'intérêt était déjà assez finement précisée au moment de l'établissement de la priore et (ii) que l'introduction des données apporte une modification sensible de l'idée de l'expert sur leur loi conjointe. C'est le résultat de la focalisation sur une modélisation initiale non asservie à des données particulières bien que couvrant l'ensemble des données disponibles. Dans cette même application, la synthèse bayésienne a permis l'introduction de données en aval des nœuds ancêtres (voir figure 6 (b)) : les données g_{its} sont introduites sur un nœud ancêtre) ; ce qui n'aurait pas été possible avec une approche classique dite de Monte-Carlo (voir figure 5 et section 2.4). Dans l'application épidémiologie (voir section 3.2), la synthèse bayésienne permet une méta-analyse ambitieuse de données non directement reliées aux paramètres d'intérêt, sans la contrainte stricte qu'engendrerait la non-identifiabilité du modèle global.

Bien entendu, la mise en œuvre de telles démarches n'est possible que par la disponibilité d'algorithmes sophistiqués de simulation, discutée dans le paragraphe suivant. Outre le résultat immédiat d'inférence sur des paramètres d'interprétation directe, la démarche oblige pour la construction du modèle central à prendre un recul bénéfique sur les objectifs poursuivis par la mise en lumière des paramètres clef, sur les mécanismes et causalités sous-jacents par l'établissement des relations, le plus souvent empiriquement modélisées, qui les relient.

La modélisation centrale sert de pivot à toute la réflexion. L'interprétation des résultats au travers de l'examen des postériores (et post-dictions) reste naturelle et facile puisqu'elle est parallèle à l'établissement de la priore concrétisée par la modélisation centrale.

Toutes ses qualités nous semblent manquer en approches statistiques fréquentistes, et nous suspectons, ne serait-ce que pour des questions d'identifiabilité des paramètres, qu'elles auraient été impossible à réaliser au même niveau de globalité pour ces deux illustrations.

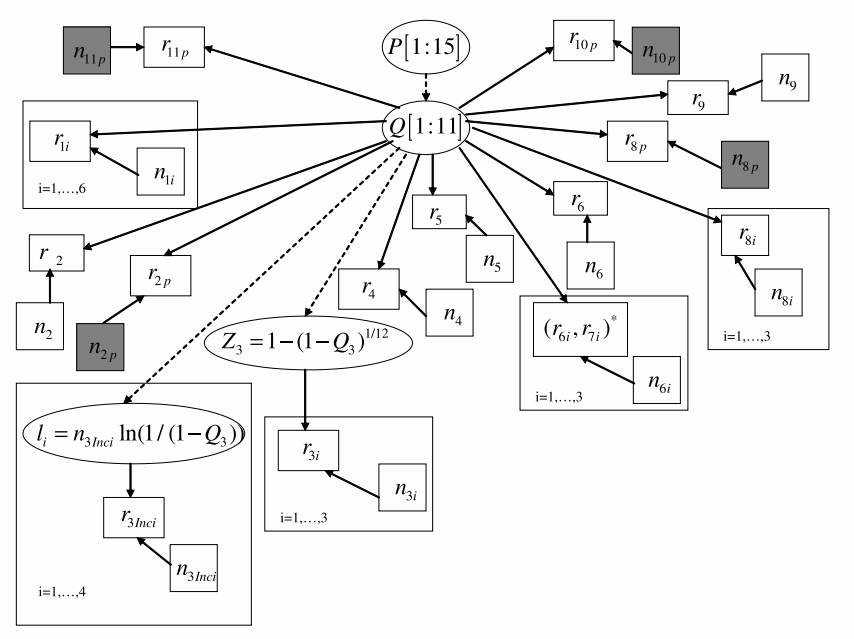

FIG. 7 – *Les probabilités fonctionnellement indépendantes de la partition complète de la population sont figurées par le vecteur P, le vecteur Q en est une transformation fonctionnelle qui prend en compte les zéros structurels de P. Ce vecteur Q de 11 composantes libres est le modèle central de cette approche, à partir duquel toutes les vraisemblances des données disponibles (entourées d'un rectangle) se déduisent soit directement (flèches continues), soit au travers d'une transformation fonctionnelle (flèches tiretées). Les données grisées correspondent à une élicitation opérée par les experts épidémiologistes. Toutes les données (et pseudo-données associées aux dires d'expert) suivent un schéma binomial (ou trinomial dans un cas). Elles se réduisent donc à une taille, n_{ki}, et un nombre de cas positifs, r_{ki}, où k est le numéro du type de données et i celui de l'étude hiérarchisée dans k.*

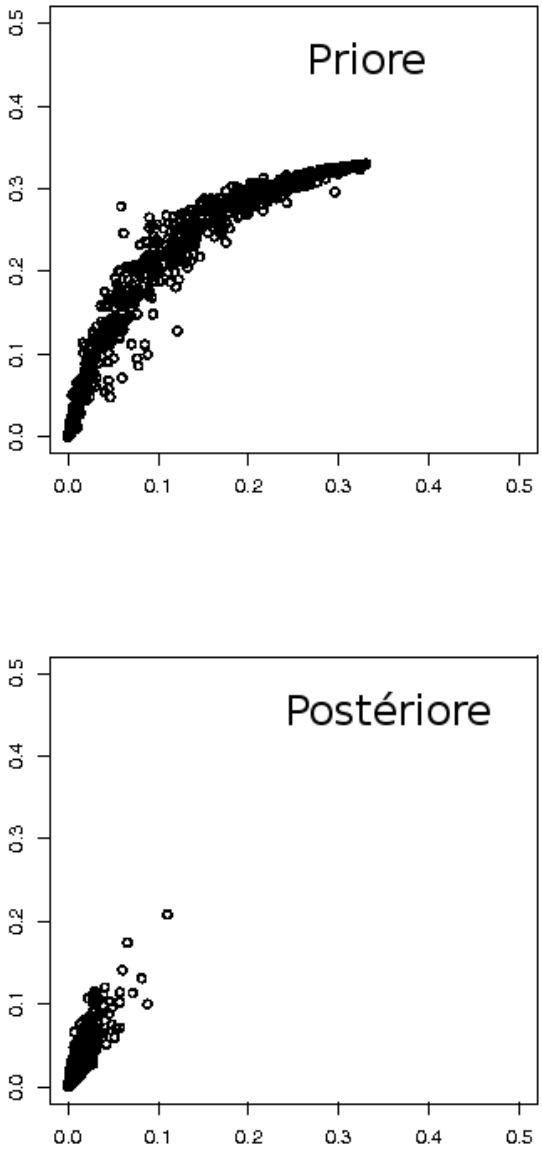

FIG. 8 – *Simulation des distributions priore et postériore de deux variables d'intérêt de l'étude campylobacter-poulet (p_{ib} en abscisses, p_{it} en ordonnées).*

4 Logiciels utiles

Dans la plupart des cas, la démarche proposée peut se réaliser avec les logiciels de la famille BUGS [Plummer (2009) et Thomas (2009)]. Utilisant une syntaxe proche du codage d'une programmation R [R Development Core Team (2008)], ils autorisent le codage de modèles très complexes en peu de lignes. La variété des distributions de probabilité disponibles est grande, elle inclut un certain nombre de distributions multivariées continues (Normale, Student, Wishart, Dirichlet) et discrète (multinomiale). Les algorithmes de simulation MCMC qu'ils mettent en œuvre sont, comme il est écrit en rouge en première page de leur aide, délicats à manier et l'utilisateur doit connaître un minimum du fonctionnement des échantillonneurs mis en œuvre pour l'obtention de résultats valides. Mention doit être faite de la possibilité d'utiliser ces logiciels depuis le langage R grâce à des paquets adaptés pour définir les entrées ou pour traiter les échantillons MCMC obtenus.

Une gêne que peut rencontrer l'utilisateur de ces logiciels pour des modèles sophistiqués est qu'il est très difficile de s'assurer que le modèle qu'on a spécifié est bien le modèle que l'on voulait définir. Ce n'est pas un langage de programmation, on ne peut donc pas effectuer de sorties intermédiaires ou capitaliser des instructions à l'aide de sous-routines. C'est l'origine du projet ReBaStaBa (Réseaux Bayésiens traités par Statistique Bayésienne) [Denis et Albert (2008) et Denis (2010)], paquet R en cours d'élaboration, où les réseaux bayésiens sont définis par des objets R associés à des fonctions pour en visualiser les propriétés (impression, représentation graphique, exploration des parentés,...), les manipuler (construction de sous-réseaux bayésiens, modification de la distribution de probabilité de quelques nœuds,...), les exporter dans des formats utilisés par d'autres paquets centrés sur les réseaux bayésiens comme Rjags, Deal ou Grappa.

5 Pour conclure

Bien que la démarche proposée nous semble maintenant simple et que nous ayons éprouvé plusieurs fois la clarification qu'elle apporte dans diverses situations, des discussions impromptues sur ce sujet avec des collègues nous ont montré que ces idées n'étaient pas si évidentes, c'est pourquoi nous avons tenté d'exposer nos arguments dans cet article.

En résumé, la proposition consiste, avant l'analyse statistique des données, à formaliser les connaissances disponibles par une distribution de probabilité sur les variables concernées. Les réseaux bayésiens sont un moyen commode pour y parvenir, mais on pourrait imaginer d'autres alternatives, par exemple des réseaux markoviens non-orientés. La formalisation obtenue sert alors de base à la paramétrisation de la vraisemblance des données disponibles. Disposant d'une distribution de probabilité conjointe complète, il semble logique de mettre en place une analyse statistique bayésienne, mais là encore on pourrait mettre en œuvre une inférence par statistique fréquentiste. Ainsi définie la démarche proposée nous semble beaucoup plus riche et plus large que la simple interprétation d'un tableau de données par un modèle hiérarchique bayésien comme on peut la trouver dans Spiegelhalter à propos de l'hépatite B [Spiegelhalter et al. (1996)].

Bien sûr, les choses ne sont jamais aussi tranchées en pratique et on commence souvent à entreprendre la définition du modèle central en fonction des données disponibles. Néanmoins,

l'effort qui consiste à mettre au point un modèle qui tienne en lui-même est un bon moyen pour prendre du recul vis-à-vis de l'information disponible.

Une des difficultés majeures de la démarche est sans doute celle de l'obtention de la distribution a posteriori des paramètres par les méthodes MCMC dans un contexte qui peut être très multivarié avec de fortes corrélations et inter-dépendances entre les paramètres. La convergence de l'algorithme peut alors être très lente. Le principe de parcimonie doit alors être appliqué au modèle pour une meilleure efficacité de la démarche.

Cette proposition de synthèse bayésienne nous semble tout de même comporter un certain nombre d'avantages décisifs :

- la statistique bayésienne est assez naturelle pour des experts peu rodés à la statistique, par exemple ils interprètent souvent les intervalles de confiance comme des intervalles de crédibilité. La mise en place d'une modélisation globale préalable permet (par des simulations si les calculs analytiques ne sont pas possibles) de montrer aux experts par des analyses de sensibilité, les conséquences de leurs dires sur des parties éloignées de ce qu'ils sont en train d'apprécier. On peut ainsi leur proposer un outil d'aide à élicitation.
- le passage d'une priore à une postériore qui peut elle-même être utilisée comme future priore, est assez cohérent avec la démarche scientifique générale d'aggrégation progressive des connaissances au fur et à mesure de la disponibilité des données.
- la focalisation sur le modèle central permet de rattacher des données aussi hétérogènes soient-elles dans leurs origines, ouvrant la voie à des méta-analyses statistiques cohérentes.

Il s'agit d'un changement de point de vue : l'objet principal n'est plus le tableau de données disponibles, mais une représentation cohérente de la réalité au travers d'une modélisation probabiliste réactualisable.

Remerciements

Nous voudrions remercier les relecteurs de la première version de ce papier pour les critiques pertinentes qu'ils ont formulées. Elles nous ont conduit à mieux placer notre démarche dans une perspective générale.

Références

Albert, I., E. Espié, A. Gallay, H. De Valk, E. Grenier, et J.-B. Denis (2009). Bayesian statistical meta-analysis of epidemiological data for QRA. In Martorell et al. (Ed.), *Safety, Reliability and Risk Analysis : Theory, Methods and Applications*, pp. 2609–2612. Taylor & Francis Group. Proceedings of ESREL'08 (Valencia, Spain), 22-25 September 2008 ; organized by the European Safety and Reliability Association and the Society for Risk Analysis.

Albert, I., E. Grenier, J.-B. Denis, et J. Rousseau (2008). Quantitative risk assessement from farm to fork and beyond : A global bayesian approach concerning food-borne diseases. *Risk Analysis 28*(2), 557–571.

Anonymous (1997). Guiding principles for monte carlo analysis (EPA/630/R-97/001). Technical report, Environmental Protection Agency (http ://www.epa.gov/NCEA/pdfs/montcarl.pdf).

Bishop, C. M. (2006). *Pattern recognition and machine learning (chapter 8)*. New York: Springer.

Carlin, P. B. et T. A. Louis (2009). *Bayesian methods for data analysis*. Boca Raton: CRC Press.

Courcoul, A., E. Vergu, J.-B. Denis, et F. Beaudeau (2009). Q fever: a Bayesian approach for a dynamic epidemiological model applied to observations on several herds. In *Proceedings of the International Society for Veterinary Epidemiology and Economics (ISVEE12)*, Durban, South Africa.

Darviche, A. (2009). *Modeling and reasoning with Bayesian networks*. Cambridge: Cambridge University Press.

Denis, J.-B. (2010). Rebastaba, réseaux bayésiens pour la statistique bayésienne. Page de présentation et de téléchargement, INRA.

Denis, J.-B. et I. Albert (2008). Construction d'un paquet R pour la manipulation de réseaux bayésiens en vue d'une inférence par statistique bayésienne. In *Actes des 4èmes Journées Francophones sur les Réseaux Bayésiens (JFRB 2008)*, Lyon, France, pp. 144–154.

Edwards, D. (2000). *Introduction to graphical modelling*. New York: Springer.

Gelman, A., J. B. Carlin, H. S. Stern, et D. B. Rubin (2006). *Bayesian data analysis*. New York: CRC Press.

Jensen, F. V. (2001). *Bayesian networks and decision graphs*. New York: Springer.

Jordan, M. I. (2004). Graphical models. *Statistical Science 19*, 140–155.

Lauritzen, S. L. (1996). *Graphical models*. Oxford: Clarendon Press.

Lauritzen, S. L. et D. J. Spiegelhalter (1988). Local computations with probabilities on graphical structures and their application to expert systems. *J. Royal Statistics Society B 50*, 157–194.

Marin, J.-M. et F. Rossi (2004). Découvrez les réseaux bayésiens. *Linux Magazine 60*, 56–65.

Pearl, J. (2000). *Causality. Models, reasoning, and inference.* Cambridge: Cambridge University Press.

Plummer, M. (2009). Jags, just another gibbs sampler. Page de présentation, IARC (http://www-fis.iarc.fr/ martyn/software/jags/).

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Robert, C. P. (2006). *Le choix bayésien : principe et pratique*. New York : CRC Press.

Spiegelhalter, D. J., N. G. Best, W. R. Gilks, et H. Inskip (1996). Hepatitis B: a case study in MCMC methods. In W. R. Gilks, S. Richardson, et D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice*, pp. 21–43. Chapman & Hall.

Thomas, A. (2009). Openbugs, bayesian inference using gibbs sampling. Page d'accueil, (http://mathstat.helsinki.fi/openbugs/).

Turner, R., D. Spiegelhalter, G. Smith, et S. Thompson (2009). Bias modelling in evidence synthesis. *J.R. Statist. Soc. A 172*(2), 21–47.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Chichester: John Wiley & Sons.

Summary

Applied statisticians involved in a field where data are rare and heterogeneous, the authors found help from two complementary but independent Bayesian approaches : the setup of a Bayesian network to model the phenomemum under study and the use of a Bayesian statistical procedure to get information from the available data sets where the Bayesian network is used to define the prior. This global and generic approach is presented through some case-studies. Contrary to the standard approach of the so-called Monte Carlo simulation, data sets can be used at every point of the chain, even downstream of other data sets. R and softwares of the BUGS family are well adapted to perform these kinds of computation.