REVUE DE STATISTIQUE APPLIQUÉE

ÉRIC PARENT
ÉTIENNE PREVOST

Inférence bayésienne de la taille d'une population de saumons par utilisation de sources multiples d'information

Revue de statistique appliquée, tome 51, n° 3 (2003), p. 5-38
<http://www.numdam.org/item?id=RSA_2003__51_3_5_0>

Rev. Statistique Appliquée, 2003, **LI** (3), 5-38

INFÉRENCE BAYÉSIENNE DE LA TAILLE D'UNE POPULATION DE SAUMONS PAR UTILISATION DE SOURCES MULTIPLES D'INFORMATION

Éric PARENT[1], Étienne PREVOST[2]

[1] *GRESE (Laboratoire de Gestion du Risque en Sciences de l'Environnement), ENGREF*
19, Avenue du Maine, F-75732 Paris Cedex 15 Parent@engref.fr
[2] *UMR Ecobiologie et Qualité des Hydrosystèmes continentaux, INRA*
65, rue de St Brieux, 35042 Rennes Cedex, Prevost@roazhon.inra.fr

RÉSUMÉ

Améliorer la gestion des stocks de saumons nécessite d'évaluer correctement le nombre total de poissons adultes qui reviennent à leur rivière d'origine pour frayer. L'inférence Bayésienne et les techniques d'estimation associées de simulation Monte Carlo par Chaines de Markov (MCMC) libèrent le modélisateur des recettes de statistique développées pour construire des estimateurs de stocks. Trop souvent, la vertu essentielle de modèles *ad hoc* (et souvent simplistes) n'a été que de mener à des expressions analytiques des estimateurs de stocks. Débarrassé de cette contrainte, le modélisateur peut aujourd'hui porter toute son attention sur le réalisme des modèles et sur l'étude des nombreuses sources d'incertitudes. L'analyse Bayésienne se présente comme un cadre de raisonnement cohérent dans lequel des inférences écologiques peuvent être bâties à partir de types variés d'informations. Ces informations peuvent être objectives et provenir d'un jeu réduit de données ou de rapports, même peu précis, de capture par pêcheurs amateurs. L'analyse Bayésienne s'appuie aussi sur des sources d'informations subjectives provenant d'expertise sur l'écosystème étudié ou sur le niveau de ressources halieutiques.

Nous montrons qu'un graphe acyclique orienté est l'outil commode de représentation des différents événements qui peuvent se produire dans une population de saumons qui remontent la rivière Scorff, utilisée comme cas d'étude. Trois types de quantités apparaissent sur cette formalisation graphique par raisonnement conditionnel : les observables, les variables latentes et les paramètres. Les techniques MCMC permettent d'atteindre facilement les distributions de probabilités *a posteriori* qui traduisent un degré de crédibilité pour le nombre de castillons et pour les autres paramètres inconnus. Les dispersions de ces distributions se réduisent fortement quand on intègre à l'analyse statistique des informations additionnelles (même de nature subjective).

Mots-clés : *Modèle graphique, Algorithme de Gibbs, Variable latente, Inférence Bayésienne, Estimation de la taille d'une population de saumons.*

ABSTRACT

Evaluating the total number of adults returning to their native river to spawn is a major issue in salmon fisheries stock assessment. Bayesian inference with Monte Carlo Markov Chain (MCMC) techniques frees the modeler from working with *ad hoc* (and often oversimplified) stock assessment models for the only reason that they led to analytical solutions. Attention can

be recentered on the realism of the model with regards to the fish behavior and on the study of the various sources of uncertainty. Bayesian analysis offers a coherent framework in which ecological inferences can be grounded on various kinds of observations, even with a limited number of data or with imprecise catch reports from anglers, as well as on subjective sources of information such as the expertise about the local ecosystem and fishery.

On the Scorff case study, we point out that a directed graphical model is convenient to describe the various events that may occur in a salmon population. This conditional reasoning formalization offers a graphical representation of the way that 3 types of quantities interfere : observed variables, latent variables and technical parameters. Bayesian posterior distributions for the number of spawners and for other basic parameters can be easily derived from MCMC techniques. Their dispersions are greatly reduced when additional information (even subjective) is brought into the analysis.

Keywords : *Graphical modeling, Gibbs sampler, Latent variable, Bayesian inference, Salmon stock assessment.*

1. Introduction

Les programmes de recherche sur la dynamique des populations de saumons et les stratégies de gestion des stocks s'appuient essentiellement sur la connaissance de la quantité de saumons adultes qui reviennent à leur rivière d'origine pour frayer. Les scientifiques et les responsables ont besoin non seulement de l'estimation de la taille de la population (la valeur la plus probable) mais aussi de l'estimation de l'incertitude la concernant. La précision avec laquelle est évaluée la taille de la population permet d'asseoir la fiabilité du savoir scientifique acquis ou de proposer des stratégies raisonnables de gestion.

Dans un environnement naturel donc incontrolé, la taille d'un stock de saumons ne s'évalue que par estimation statistique. Des techniques statistiques sont utilisées couramment pour estimer les tailles de population de poissons au travers de modèles expérimentaux : méthodes de capture/ marquage /recapture (Clobert and Pradel, 1993), prélèvements successifs etc... Seber (1982) donne une revue de la littérature sur les méthodes et les modèles. Malheureusement de nombreux problèmes d'estimation de la taille des stocks ne correspondent pas aux modèles académiques : les hypothèses fondatrices ne sont pas vérifiées, les conditions asymptotiques requises pour dériver les estimations ne correspondent pas aux données rassemblées et finalement toute l'information disponible n'est pas utilisée. L'expérience de marquage/recapture sur la rivière Scorff, présentée dans la première partie de cet article illustre bien tous ces problèmes. Des pertes imprévues surviennent avant la recapture des individus marqués (l'hypothèse d'une population fermée est manifestement violée). Peu de poissons marqués sont recapturés de telle sorte que les conditions d'utilisation de l'estimation de la variance basée sur les propriétés asymptotiques ne sont pas remplies. Au-delà du nombre de poissons marqués et recapturés, l'information disponible se trouve dans les connaissances locales de l'écosystème et de la pêche, la littérature scientifique, les observations des années précédentes, les informations complémentaires rassemblées sur place par les techniciens et les pêcheurs (découverte de poissons morts, déclaration de capture, etc.). La méthode d'évaluation statistique standard ignore ces sources d'information et considère chaque estimation annuelle comme si c'était la première expérience jamais conduite...

Le paradigme Bayésien offre un champ de réflexion fertile pour le «détective écologique» (Bernier *et al.*, 2000). Le processus d'inférence Bayésienne est développé dans la seconde partie de cet article. Il s'appuie sur un modèle de probabilité qui reproduit le comportement du poisson. Les données sont rassemblées dans des procédures d'observations qui ne sont pas toujours conventionnelles. Un tel modèle fait la distinction entre quantités observables et non observables. Les quantités non-observables incluent à la fois des paramètres (coefficients exprimant les caractéristiques permanentes de la population de saumon) et des variables latentes (quantités «physiques» qui n'ont pas été observées pour une raison ou pour une autre). Une représentation graphique du modèle basé sur le raisonnement conditionnel en facilite la compréhension, la communication et le traitement Bayésien. En résumé, l'inférence Bayésienne a pour objet de calculer la loi de probabilité des quantités inconnues, telle que la taille du stock, sous la forme d'une fonction des quantités observées. À l'inverse de l'approche utilisée en statistique fréquentiste, l'analyse Bayésienne permet d'intégrer toutes les sources disponibles d'information. Par conséquent, l'incertitude qui porte sur l'abondance du stock et sur d'autres paramètres intéressant le biologiste, comme le taux de survie et le rendement de la capture, peut s'en trouver réduite. De récents progrès dans le calcul Bayésien, les algorithmes markoviens de simulation Monte Carlo Markov (MCMC) rendent facile l'estimation Bayésienne. Parmi les techniques MCMC, l'échantillonnage de Gibbs est particulièrement adapté quand le raisonnement conditionnel est exprimé sous la forme d'un graphe acyclique. La seconde partie de l'article montre que travailler avec la modélisation graphique et les techniques MCMC bouleverse les étapes habituelles de la modélisation statistique, fut-elle classique ou Bayésienne (Bernardo and Smith,1994) : il n'est plus besoin de rechercher la vraisemblance, ni même d'écrire explicitement la formule de Bayes pour obtenir la loi *a posteriori* des quantités inconnues.

Les résultats numériques de l'analyse Bayésienne sont présentés dans la troisième partie de cet article. On trouve que l'on peut facilement calculer les intervalles de crédibilité pour le nombre de reproducteurs et les autres paramètres d'intérêt et que la largeur de ces intervalles se réduit sensiblement quand on intègre toutes les sources d'information. Finalement les perspectives et les limites des procédures MCMC, des variables latentes et de l'analyse Bayésienne sont discutées à la lumière de ce cas illustratif.

2. Présentation du problème

2.1. Les trois dernières étapes du cycle de vie du saumon : remonter la rivière, échapper aux pêcheurs à la ligne et survivre jusqu'à la saison du frai

Les saumons atlantiques (*Salmo salar*) qui reviennent adultes dans les rivières de Bretagne (France), sont répartis en deux catégories : le saumon de printemps qui a passé deux années en mer (exceptionnellement trois) et les castillons qui reviennent dans leur rivière natale l'année qui suit leur migration vers la mer.

Les castillons constituent l'essentiel des adultes (approx. 90 %) qui reviennent dans la rivière, principalement de la fin du printemps à la première moitié de l'été. Sur la rivière Scorff un dispositif expérimental de contrôle des migrations a été installé

et les adultes de retour sont dénombrés par la technique du marquage/recapture. Le marquage est opéré dans un dispositif de piégage situé à l'embouchure de la rivière. L'efficacité du piège varie selon le débit de la rivière. La période de retour des saumons de printemps correspond à un débit plus haut que pour les castillons. Par conséquent, l'estimation des retours se fait séparément pour les saumons de printemps et pour les castillons. L'étude de cas présentée ici ne traite que du retour des castillons. La figure 1 décrit le sort d'un saumon rentrant dans sa rivière d'origine après son voyage dans l'Atlantique. Trois événements principaux peuvent arriver au candidat reproducteur. Premièrement, à l'entrée dans le Scorff, le saumon peut-être capturé, marqué et relaché (première étape de la procédure d'estimation du stock). Puis, une certaine quantité d'individus (marqués ou non) sera prélevée par les pêcheurs à la ligne. La loi Française exige que la prise de saumon soit officiellement déclarée, mais cette ordonnance n'est pas toujours respectée. Une étude locale supplémentaire permet de compléter ces renseignements. Ces deux sources permettent d'otenir une première évaluation du nombre de saumons «réellement capturés» et un certain nombre de saumons prélevés est apporté aux techniciens de l'Institut National de la Recherche Agronomique pour identification du marquage. En fin de compte, le poisson qui a échappé à la pêche à la ligne devra survivre jusqu'à la saison de reproduction. Pendant le frai hivernal, les chercheurs se rendent sur les sites de reproduction et complètent les études statistiques par une phase de recapture.

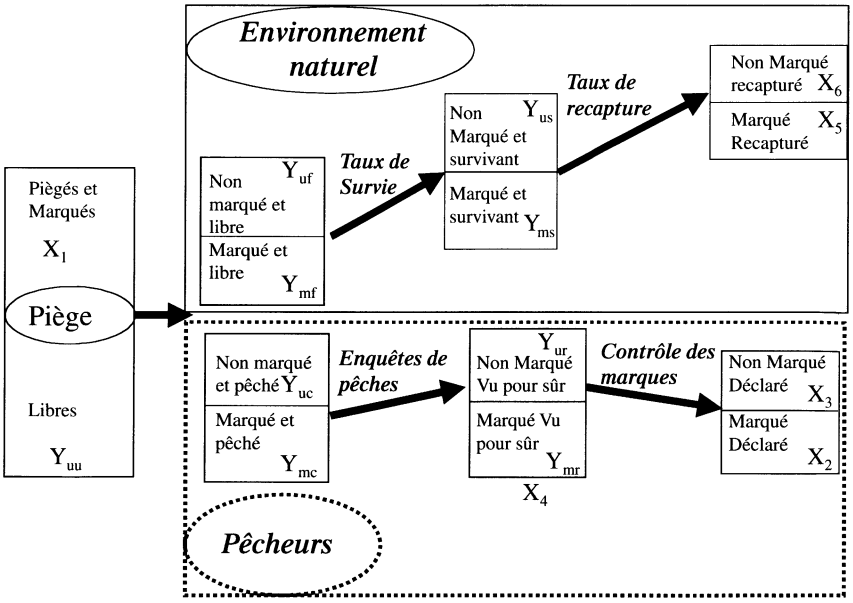

FIGURE 1
Le destin d'un saumon qui revient remonter le Scorff

2.2. Variables observées

Les données de la table 1 concernent six années (une par ligne) et six variables (en colonnes). Les données de la première année (1994) sont exclues de l'étude car elles sont significativement différentes des autres. L'efficacité du piège et la recapture au moment du frai ont été moins bonnes.

Les variables observées portent les informations suivantes :

X_1 : Nombre d'individus capturés, marqués et relachés,

X_2 : Nombre de poissons marqués, péchés à la ligne et rapportés par les pêcheurs pour la détection du marquage,

X_3 : Nombre de poissons non marqués, péchés à la ligne et rapportés par les pêcheurs pour la détection du marquage

X_4 : Total des poissons marqués et non marqués provenant d'observations sur les sites de pêche (par conséquent $X_4 > X_2 + X_3$)

X_5 : Nombre de poissons marqués qui survivent et sont recapturés pendant ou après le frai,

X_6 : Nombre de poissons non marqués, recapturés pendant ou après le frai.

TABLE 1
Données du Scorff

Année	X_1	X_2	X_3	X_4	X_5	X_6
1994	156	3	14	42	4	14
1995	500	39	10	75	31	28
1996	502	25	8	87	45	14
1997	320	17	7	33	19	9
1998	442	50	5	66	56	13
1999	167	16	4	24	16	11

2.3. Modèle stochastique du comportement du saumon

2.3.1. Les paramètres techniques sont inconnus mais supposés stationnaires

De telles quantités inconnues doivent être introduites au cours de la modélisation, pour représenter les caractéristiques stochastiques qui régissent le comportement individuel. Ces quantités sont censées rester identiques d'un poisson à l'autre. Les paramètres techniques suivants, inconnus, sont conceptuellement essentiels pour les biologistes :

κ : Nombre de castillons qui remontent le courant,

θ : Probabilité qu'un castillon soit capturé et marqué, au passage du piège,

α : Probabilité qu'un saumon non pêché survive jusqu'à la période de reproduction,

β : Probabilité qu'un castillon soit prélevé par les pécheurs,

τ : Probabilité qu'un saumon pêché soit enregistré comme prise certaine,

δ : Probabilité qu'un saumon pêché et enregistré soit déclaré et le marquage vérifié par les techniciens,

π : Probabilité qu'un castillon soit recapturé après la période de reproduction.

2.3.2. L'expertise a priori est utile pour préciser certains paramètres inconnus

Les paramètres inconnus sont souvent moins inconnus qu'ils ne semblent l'être à première vue : même sans donnée mesurée, tous les sous-ensembles de valeurs possibles sont loin d'être affectés d'un même degré de crédibilité. D'un point de vue technique, l'introduction explicite d'un «prior» constitue la principale différence opérationnelle entre les statistiques conventionnelles et l'approche Bayésienne développée ci-après. Un prior est une fonction de distribution de probabilité qui représente la connaissance des paramètres disponibles avant le recueil des données. Il peut s'interpréter comme un pari sur l'importance relative accordée aux valeurs plausibles des paramètres. Il n'est, en général, pas justifié d'avancer une équidistribution sur les valeurs du stock de saumons car les praticiens ont toujours une certaine connaissance des caractéristiques de la population étudiée (voir Punt et Hilborn, 1997, qui donnent des arguments pour l'emploi de priors informatifs au lieu de priors vagues dans le cas d'évaluation des stocks de pêche). Pour le cas du Scorff, la connaissance *a priori* (noté $\mathbf{X}_0$ dans le raisonnement conditionnel) peut-être résumée comme suit :

Étant donné la taille de la rivière les données antérieures sur la production juvénile (Baglinière and Champigneulle, 1986) dans la rivière et le nombre de survivants après le séjour en mer (Potter and Crozier, 2000), les experts sont prêts à parier à 9 contre 1 que le nombre de saumons rentrant dans le Scorff κ se situe dans l'intervalle $[100, 3000]$ avec une valeur hautement probable autour de 700 individus.

On ne connait guère la probabilité de capture θ au piège près de l'embouchure du Scorff : on pourrait imaginer une répartition symétrique avec 0.5 comme moyenne et seulement $10\ \%$ de chances d'être inférieure à 0.1 ou supérieure à 0.9.

La première estimation du taux de survie des saumons α dans la rivière est supérieure à 0.9. Les experts sont pratiquement sûrs (avec une probabilité *a priori* de 0.9) que α est supérieur à 0.75.

Le taux d'exploitation de la pêche à la ligne β est sans doute situé autour de $0.1 - 0.3$. Il semble peu crédible (moins de $10\ \%$ de chance) que β dépasse 0.7.

La probabilité τ, qu'un saumon attrapé soit reconnu par les contrôles locaux comme prise certaine est supérieure à 0.9 et il semble hautement improbable ($5\ \%$) qu'elle soit inférieure à 0.5.

On sait peu de choses sur la probabilité δ qu'un saumon reconnu soit présenté au contrôle du marquage. Une répartition symétrique avec 0.5 comme moyenne et

seulement 10 % de chances d'être inférieure à 0.1 ou supérieure à 0.9 traduirait cette faiblesse du prior.

En considérant le nombre de sites étudiés et les efforts de survie durant la recapture, la probabilité de recapture π est très vraisemblablement inférieure à 0.25, peu probablement comprise entre 0.25 et 0.5 et il est presque impossible qu'elle soit supérieure à 0.5. Dans ce qui précède «très vraisemblablement» signifie qu'il y a 9 chances contre 1, «presque impossible» représente moins de 1 % de chance et la probabilité restante (environ 9 %) est représentée par les valeurs «improbables».

2.3.3. Elicitation a priori

La figure 2 illustre une loi de probabilité discrète de forme acceptable pour représenter le degré de crédibilité portant sur κ donné par l'expertise $\mathbf{X_0}$. Cette courbe a été obtenue par une «discrètisation» de la fonction gamma avec coefficients 2.4 et 500, tronquée à l'intervalle $[0, 4\,000]$, en raison des ressources limitées de la rivière. Tronquer au delà de 4 000 permet aussi un calcul d'intégration plus commode, mais une analyse de sensibilité montre que c'est largement justifié. Cette courbe présente une valeur maximum aux environs de 700 et met 90 % de chances dans l'intervalle $[100, 3\,000]$.

$$p(\kappa|\mathbf{X}_0) = \frac{\kappa^{2.4} \exp\left(-\frac{\kappa}{500}\right)}{\sum_{z=0}^{z=4000} z^{2.4} \exp\left(-\frac{z}{500}\right)} \tag{1}$$

Les six autres paramètres $\theta, \alpha, \beta, \tau, \delta, \pi$ sont des probabilités, appartenant par conséquent à l'intervalle $[0, 1]$. La forme paramétrique de la loi bêta (2) avec deux coefficients adéquats $a_{\mathbf{X}_0}$ et $b_{\mathbf{X}_0}$ peut représenter une large variété de comportements pour une quantité aléatoire variant entre 0 et 1.

$$p(z|\mathbf{X}_0) = \Gamma(a_{\mathbf{X}_0} + b_{\mathbf{X}_0}) \frac{z^{a_{\mathbf{X}_0}-1}(1-z)^{b_{\mathbf{X}_0}-1}}{\Gamma(a_{\mathbf{X}_0})\Gamma(b_{\mathbf{X}_0})} \tag{2}$$

Supposons $a_{\mathbf{X}_0} > 1$ et $b_{\mathbf{X}_0} > 1$ de telle sorte que la distribution (2) soit unimodale. La figure 3 et la table 2 montrent les résultats de l'élicitation de la loi de probabilité bêta($a_{\mathbf{X}_0}$, $b_{\mathbf{X}_0}$) pour traduire l'expertise à propos des différents paramètres techniques. Le mode de cette fonction $\frac{a_{\mathbf{X}_0} - 1}{a_{\mathbf{X}_0} + b_{\mathbf{X}_0} - 2}$, donne une première relation linéaire entre $a_{\mathbf{X}_0}$ et $b_{\mathbf{X}_0}$ quand l'expert indique la valeur la plus probable du paramètre. Si un intervalle de crédibilité est donné pour la loi *a priori*, les paramètres sont alors ajustés afin de correspondre aux propriétés de la probabilité par une méthode (unidimensionelle) d'approximations successives.

Comme la connaissance *a priori* de chaque paramètre est établie indépendamment, le prior conjoint est le produit de tous les priors.

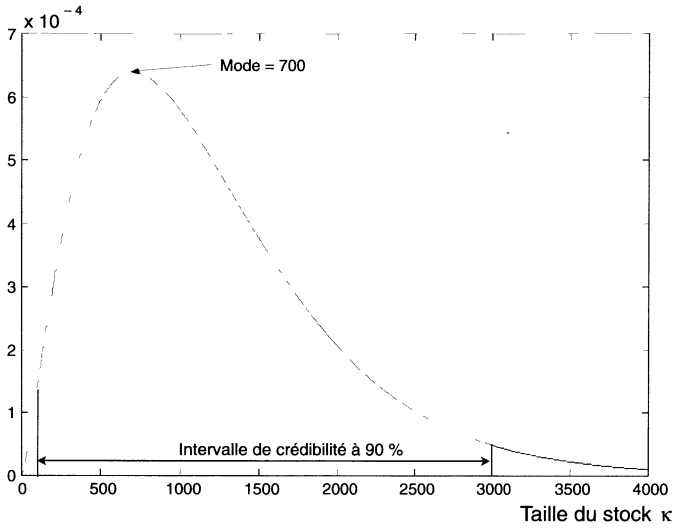

FIGURE 2
Loi a priori pour la taille du stock , paramètre κ

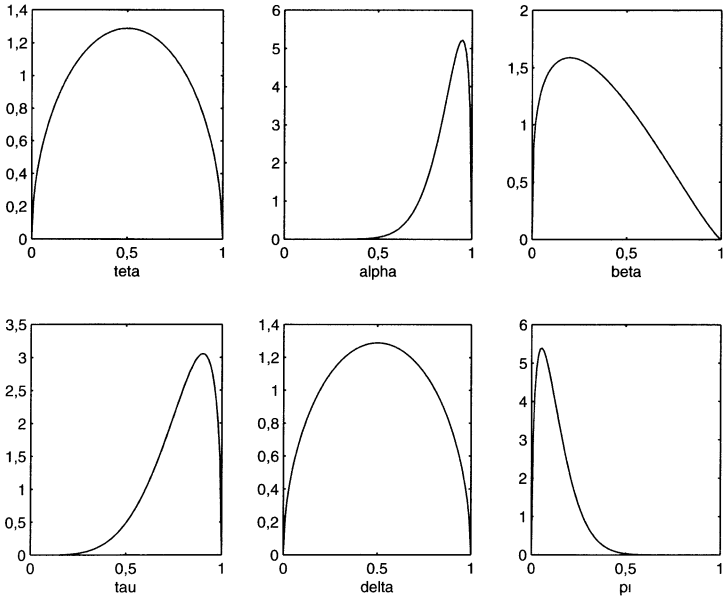

FIGURE 3
Loi a priori pour les paramètres descriptifs de comportement

TABLE 2
L'expertise a priori X_0 est encodée grâce à des distributions de probabilité de type bêta

Paramètre	X_0=$\{$*expertise préalable sur les parmètres*$\}$	a_{X_0}	b_{X_0}	Interprétation
θ	$p(\theta\|X_0)$ sym.; $p(\theta \in [0.1, 0.9]\|X_0) = 0.9$	1.53	1.53	efficacité du piège
α	$Mode(\alpha\|X_0) \approx 0.9$; $p(\alpha > 0.75\|X_0) = 0.9$	10	1.5	taux de survie
β	$Mode(\beta\|X_0) \approx 0.2$; $p(\beta > 0.7\|X_0) \leqslant 0.1$	1.3	2.2	taux de capture
τ	$Mode(\tau\|X_0) \approx 0.9$; $p(\tau < 0.5\|X_0) = 0.05$	5.5	1.5	eff. suivie su site
δ	$p(\delta\|X_0)$ sym.; $p(\delta \in [0.1, 0.9]\|X_0) = 0.9$;	1.53	1.53	prob (déclaration)
π	$Mode(\pi\|X_0) \approx 0.2$; $p(\pi \leqslant 0.25\|X_0) = 0.9$ $p(0.25 \leqslant \pi \leqslant 0.5\|X_0) = 0.09$; $p(\pi > 0.5\|X_0) = 0.01$	1.6	11	taux de recapture

2.3.4. Variables latentes

Les paramètres inconnus et les variables observées ne sont pas suffisants pour décrire les pérégrinations d'un saumon. Des variables latentes, *i.e.* des quantités intermédiaires reliées à des variables non-observées ayant une signification physique, sont alors introduites. Elles sont utiles pour aider à comprendre les étapes intermédiaires de la modélisation conditionnelle. Évidemment le modèle doit être complètement défini : les distributions conditionnelles des variables latentes sachant les paramètres et les variables observables doivent être précisées. Les variables latentes suivantes présentent un intérêt particulier pour la modélisation :

Y_{uu} : saumons non capturés, par conséquent non marqués (indice$_{uu}$, pour *unmarked, uncaptured)*

Y_{mc} : individus marqués pêchés à la ligne,

Y_{uc} : individus non marqués pêchés à la ligne, *unmarked, captured)*

Y_{mf} : individus marqués restés libres pendant la période de pêche, *(marked, free)*

Y_{uf} : individus non marqués restés libres pendant la période de pêche,

Y_{mr} : individus marqués enregistrés comme réellement attrapés, *(marked, registered)*

Y_{ur} : individus non marqués enregistrés comme réellement attrapés,

Y_{ms} : castillons marqués survivants jusqu'au frai,

Y_{us} : castillons non marqués survivants jusqu'au frai.

Certaines combinaisons de variables latentes sont importantes pour établir les compte-rendus des scientifiques. À titre d'exemple, scientifiques et responsables de la pêche aimeraient connaître le champ des valeurs crédibles pour $Y_{mc} + Y_{uc}$,

nombre total de saumons attrapés par les pêcheurs à la ligne. D'autre part $Y_{ms} + Y_{us}$, qui représente «l'échappement», apparaît comme une valeur clé pour connaître la pérennité de l'état du stock.

2.3.5. Le modèle statistique sous la forme d'un graphe acyclique orienté

Les variables latentes et observées ainsi que les paramètres se combinent dans un modèle statistique. Par exemple, nous exprimerons que la variable aléatoire X_1 suit une distribution binomiale de coefficient κ (nombre d'essais) et θ (probabilité de réussite) par la notation : $X_1 \sim B(\kappa, \theta)$. Les équations du modèle comprennent des équations déterministes de bilan et des équations stochastiques de comportement binomial. Elles s'écrivent ainsi :

$$\left.\begin{aligned} X_1 &\sim B(\kappa, \theta) \\ Y_{uu} &= \kappa - X_1 \\ Y_{mc} &\sim B(X_1, \beta),\ Y_{uc} \sim B(Y_{uu}, \beta) \\ Y_{mf} &= X_1 - Y_{mc},\ Y_{uf} = Y_{uu} - Y_{uc} \end{aligned}\right\} \tag{3}$$

$$\left.\begin{aligned} Y_{mr} &\sim B(Y_{mc}, \tau) \\ X_4 &\sim B(Y_{uc} + Y_{mc}, \tau) \\ Y_{ur} &= X_4 - Y_{mr} \\ X_2 &\sim B(Y_{mr}, \delta),\ X_3 \sim B(Y_{ur}, \delta) \\ Y_{ms} &\sim B(Y_{mf}, \alpha),\ Y_{us} \sim B(Y_{uf}, \alpha) \\ X_5 &\sim B(Y_{ms}, \pi),\ X_6 \sim B(Y_{us}, \pi) \end{aligned}\right\} \tag{4}$$

Les équations (3) et (4) permettent la représentation du modèle sur un graphe orienté. Un état de l'art sur ces nouveaux outils Bayésiens se trouve dans Gilks *et al.* (1994) et dans Spiegelhalter *et al.* (1996a). Ces auteurs ont développé le logiciel BUGS (Spiegelhalter *et al*, 1996b) basé sur la représentation graphique. Ce logiciel, gratuit, peut être consulté sur le site web : http ://www.mrc-bsu.cam.ac.uk/bugs. L'intérêt scientifique de BUGS pour la gestion des stocks a été précédemment mohtré par Meyer et Millar (1999, 2000) par l'étude d'un cas basé sur les modèles dynamiques d'état.

La figure 4 représente toutes les quantités par des noeuds (soit stochastiques, soit déterministes) sur un graphe orienté d'influence, où les flèches pénètrent dans un noeud depuis les variables qui exercent une influence directe sur celui-ci (les noeuds représentant l'ascendance directe par une liaison stochastique sont appelés «parents» dans ce qui suit). Les descendants d'un noeud sont soit des variables latentes soit des variables observées. Des graphes acycliques orientés permettent une interprétation sans ambiguité des termes «parents», «descendants», «ancêtres», «fils», etc... Les associations entre grandeurs reliées par des équations déterministes de bilan ne

participent pas à ces définitions de parenté : le noeud déterministe correspondant est situé au même niveau que la variable stochastique à laquelle il est rattaché. La figure 5 donne le graphe acyclique orienté qui correspond au graphe d'influence de la figure 4 en effectuant l'élimination des noeuds déterministes : seules sont conservées les quantités aléatoires sur lesquelles portera l'inférence Bayésienne. Pour la commodité du dessin, on a associé les variables (Y_{mc}, Y_{uc}) en un même noeud.

Le graphe de la figure 5 représente le raisonnement conditionnel sur lequel le modèle est basé : les flèches du raisonnement conditionel descendent des paramètres conceptuels jusqu'aux quantités observées.

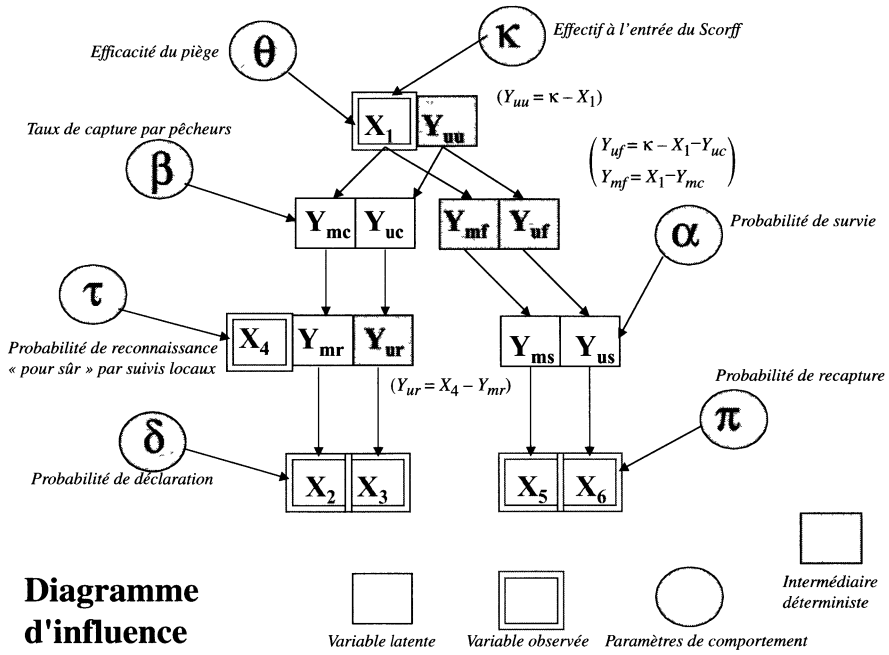

FIGURE 4
La vie d'un saumon après sa remontée dans le Scorff sous la forme d'un graphe d'influence

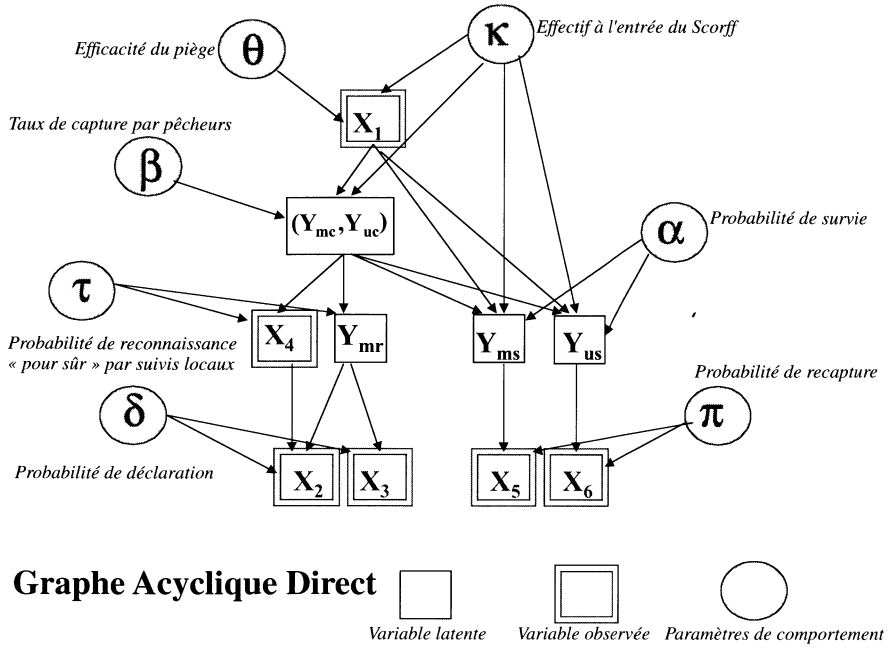

FIGURE 5
La vie d'un saumon après sa remontée dans le Scorff sous la forme d'un graphe acyclique direct

Les (hyper)paramètres des lois *a priori* complètent le graphe comme noeuds parents pour les paramètres du modèle. L'inférence Bayésienne consiste à exprimer formellement que l'information apportée par les données sert à mettre à jour la connaissance *a priori* des paramètres. Le savoir *a posteriori* est traduit par une loi de probabilité, fonction des quantités observées. L'obtention de cette loi *a posteriori* est facilitée grâce à la structure spéciale du graphe. L'indépendance conditionnelle représentée Figure 5 s'exprime par la propriété markovienne : la probabilité conditionnelle d'un noeud sachant tous ses ascendants ne dépend que de ses noeuds parents. Ainsi par exemple :

$$p(X_5\,|Y_{ms}\,, Y_{mf}, X_1, \alpha, \pi, \kappa, \theta) = p(X_5\,|Y_{ms}\,, \pi)$$

En conséquence, la distribution conjointe de toutes les quantités q du problème, $q\epsilon\ Q, =$ {variables observées, variables latentes et paramètres } est une simple factorisation :

$$p(Q) = \prod_{q \in Q} p(q|parents[q]) \tag{5}$$

Notons que les paramètres sont des ancêtres pour tous les noeuds (Ils se situent à l'extérieur du graphe de la figure 5). La formule (5) n'est rien d'autre

que la factorisation de la vraisemblance étendue aux variables latentes par les lois de probabilité *a priori*. En intégrant sur les variables latentes dans les formules (3) et (4), on obtient la loi jointe des observables et des paramètres. La loi *a posteriori* des paramètres étant donné les observations pourrait être formellement dérivée en appliquant ensuite le théorème de Bayes (ce qui demande cette fois, d'intégrer sur les paramètres). La section suivante montre que travailler avec les modèles graphiques et l'algorithme de Gibbs court-circuite toutes ces étapes de la modélisation Bayésienne et évite aussi l'évaluation numérique de lourdes intégrales multidimensionnelles.

2.3.6. Le modèle interannuel est un «empilement» des modèles annuels

Il n'y a pas de difficulté conceptuelle supplémentaire pour établir un modèle interannuel. Il faut souligner la très forte hypothèse de stabilité des paramètres, permettant une cohérence interannuelle (donc un transfert d'information d'année en année) en partageant les valeurs communes de $\theta, \alpha, \beta, \tau, \delta$ et π. Cette hypothèse est particulièrement discutable sur l'efficacité de la capture θ et la probabilité de recapture π qui peuvent certainement varier d'une année sur l'autre en fonction du débit de la rivière et des conditions hydrométéorologiques.

3. Chaîne de Monte Carlo Markov - Inférence Bayésienne

Selon l'approche Bayésienne, l'inférence consiste à mettre à jour, la loi *a priori* du paramètre $p(\kappa, \theta, \alpha, \beta, \tau, \delta, \pi|\mathbf{X}_0)$, (où $\mathbf{X}_0$ rappelle que l'on conditionne sur un savoir initial et des hypothèses de construction) en une loi *a posteriori* $p(\kappa, \theta, \alpha, \beta, \tau, \delta, \pi|\mathbf{X}_0, X_1, X_2, \ldots, X_6)$ en tenant compte des quantités observées $(X_1, X_2, \ldots, X_6)$. En tant que densité de probabilité multivariable, cette expression est bien souvent impossible à utiliser : même si elle peut-être déduite des équations du modèle (3) à (5) et des lois *a priori* $p(\kappa, \theta, \alpha, \beta, \tau, \delta, \pi)$, elle fait appel à des intégrales multiples qui proviennent de l'intégration des variables latentes ou du dénominateur de la formule de Bayes. Pendant longtemps, l'évaluation de telles intégrales a limité l'inférence Bayésienne à l'usage d'exemples académiques, insuffisamment réalistes pour traiter des problèmes opérationnels. De fait, l'obtention analytique de la loi *a posteriori* $p(\kappa, \theta, \alpha, \beta, \tau, \delta, \pi|X_1, X_2, \ldots, X_6)$ n'apporte qu'une satisfaction mathématique. En pratique, il est utile de dériver les moyennes, d'évaluer les intervalles de crédibilité des paramètres, etc... Ces opérations peuvent se faire de façon empirique, sans recours à un calcul d'intégrale, si un échantillon des valeurs du paramètre tiré dans la loi *a posteriori* est disponible. Par conséquent, un regain d'intérêt s'est manifesté pour de vieux algorithmes (Metropolis *et al.*, 1953; Hastings, 1970), fondé sur la simulation plutôt que l'approximation numérique, amplifié par le developpement d'ordinateurs toujours plus puissants. Les difficultés pratiques du calcul Bayésien sont aujourd'hui maîtrisées grâce à l'utilisation d'algorithmes Monte Carlo à base de chaînes de Markov. Ces algorithmes permettent de générer un «pseudo-échantillon» qui a toutes les propriétés statistiques d'un échantillon de la loi *a posteriori* (Kass *et al.*, 1996; Brooks, 1998).

Dans les ouvrages de référence récents, sur les méthodes MCMC (Gelman *et al.*, 1995; Tanner, 1996; Robert et Casella, 1999), deux types de méthodes sont

mis en vedette : l'algorithme de Metropolis-Hastings et l'échantillonnage de Gibbs. L'algorithme de Metropolis-Hastings est général. Il peut simuler n'importe quelle distribution de la loi *a posteriori* conjointe des paramètres du modèle, quand la distribution *a priori* et la vraisemblance sont connues. L'algorithme explore le domaine de définition du paramètre, en utilisant une stratégie spécifique itérative aléatoire. Au sens mathématique, cet algorithme est une chaine de Markov homogène et positive qui converge sous des conditions très générales vers la limite désirée de répartition aléatoire des valeurs, à savoir la distribution *a posteriori* du paramètre. D'un autre coté, l'échantillonnage de Gibbs (Geman and Geman, 1984; voir aussi Casella et George, 1992, pour une revue de littérature sur le sujet) fonctionne seulement quand toutes les distributions conditionnelles d'un paramètre, étant donné les autres paramètres et les données (conditionnelles complètes), sont disponibles pour la simulation stochastique. Ceci est le cas lorsque le modèle peut-être représenté par un arbre orienté comme dans la figure 5. Le problème est ici d'autant plus simplifié que les conditionnelles complètes n'impliquent seulement que des distributions bêta et binomiales.

3.1. L'échantillonnage de Gibbs divise un problème complexe en plusieurs sous-problèmes simples

Supposons que l'on veuille générer un échantillon de triplets $\{(a^1, b^1, c^1), (a^2, b^2, c^2), ...\}$ pour la loi conjointe $p(A, B, C)$ de trois variables aléatoires A, B, C. Si les conditionnelles complètes, *i.e.* des fonctions de densité de probabilité $p(A\,|B, C)$, $p(B\,|A, C)$, $p(C\,|A, B)$ sont de forme facilement simulable, l'algorithme de Gibbs, à partir d'un triplet initial (a^0, b^0, c^0) effectue des itérations comme suit :

À l'étape i, on a déjà obtenu $(a^{i-1}, b^{i-1}, c^{i-1})$, à l'étape précédente, on génère alors :

$$a^i \sim p(A\,|B = b^{i-1}, C = c^{i-1}),$$

$$b^i \sim p(B\,|A = a^i, C = c^{i-1}),$$

$$c^i \sim p(C\,|A = a^i, B = b^i).$$

L'itération de ces trois phases de l'algorithme stochastique garantit que la chaîne d'échantillonnage de Gibbs converge ergodiquement vers la distribution $p(A, B, C)$. Les démonstrations des propriétés ergodiques des Chaines de Monte Carlo Markov et les applications pratiques de ces techniques sont désormais largement publiées et il n'est pas besoin de les rappeler ici. On peut en trouver les détails dans Robert et Casella (1999) ou dans Tanner (1996).

Dans la pratique, cela signifie qu'après avoir écarté les triplets dans une période initiale de «chauffe», les triplets générés ensuite se comporteront comme un échantillon aléatoire de (A, B, C) pour tout calcul statistique caractéristique, comme l'évaluation de la moyenne d'une fonction mesurable de (A, B, C). Naturellement, la propriété peut se généraliser pour plus de trois variables. Pour en revenir au problème de déduction de la taille de la population de saumons,

ce résultat s'applique comme suit : en générant les paramètres tour à tour, un à un selon les sept conditionnelles complètes $p(\kappa \,|\theta, \alpha, \beta, \tau, \delta, \pi, X_1, X_2, \ldots, X_6)$, $p(\theta \,|\kappa, \alpha, \beta, \tau, \delta, \pi, X_1, X_2, \ldots, X_6)$, ... $p(\pi \,|\kappa, \theta, \alpha, \beta, \tau, \delta, X_1, X_2, \ldots, X_6)$, on peut obtenir un échantillon de 7-uplets provenant de $p(\kappa, \theta, \alpha, \beta, \tau, \delta, \pi \,|X_1, X_2, \ldots, X_6)$, la loi *a posteriori* conjointe des paramètres.

3.2. Dans un modèle graphique orienté, les conditionnelles complètes impliquent seulement les noeuds parent et fils

Considérons comme dans la figure 6, une branche dans un modèle graphique orienté avec la variable aléatoire éventuellement multidimensionnelle A, regroupant l'ensemble des noeuds parents de B, qui à son tour est un noeud parent pour la quantité aléatoire C (éventuellement multidimensionnelle). En travaillant avec toutes les autres variables fixées, notées ci-après $(ABC)^-$, la structure orientée du modèle traduit la représentation graphique de la propriété d'indépendance conditionnelle pour la conditionnelle complète de C : $p(C \,|A, B, (ABC)^-) = p(C \,|B, (ABC)^-)$

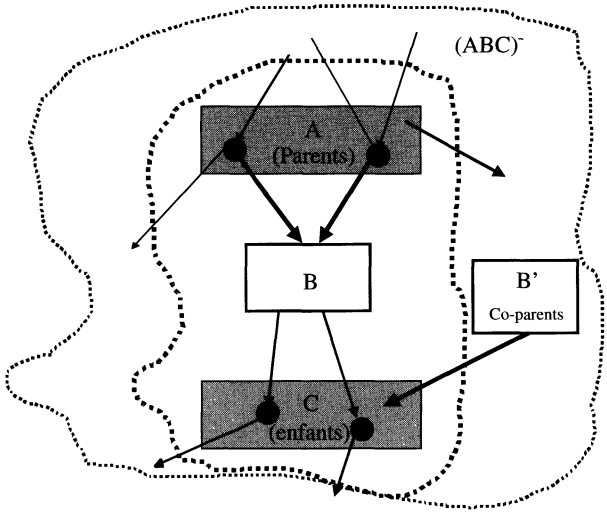

FIGURE 6
La mise à jour Bayésienne tire parti de la structure conditionnelle

Comme le noeud B est situé entre les ensembles de noeuds A et C, la loi de Bayes implique que :

$$p(B \,|A, C, (ABC)^-) = \frac{p(B \,|A, (ABC)^-)p(C \,|A, B, (ABC)^-)}{p(C \,|A, (ABC)^-)} \propto p(B \,|A, (ABC)^-)p(C \,|B, (ABC)^-) \tag{6}$$

$$p(B \,|A, C, (ABC)^-) \propto p(B \,|A)p(C \,|B, (ABC)^-) \tag{7}$$

Quand on considère cette expression en tant que fonction de B, le dénominateur de Bayes ne dépend pas de B : il s'agit d'une constante de normalisation qui garantit que l'intégrale sur tout le domaine de B vaudra 1. Le numérateur est exprimé par deux probabilités conditionnelles qui suivent la direction des flèches : puisque c'est cette démarche même qui a créé le modèle, ces deux expressions sont donc explicitement connues! Par conséquent, sur un modèle graphique orienté, les conditionnelles complètes sont connues à une constante normalisante près. L'échantillonnage issu de chacune de ces distributions conditionnelles complètes *a posteriori* n'est pas difficile à obtenir puisque chacune d'elles est une loi de probabilité et la constante normalisante est au pire une intégrale monodimensionnelle (que l'on peut brutalement calculer point par point ou atteindre par des algorithmes de simulation rapides). On peut aller plus loin en appelant B' les parents autres que B du noeud C. Quand dans l'expression (6), on va chercher à calculer $p(C\,|B,(ABC)^-)$, il suffit de réitérer le raisonnement précédent avec C jouant le rôle de B et $(B, B\prime)$ jouant le rôle de A. Il vient :

$$p(C\,|B,(ABC)^-) = Const \times p(C\,|B,B') \tag{8}$$

Dans l'équation (8), le terme *Const* n'est pas une fonction de B

Finalement, en combinant (6) et (8) et en ne gardant que les termes dépendants de B, la structure orientée du graphe permet une écriture «locale» de la règle de Bayes :

$$p(B\,|A,C,(ABC)^-) = \frac{p(B\,|A)p(C\,|B,B')}{\int\limits_z p(z\,|A)p(C\,|z,B')dz} \tag{9}$$

Dans cette écriture locale de la règle de Bayes, la relation liant un noeud B à ses parents A joue le rôle du prior et la relation liant ses enfants C aux parents (B, B') joue le rôle de la vraisemblance.

Selon l'équation (9), la conditionnelle complète de chaque quantité aléatoire formant un noeud du graphe (observables, variable latente ou paramètre) du graphe est uniquement fonction des parents de ce noeud, de ses enfants et des co-parents de ses enfants. Cet ensemble conditionnant forme la couverture markovienne du noeud. À partir du graphe acyclique orienté direct, si on rajoute des liens fictifs entre les variables qui ont les mêmes descendants directs et si on transforme tous les liens en liens non orientés, on obtient un graphe qui exprime la structure de voisinage de chaque noeud associée au conditionnement. La figure 7 réalise cette transformation à partir du graphe acyclique orienté direct de la figure 5. Spiegelhalter *et al.* (1996a) appellent avec humour cette opération «moralisation» car elle consiste à regrouper les familles en «mariant» les parents.

D'un point de vue théorique, cette opération s'apparente à la mise en relation bijective entre un champ de Gibbs et un champ de Markov latticiels : (théorème de Hammersley-Clifford dans Guyon, 1995). Le problème spécifique a d'abord été présenté sous la forme de graphe orienté acyclique, structure proche de celle d'un champ de Gibbs latticiel (figure 5); on passe maintenant pour la suite de l'article à sa représentation sous la forme d'un champ de Markov (figure 7).

D'un point de vue pratique, malgré son aspect rébarbatif, la figure 7 est des plus utiles : quand on arrive à isoler une quantité des autres par une frontière conditionnante

(la couverture markovienne), on en déduit l'indépendance (conditionnelle à la frontière) des deux parties ainsi délimitées. Par exemple les noeuds κ et X_1 séparent le noeud θ des autres quantités : on en déduit que la conditionnelle complète de θ ne dépendra que de κ et X_1. En s'appuyant sur la figure 7, la section suivante recherche, pour chaque variable, ses voisines, de façon à l'isoler et exprimer ou reconnaître la conditionnelle complète. La table 3 donne pour chaque variable d'intérêt de l'inférence bayésienne (c-a-d paramètre ou variable latente stochastique), la couverture markovienne associée. Cette table se construit par lecture de voisinages à partir de la figure 7, sauf les deux dernières lignes précisant la couverture de chacune des composantes du noeud (Y_{mc}, Y_{uc}) pour lesquelles il faut construire un graphe un peu plus détaillé que celui donné par la figure 5.

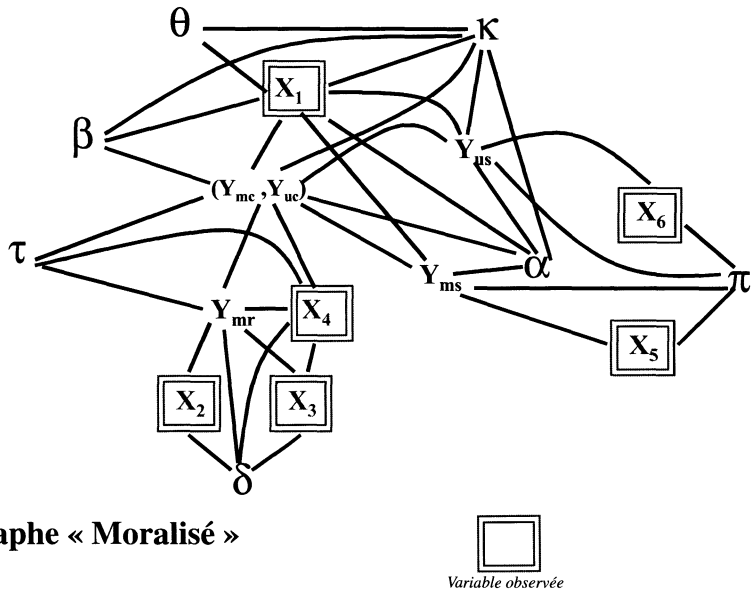

FIGURE 7
La vie d'un saumon après sa remontée dans le Scorff sous la forme d'un graphe «moralisé»

TABLE 3
Noeuds figurant dans les conditionnelles complètes

Nœud à mettre à jour	Liste des variables de sa clique intervenant dans la conditionnelle complète
θ	κ, X_1
β	κ, X_1, Y_{mc}, Y_{uc}
τ	Y_{mc}, Y_{uc}, Y_{mr}, X_4
δ	Y_{mr}, X_4, X_2, X_3
π	Y_{ms}, Y_{us}, X_5, X_6
α	X_1, Y_{mc}, Y_{uc}, κ, Y_{ms}, Y_{us}
Y_{ms}	X_1, Y_{mc}, Y_{uc}, α, π, X_5
Y_{us}	X_1, κ, Y_{mc}, Y_{uc}, α, π, X_6
κ	θ, β, X_1, Y_{us}, α, Y_{mc}, Y_{uc}
Y_{mr}	X_4, X_2, X_3, τ, δ, Y_{mc}, Y_{uc}
(Y_{mc}, Y_{uc})	β, X_1, κ, Y_{us}, α, Y_{ms}, X_4, Y_{mr}, τ
Y_{mc}	β, X_1, Y_{us}, α, Y_{ms}, Y_{mr}, τ
Y_{uc}	β, X_1, κ, Y_{us}, α, X_4, Y_{mr}, τ

Dans la section suivante, pour certains noeuds, on reconnaitra dans l'équation (9) une structure connue (par conjugaison); par contre, pour d'autres noeuds, la forme de leur conditionnelle complète ne sera pas dans la bibliothèque des lois de probabilité standard et il faudra l'expliciter.

3.3. Actualisation Bayésienne des éléments d'un modèle graphique orienté par l'échantillonnage de Gibbs

On remarquera que seuls les noeuds stochastiques (à l'exception des observables qui sont des noeuds terminaux) peuvent être mis à jour par le théorème de Bayes (9). Les noeuds déterministes ne sont que des quantités intermédiaires.

La marginalisation permet de ne pas tenir compte des variables latentes. L'approche Bayésienne traite les variables latentes comme les autres paramètres. Leurs distributions conditionnelles complètes sont évaluées. Par conséquent l'échantillonnage de Gibbs générera un «pseudo-échantillon» de :

$$p(\kappa,\theta,\alpha,\beta,\tau,\delta,\pi,Y_{uu},Y_{mc},Y_{uc},Y_{mf},Y_{uf},Y_{mr},Y_{ur},Y_{ms},Y_{us}\,|\mathbf{X_0},X_1,X_2,\ldots,X_6) \tag{10}$$

De cet échantillon, on extraira simplement les valeurs des paramètres intéressants (et on oubliera celles des variables latentes) afin d'obtenir un échantillon issu de :

$$p(\kappa, \theta, \alpha, \beta, \tau, \delta, \pi \,|\mathbf{X}_0, X_1, X_2, \ldots, X_6) \tag{11}$$

Les propriétés conjuguées des lois binomiales bêta rendent les mises à jour Bayésiennes plus faciles. Par exemple, la figure 7 montre que la distribution conditionnelle complète du paramètre π quantifiant la probabilité de recapture dépend seulement du prior et de la loi de probabilité qui relie le noeud π aux quantités observées X_5, X_6 : en effet, les noeuds X_5, X_6 isolent π du reste du monde. Les lois *a priori* bêta ont été calées pour tous les paramètres compris entre 0 et 1 pour représenter les croyances *a priori* sur les valeurs possibles de ces paramètres.

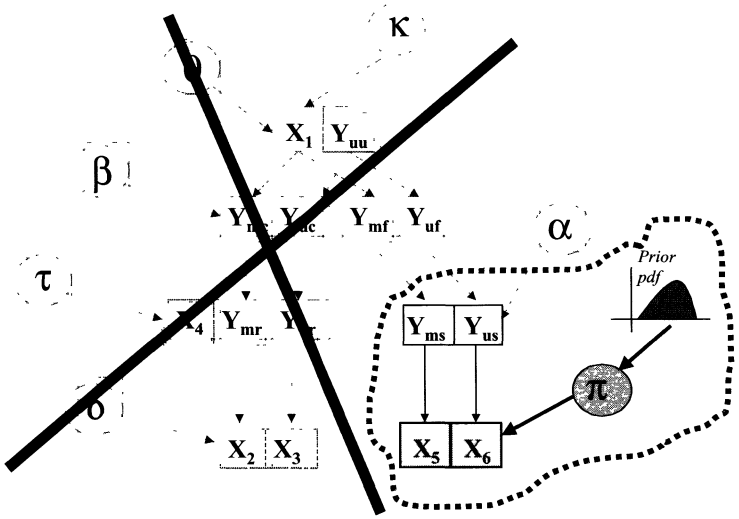

FIGURE 8
Inférence Bayésienne par échantillonneur de Gibbs pour le paramètre π

Le prior de π est une bêta($a_{\mathbf{X}_0}$,$b_{\mathbf{X}_0}$), avec $a_{\mathbf{X}_0} = 1.6$ et $b_{\mathbf{X}_0} = 11$ donné par l'équation (2).

X_5, X_6 sont des variables binomiales (conditionnellement indépendantes sachant π)

$$p(X_5, X_6 \,|\pi, Y_{ms}, Y_{us})$$

$$= \frac{\Gamma(Y_{ms}+1)\Gamma(Y_{us}+1)}{\Gamma(X_5+1)\Gamma(X_6+1)} \frac{\pi^{X_5+X_6}(1-\pi)^{Y_{ms}+Y_{us}-X_5-X_6}}{\Gamma(Y_{ms}-X_5+1)\Gamma(Y_{us}-X_6+1)} \tag{12}$$

D'après le théorème de Bayes (9), la conditionnelle complète *a posteriori* de π peut s'écrire :

$$p(\pi|\mathbf{X}_0, X_5, X_6, Y_{ms}, Y_{us}) \propto p(X_5, X_6\,|\pi, Y_{ms}, Y_{us}\,)p(\pi\,|\mathbf{X}_0) \\ \propto \pi^{X_5+X_6+a_{\mathbf{X}_0}-1}(1-\pi)^{Y_{ms}+Y_{us}-X_5-X_6+b_{\mathbf{X}_0}-1} \tag{13}$$

Considérant cette expression comme une fonction de π, on reconnait une distribution de la même famille que le prior, c-a-d une fonction bêta avec des coefficients mis à jour $X_5 + X_6 + a_{\mathbf{X}_0}$ et $Y_{ms} + Y_{us} - X_5 - X_6 + b_{\mathbf{X}_0}$.

Tous les autres paramètres de probabilité peuvent être un à un facilement isolés d'un grand nombre d'autres noeuds du graphe de la figure 7 en ne conservant que les grandeurs qui interviennent dans le nombre d'essais ou dans le nombre de succès du tirage binomial dont ils sont paramètres. Ils obéissent à un système similaire de mise à jour : un prior bêta donnera un posterior de même type quand on conditionnera sur des résultats d'un tirage binomial. Pour $\theta, \alpha, \beta, \tau, \delta, \pi$, on choisit également des lois bêta. Notons qu'on dispose de générateurs aléatoires performants pour la loi bêta.

Conditionnelles complètes non explicites

Actualisation de la taille du stock. L'évaluation de la conditionnelle complète de κ est un peu plus complexe, car la loi *a priori* n'est pas une fonction de distribution standard. Sur la figure 7, on voit que les voisins de κ sont $(\beta, \theta, X_1, Y_{mc}, Y_{uc}, Y_{us}, \alpha)$. Toutes ces grandeurs conditionnantes sont présentes, car si on retourne à la figure 4, κ est une partie des noeuds de bilan déterministe Y_{uu} et Y_{uf} , de telle sorte que la recherche des noeuds stochastiques descendants ne prend fin qu'avec Y_{uc} et Y_{us}, qui sont partie prenante dans l'expression de la conditionnelle complète κ. La figure 9 montre la partie de l'arbre orienté concernée par cette mise à jour Bayésienne. La formule de la conditionnelle complète pour κ est donnée dans l'appendice A.

Mise à jour des variables latentes.

L'évaluation de la conditionnelle complète des variables latentes s'effectue selon la même démarche. Par exemple, la figure 7 nous dit que «seulement» $(\beta, X_1, \kappa, Y_{us}, \alpha, Y_{ms}, X_4, Y_{mr}, \tau\,)$ vont intervenir dans la conditionnelle complète du couple (Y_{mc}, Y_{uc}). Le détail donné à la figure 10 pour la seule variable Y_{mc} montre que la mise à jour de Y_{mc} implique ses noeuds parents β et X_1 aussi bien que ses noeuds descendants Y_{mr} et Y_{ms} (via le noeud de bilan déterministe $Y_{mf} = X_1 - Y_{mc}$) ainsi que τ qui, avec Y_{mc} , est co-parent de Y_{mr}. Les détails mathématiques sont donnés dans l'appendice B.

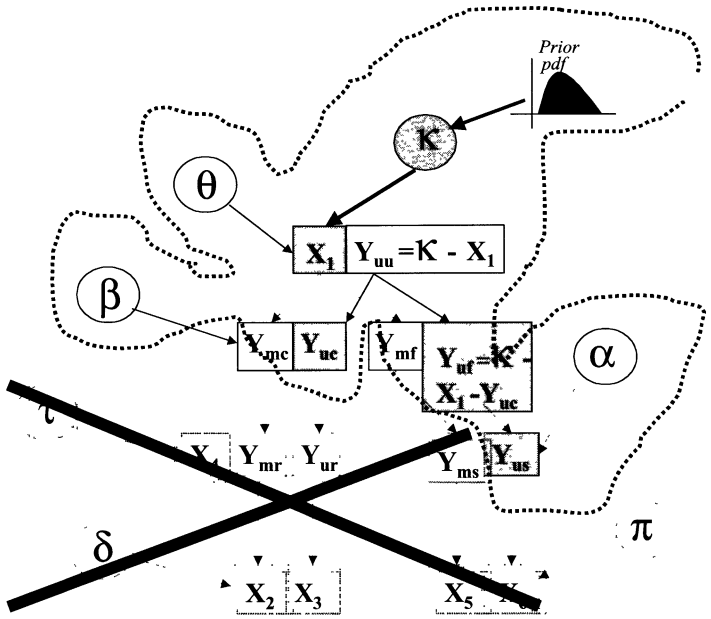

FIGURE 9
Conditionnelle complète pour κ, la taille du stock

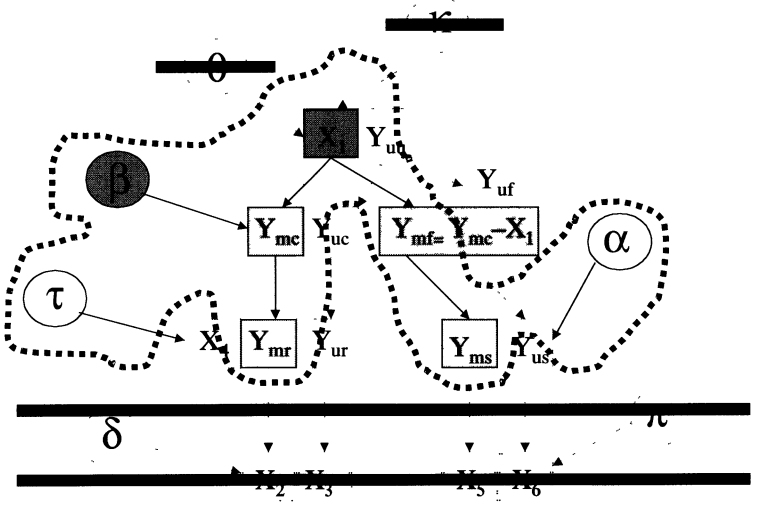

FIGURE 10
Éléments du graphe impliqués dans le calcul de la conditionnelle complète de Y_{mc}

4. Résultats Numériques

4.1. Année 1995

4.1.1. Calcul MCMC

La version actuelle de WINBUGS (édition 1.3, Feb 2000) rencontre des difficultés pour réaliser le tirage binomial de X_2 tout en assurant la positivité de $X_4 - Y_{mr} = Y_{ur}$. Un programme MATLAB, itérant la mise à jour des paramètres et des variables latentes selon des lois bêta et les équations données dans les annexes, a été utilisé pour fournir les résultats ci-après. Trois chaînes de 100 000 échantillons sont générées par l'algorithme de Gibbs mais seules les 5000 dernières valeurs sont conservées. Le diagnostic de Gelman et Rubin (1992) basé sur une analyse classique de variance pour comparer les variances inter et intra-chaînes est satisfait pour tous les paramètres. Cependant, l'autocorrélation reste particulièrement forte parmi les échantillons pour β, π et τ, ce qui indique que l'exploration MCMC de leur domaine est lente mais le mélange correct entre les trois chaînes permet de conclure qu'une exploration adéquate du domaine *a posteriori* a été réalisée d'après ce grand nombre d'itérations. Les estimations empiriques de probabilité données dans les figures 11 et 12 et l'intervalle de crédibilité à 90 % de la Table 4 proviennent directement de cet échantillonnage MCMC.

TABLE 4
Intervalles de crédibilité pour les paramètres avec prise en compte de l'année 1995 seulement.

Paramètre	Moyenne	Écart-type	95 % quantile	5 % quantile
θ	0.67	0.04	0.74	0.61
α	0.81	0.10	0.95	0.62
β	0.11	0.02	0.15	0.09
τ	0.89	0.09	0.99	0.70
δ	0.65	0.06	0.73	0.56
π	0.11	0.02	0.15	0.08
κ	747	41	816	680
$Y_{mc} + Y_{uc}$ (pêche effective)	85	12	109	75
$Y_{ms} + Y_{us}$ (échappement)	534	70	640	410

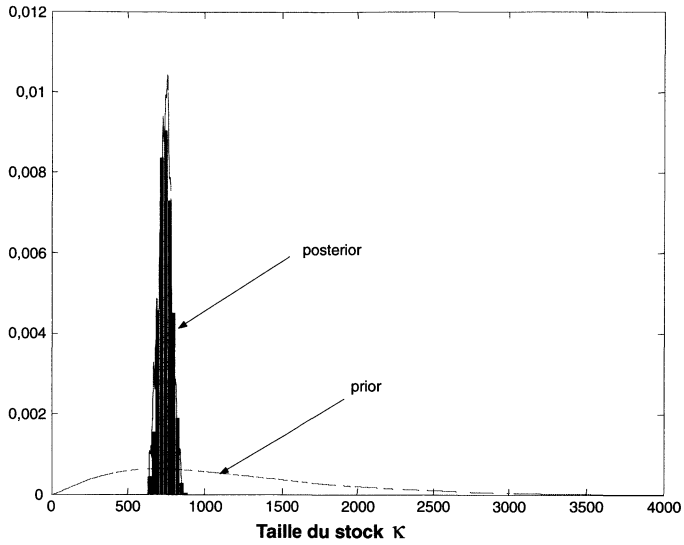

FIGURE 11
Inférence Bayésienne pour la taille du stock, paramètre κ
(prise en compte de l'année 1995 seulement)

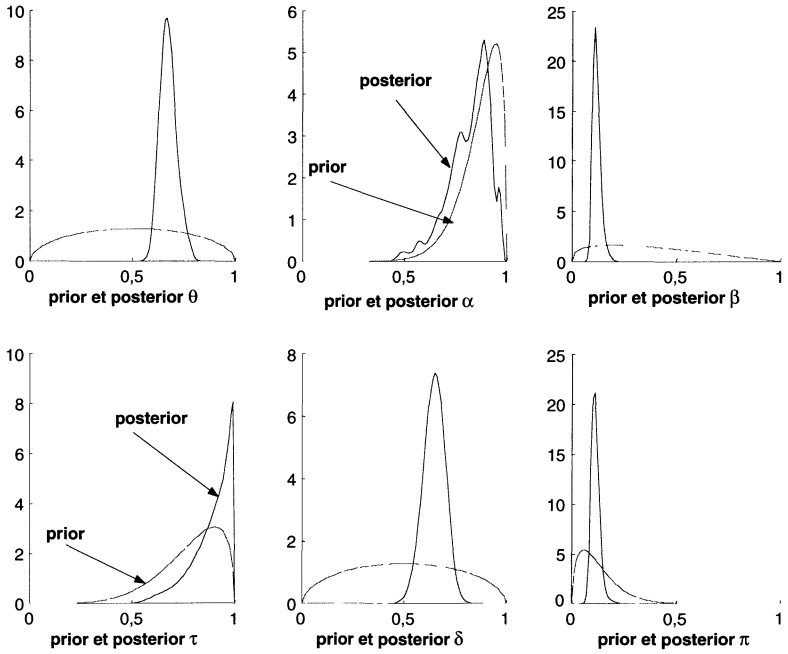

FIGURE 12
Inférence pour les paramètres de comportement
(prise en compte de l'année 1995 seulement)

4.1.2. L'inférence Bayésienne

Un simple coup d'oeil au prior et à la probabilité *a posteriori* pour chacun des paramètres basiques (figures 11 et 12) montre que, pour la majorité d'entre eux, l'incertitude initiale est considérablement reduite. L'efficacité du piège θ est supérieure à 0.5. Le dispositif de capture crée un fort courant qui attire les castillons de retour. La proportion prélevée par les pêcheurs à la ligne est d'environ 10%. Seuls, le taux de survivants α et l'efficacité de l'enregistrement τ restent très imprécis. La probabilité *a posteriori* α est semblable à son prior. Cela s'explique en revenant au diagramme d'influence de la figure 4 : aucune information en provenance de données n'est reliée directement à α . Le mode *a posteriori* de δ est très différent de son emplacement *a priori*. Cette différence révèle un trait spécifique du Scorff qui ne s'explique pas par l'expertise *a priori* ni selon les hypothèses du modèle.

L'emploi de l'échantillonnage de Gibbs peut aussi être utile pour étudier la covariation entre les paramètres. La matrice de corrélation donnée Table 5 montre que l'évaluation *a posteriori* du taux de survivants α ne peut se faire indépendemment de l'information concernant l'efficacité de la recapture π. Comme on peut s'y attendre, l'influence de l'action des pêcheurs β et la probabilité d'enregistrement τ sont partiellement confondues : leur corrélation vaut en moyenne -0.7. Elle est négative car l'essentiel de l'information est apporté par X_1 et X_4 : à X_1 et X_4 connus, X_1 renseigne fortement sur κ et si on fait le pari que β est grand, il faut alors en même temps faire le pari que τ est petit car $E(X_4\,|\kappa,\beta,\tau) = \kappa\beta\tau$. La relation entre θ et κ est issue de l'hypothèse binomiale $E(X_1\,|\theta,\kappa) = \kappa\theta$.

TABLE 5
Matrice de corrélation a posteriori entre les paramètres (prise en compte de l'année 1995 seulement)

Matrice de corrélation	κ	α	β	δ	τ	π	θ
κ	1						
α	- 0.21	1					
β	- 0.28	0.08	1				
δ	0.00	0.01	- 0.01	1			
τ	- 0.06	0.02	**- 0.70**	0.01	1		
π	- 0.15	**- 0.67**	0.11	0.03	- 0.09	1	
θ	**- 0.91**	0.19	0.26	0.00	0.05	0.14	1

4.2. 5 années de données

Les figures 13 et 14 rapportent les résultats des calculs Bayésiens tenant compte des cinq dernières années de données de la table 1 selon le modèle interannuel. En

comparant la table 4 et la table 6 on s'aperçoit que les écarts-types se réduisent quand on intègre plus d'information dans l'analyse. Cela est dû à un effet « boule de neige » : l'information supplémentaire est véhiculée d'une année sur l'autre par l'intermédiaire des paramètres communs $(\pi, \theta, \alpha, \beta, \tau, \delta)$ jusqu'à diminuer le domaine d'incertitude attaché aux valeurs plausibles de la taille de chacun des stocks annuels. Notons que les écarts-type se réduisent tous quand on passe au modèle sur 5 ans sauf la probabilité de recapture π qui, même si elle est en moyenne plus élevée, se retrouve bien plus mal déterminée. L'intervalle de crédibilité est d'une longueur deux fois plus importante et disjoint de celui obtenu en 1995. Ceci est l'indication d'une variabilité interannuelle de la pêche de recapture qu'on retrouve dans les faits : en décembre, durant la période de frai, le comptage des reproducteurs s'effectue dans l'eau glacée, la nuit : la lumière les attire et l'on repère mieux à la lampe torche ces poissons de très grande taille engagés dans leur activité reproductrice avant de mourir. Par conséquent la proportion de capture π est très fortement influencée par la date de la pêche et les conditions hydrométéoclimatiques. Si le courant est fort ou que les techniciens évaluent mal la date de frai, la plupart des reproducteurs sont invisibles. Le modèle interannuel est donc peu réaliste vis-à-vis de la non stationnarité de ce paramètre π .

TABLE 6
Intervalles de crédibilité pour les paramètres, avec prise en compte des cinq années de données

Paramètre	Moyenne	Écart-type	95 % quantile	5 % quantile
θ	0.72	0.02	0.76	0.69
α	0.41	0.11	0.58	0.24
β	0.11	0.01	0.13	0.10
τ	0.99	0.01	1	0.90
δ	0.63	0.03	0.68	0.59
π	0.27	0.08	0.32	0.16
κ_{1995}	700	25	740	660
κ_{1996}	695	26	740	650
κ_{1997}	430	17	460	400
κ_{1998}	590	21	625	560
κ_{1999}	235	11	250	220

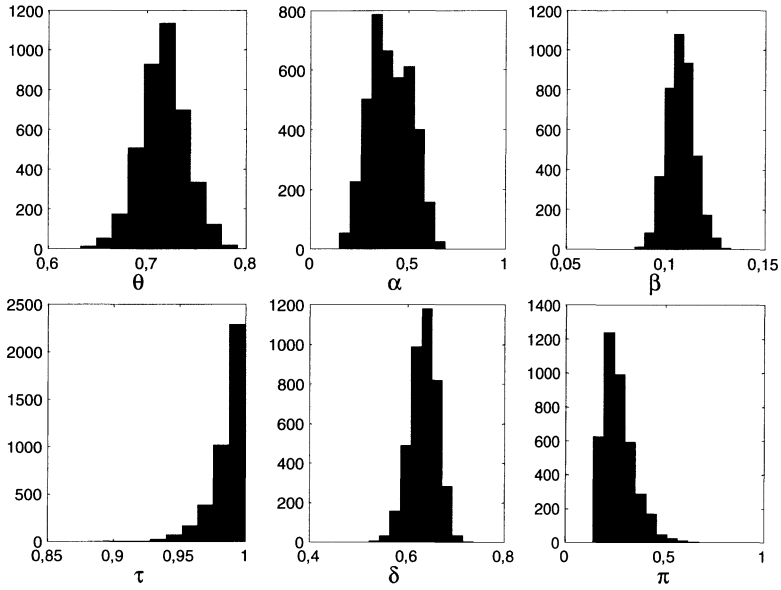

FIGURE 13
Lois marginales a posteriori des paramètre de comportement (prise en compte de la période 1995-1999)

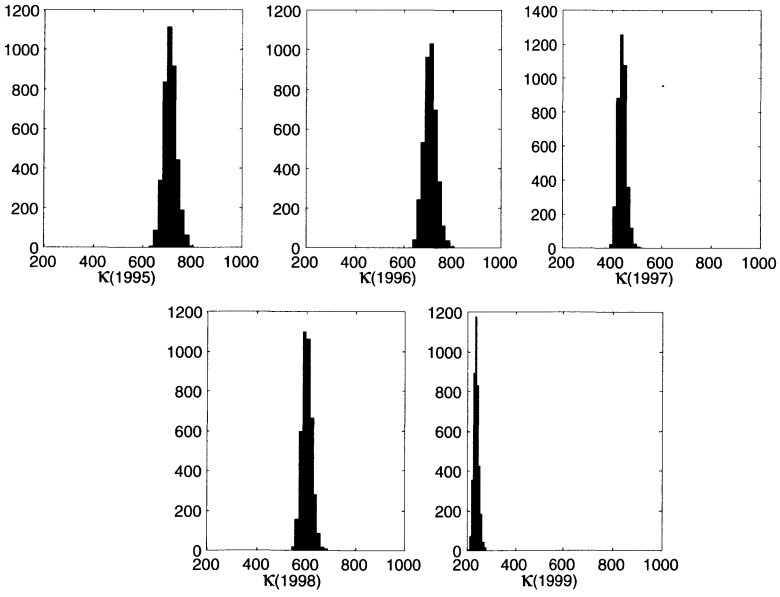

FIGURE 14
Lois marginales a posteriori pour la taille du stock chaque année de la période 1995-1999 (paramètre κ_t)

5. Discussion

La mise au point d'un modèle statistique est un acte des plus subjectifs, que ce soit en analyse statistique Bayésienne ou classique. Le cas du Scorff illustre comment la perspective Bayésienne tire parti de structures conditionnelles représentées par un modèle graphique (graphe acyclique orienté). La modélisation conditionnelle se déroule très simplement après que les variables latentes, les paramètres modèles et les variables observées aient été identifiés. Ces trois éléments constituants donnent beaucoup de liberté pour créer les modèles de représentation d'un problème réel. Conséquence de la structure conditionnelle, au grand dam de certains statisticiens, l'écriture de la vraisemblance complète est inutile. Encourant de plus la malédiction de certains Bayésiens, on se passe aussi de l'écriture de la formule de Bayes pour tout le vecteur des paramètres.

L'expertise *a priori* et toutes les données relatives au problème – même si elles ne font pas partie du dispositif expérimental – fournissent une information précieuse utilisable pour réduire l'incertitude. Dans l'exemple des saumons du Scorff , la taille du stock de castillons d'année en année et leurs intervalles de crédibilité peuvent être évalués en intégrant à l'étude de telles informations. Les études Bayésiennes présentent l'avantage d'identifier clairement les hypothèses *a priori* qui peuvent alors être critiquées. D'un point de vue méthodologique, c'est là que réside la principale différence entre les statistiques conventionnelles et les statistiques Bayésiennes. Par exemple, le paramètre δ fait apparaître un conflit entre le prior et la loi *a posteriori*. Cette discordance apporte au statisticien des éléments de modification. Deux diagnostics sont possibles : (i) le prior issu des connaissances locales n'est pas représentatif de la situation à la lumière des données, (ii) le modèle est mal déterminé à certains égards. Ces alternatives devront être vérifiées et pourront servir de point de départ à une analyse plus fine et plus poussée. Néanmoins il n'existe qu'une modeste littérature sur l'élicitation de priors. Les efforts doivent être poursuivis pour améliorer la prise en compte du savoir *a priori* dans l'élicitation de priors, en suivant les recommandations de Berger (1980). Les méthodes pour tester de façon approfondie comment les hypothèses de structure du prior contraignent les probabilités *a priori* ne sont pas abordées dans cette application du Scorff mais les lecteurs intéressés peuvent se reporter à Gelman *et al.* (1995).

L'hypothèse de stationnarité des paramètres $(\pi, \theta, \alpha, \beta, \tau, \delta)$ du modèle est très discutable. On surestime sans doute la précision des estimateurs du nombre de géniteurs qui remontent la rivière. Faire l'hypothèse d'un comportement interannuel stationnaire est certes commode et parcimonieux mais c'est une simplification bien osée du monde réel. Imaginons que les paramètres varient d'une année sur l'autre. Une certaine proximité interannuelle peut-être conservée en supposant que les valeurs annuelles des paramètres proviennent d'une même distribution. Une telle structure hiérarchique donne un modèle souvent plus réaliste et fournit un moyen classique de représenter la surdispersion : d'une année à l'autre les conditions environnementales peuvent varier, mais il reste des traits communs qui garantissent une certaine cohérence : le type de pêche pratiquée, le protocole expérimental, la population de saumons, etc... En s'appuyant sur des analyses additionnelles, non reportées ici, nous considérons néanmoins qu'avec seulement six années de données, la complexité

supplémentaire introduite par une structure hiérarchique ne vaut pas la peine car, dans ce cas particulier à peu de données, elle n'apporte guère de bénéfice en terme de qualité de l'estimation, de compréhension du modèle et d'interprétation écologique.

Il y a bien plus de paramètres dans le cas d'étude que de données rassemblées au cours d'une année d'observation. Cet excès de «paramètrisation» est dangereux car il peut engendrer des indéterminations et un effet de confusion du rôle des paramètres, pour lequel les procédures d'estimations classiques ne fournissent pas de réponse générale. Dans la procédure Bayésienne, grâce à l'utilisation des priors, le calcul d'inférence se fait toujours selon une même logique, et la faiblesse de structure des modèles est révélée à l'examen comparé du prior et du posterior. Pour sûr, le modèle décrit par le système d'équation (3) est surparamétré puisqu'aucune information, à part les priors, ne permet de faire l'inférence séparée de α et de π à partir des données : c'est seulement le produit $\alpha\ \pi$ qui compte pour expliquer les données. Le cadre Bayésien s'accomode d'un tel lien entre un couple de paramètres. Plus généralement, la matrice de variance-covariance entre les paramètres permet de détecter quels paramètres produisent des effets confondus, mais même une sévère confusion comme ci-dessus, n'est pas un problème pour conduire l'inférence Bayésienne. La modélisation en écologie repose d'ailleurs en équilibre instable entre des modèles réalistes mais souvent surparamétrés et des modèles parcimonieux trop rustiques ou avec des coefficients de réglage dont les valeurs ont été imposées par la littérature sans possibilité de validation réelle. C'est bien souvent en fixant certains paramètres que l'approche statistique fréquentiste surmonte les problèmes engendrés par la confusion des effets. L'approche Bayésienne fournit un moyen intelligent et cohérent de sortir du dilemne précédent en s'appuyant sur des priors fondés sur l'expertise du praticien.

Les techniques d'estimations Bayésiennes par MCMC, et particulièrement l'échantillonneur de Gibbs sont les outils appropriés pour les modèles conditionnels graphiques. L'utilisation de variables latentes permet de représenter par simulation des événements cachés ou inobservés («données» manquantes) en complément de l'échantillon systématique des données observées. Les praticiens comprennent généralement fort bien la signification des variables latentes. Il leur semble de plus naturel, d'introduire dans le modèle des variables qu'ils savent interpréter, même si elles sont cachées. L'approche Bayésienne resserre les liens entre théoriciens et scientifiques de terrain : la modélisation graphique est un outil de communication pour se mettre d'accord sur la structure d'un modèle. D'ailleurs, tester différentes propositions de modèles n'est pas un problème : avec les techniques MCMC, l'inférence statistique Bayésienne – c-a-d obtenir la distribution *a posteriori* des paramètres et en tirer les caractéristiques statistiques – se conduit souvent en temps réel. Le praticien peut suggérer tout type de modèle et de prior et même l'élaborer par étape. L'énergie et le temps ainsi gagnés peuvent être réinvestis pour travailler sur le réalisme du modèle (dans cet exemple du Scorff, décrire au mieux le comportement du poisson) et sur l'étude des diverses sources d'incertitude.

Remerciements

Notre reconnaissance va à Jacques Bernier et Lucien Duckstein pour les nombreuses discussions et relectures ainsi qu'à Pierre Cazes pour ses corrections d'épreuves méticuleuses et avisées. L'étude du Scorff est issue d'un projet commun entre l'Institut de Recherche Agronomique (INRA), le Conseil Supérieur de la Pêche, et La Fédération de Pêche et de Protection des écosystèmes aquatiques du Morbihan. La collecte des informations sur le terrain a été effectuée par les techniciens de la station expérimentale du Moulin des Princes, Nicolas Jeannot et François Burban, aidés de Jean-Yves Moelo.

Références

BAGLINIÈRE J.-L., CHAMPIGNEULLE A. (1986), Population estimates of juvenile Atlantic salmon, Salmo salar, as indices of smolt production. J. Fish Biol. 29 467-482.

BERGER J.O. (1980), Statistical decision theory and Bayesian analysis. Springer Verlag, New York.

BERNARDO J.M., SMITH A.F.M. (1994), Bayesian theory. John Wiley and Sons, Londres.

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

BROOKS S.P. (1998), Markov chain Monte Carlo method and its application. The Statistician 47(1) 69-100.

CASELLA G., GEORGE E. I. (1992), Explaining the Gibbs Sampler. The American Statistician 46(3) 167-174.

CLOBERT J., PRADEL R. (1993), Modelling some demographic parameters in animal populations studied by capture-mark-recapture : review and perspectives *In* Biometrie et Environnement. *Edited by* J-D. Lebreton and B. Asselain. Masson : Paris. pp.151-174.

GELMAN A., RUBIN D.B. (1992), Inference from iterative simulation using multiple sequences. Statist. Sci. 7 457-511.

GELMAN A., CARLIN J.B., STERN H.S., RUBIN D.B. (1995), Bayesian data analysis. Chapman and Hall, Londres.

GEMAN S., GEMAN D. (1984), Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6 721-741.

GILKS W.R., THOMAS A., SPIELGELHALTER D. J. (1994), A language and program for complex Bayesian modelling. Statistica 43 169-178.

GUYON X. (1995), Random Fieds on a network : modeling, statistics and applications, Springer.

HASTINGS W. K. (1970), Monte Carlo sampling methods using Markov chains and their application. Biometrika 57 97-109.

KASS R.E., CARLIN B.P., GELMAN A., NEAL R.M. (1996), Markov Chain Monte Carlo in practice : a roundtable discussion. Proceedings of the Joint Statistical Meetings. pp.1–8.

METROPOLIS N., ROSENBLUTH A. W., ROSENBLUTH M.N., TELLER E. (1953), Equations of state calculations by fast computing machines. J. Chem. Phys. 21 1087-1092.

MEYER R., MILLAR R. B. (1999), BUGS in Bayesian stock Assessments. Can. J; Fish. Aquat. Sci. 56 1078-1087.

MEYER R., MILLAR R. B. (2000), Non-linear State Space Modelling of Fisheries Biomass Dynamics by using Metropolis-Hastings within-Gibbs Sampling. Journal of the Royal Statistical Society C 49 327-342.

POTTER E.C.E., CROZIER W.W. (2000), A perspective on the marine survival of Atlantic salmon. 19-36. The ocean life of Atlantic salmon - Environmental and biological factors influencing survival. In Fishing News Books. *Edited by* Mills D., Blackwell Science, Oxford.

PUNT A. E., HILBORN R. (1997), Fisheries stock assessment and decision analysis : the Bayesian approach. Reviews in Fish Biology and Fisheries 7 35-63.

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

SEBER G.A.F. (1982), The estimation of animal abundance and related parameters. Charles Griffin and Co. Ltd, Londres et High Wycombe.

SPIEGELHALTER D.J., THOMAS A., BEST N. G. (1996a), Computation on Bayesian Graphical Models. In *Bayesian Statistics* 5. pp. 407-425 *Edited by* J.M. Bernardo, J.O. Berger, A. P. Dawid and A.F.M. Smith. Oxford University Press.

SPIEGELHALTER D.J., THOMAS A., BEST N. G., GILKS W.R. (1996b), *BUGS 0.5,* Bayesian Inference using Gibbs Sampling. Manual Cambridge, UK : MRC Biostatistics Unit.

TANNER M.A. (1996), Tools for statistical inference : methods for the exploration of posterior distribution and likelihood functions. Springer Verlag, New York.

Appendice A

Le paramètre κ de la taille du stock formule conditionnelle complète

(θ,α,β) sont des noeuds parents pour κ. (X_1,Y_{uc},Y_{us}) sont des noeuds stochastiques fils issus de κ

H représente les quantités complémentaires de (κ, θ,α,β,X_1,Y_{uc},Y_{us}) dans les noeuds stochastiques.

$$\mathbf{H} = (\kappa, \theta, \alpha, \beta, X_1, Y_{uc}, Y_{us})^- =$$

$$\mathbf{X_0}, X_2, X_3, X_4, X_5, X_6, \tau, \delta, \pi, Y_{mc}, Y_{mr}, Y_{ms}$$

Le théorème de Bayes s'écrit :

$$p(\kappa\,|\mathbf{H},\theta,\alpha,\beta,X_1,Y_{uc},Y_{us}\,) =$$

$$\frac{p(\kappa,X_1,Y_{uc},Y_{us}\,|\mathbf{H},\theta,\alpha,\beta\,)}{\sum\limits_{\kappa} p(\kappa,X_1,Y_{uc},Y_{us}\,|\mathbf{H},\theta,\alpha,\beta\,)}$$

En utilisant la loi de probabilité complète

$$p(\kappa,X_1,Y_{uc},Y_{us}\,|\mathbf{H},\theta,\alpha,\beta\,) =$$

$$p(Y_{uc},Y_{us}\,|\mathbf{H},X_1,\kappa,\theta,\alpha,\beta\,)p(X_1\,|\mathbf{H},\kappa,\theta,\alpha,\beta\,)p(\kappa\,|\mathbf{H},\theta,\alpha,\beta\,)$$

La structure conditionnelle spéciale du graphe orienté (Figure 4) entraîne les simplifications suivantes :

$$p(\kappa,X_1,Y_{uc},Y_{us}|\mathbf{H},\theta,\alpha,\beta) =$$

$$p(Y_{uc}|\mathbf{X_0},X_1,\kappa,\beta)p(Y_{us}|\mathbf{X_0},X_1,Y_{uc},\kappa,\alpha)p(X_1|\mathbf{X_0},\kappa,\theta)p(\kappa|\mathbf{X_0})$$

Pour obtenir la conditionnelle complète de κ, il n'y a qu'à identifier quelles parts de $p(\kappa,X_1,Y_{uc},Y_{us}\,|\mathbf{X_0},\theta,\alpha,\beta\,)$ sont fonctions de κ :

$$p(\kappa\,|\mathbf{H},\theta,\alpha,\beta,X_1,Y_{uc},Y_{us}\,) \propto p(\kappa,X_1,Y_{uc},Y_{us}\,|\mathbf{H},\theta,\alpha,\beta\,)$$

En repérant les termes qui concernent κ dans les diverses lois de probabilité rencontrées dans $p(\kappa,X_1,Y_{uc},Y_{us}\,|\mathbf{H},\theta,\alpha,\beta\,)$ on voit que :

$$p(\kappa\,|\mathbf{H},\theta,\alpha,\beta,X_1,Y_{uc},Y_{us}\,) \propto \frac{\Gamma(1+\kappa-X_1)}{\Gamma(1+\kappa-X_1-Y_{uc})}(1-\beta)^{\kappa-X_1-Y_{uc}}\beta^{Y_{uc}}$$
$$\times\, p(Y_{us}\,|\mathbf{X_0},X_1,Y_{uc},\kappa,\alpha\,)p(X_1\,|\mathbf{X_0},\kappa,\theta\,)p(\kappa\,|\mathbf{X_0}\,)$$

$$p(Y_{us}\,|\mathbf{X_0},X_1,Y_{uc},\kappa,\alpha\,)p(X_1\,|\mathbf{X_0},\kappa,\theta\,)p(\kappa\,|\mathbf{X_0}\,)$$

$$\propto \frac{\Gamma(1+\kappa-X_1-Y_{uc})}{\Gamma(1+\kappa-X_1-Y_{uc}-Y_{us})}(1-\alpha)^{\kappa-X_1-Y_{uc}-Y_{us}}\alpha^{Y_{us}}$$
$$\times\, \frac{\Gamma(1+\kappa)}{\Gamma(1+\kappa-X_1)}\theta^{X_1}(1-\theta)^{\kappa-X_1}p(\kappa\,|\mathbf{X_0}\,)$$

D'après la distribution d'échantillonnage de Y_{us} voir équation (4), $\kappa \geqslant X_1 + Y_{uc} + Y_{us}$ la constante normalisante est obtenue en intégrant sur toutes les valeurs possibles de κ

$$p(\kappa\,|\mathbf{H},\theta,\alpha,\beta,X_1,Y_{uc},Y_{us}) = \frac{(1-\beta)^{\kappa}(1-\alpha)^{\kappa}(1-\theta)^{\kappa}p(\kappa\,|\mathbf{X_0})\Gamma(1+\kappa)/\Gamma(1+\kappa-X_1-Y_{uc}-Y_{us})}{\sum_{\kappa=X_1+Y_{uc}+Y_{us}}^{4000}(1-\beta)^{\kappa}(1-\alpha)^{\kappa}(1-\theta)^{\kappa}p(\kappa\,|\mathbf{X_0})\Gamma(1+\kappa)/\Gamma(1+\kappa-X_1-Y_{uc}-Y_{us})}$$

Appendice B

Conditionnelle complète pour la variable latente Y_{mc}

On suit le même raisonnement que dans l'appendice A. Réglons d'abord le problème des bornes de variations *a posteriori*. Comme la grandeur Y_{mc} est un nombre d'essai d'une loi binomiale de paramètre τ qui produit Y_{mr} succès, on a la relation $Y_{mc} > Y_{mr}$. Comme la grandeur Y_{mc} est également le nombre de succès d'une loi binomiale de paramètre β avec X_1 essais qui ont produit au moins Y_{ms} échecs, on a la relation $Y_{mc} < X_1 - Y_{ms}$

$$\mathbf{H} = (Y_{mc},\tau,\alpha,\beta,X_1,Y_{mr},Y_{ms})^{-} = \mathbf{X_0},X_2,X_{3,}X_4,X_5,X_6,\theta,\delta,\pi,Y_{uc},Y_{us}$$

$$p(Y_{mc}\,|\mathbf{H},\tau,\alpha,\beta,X_1,Y_{mr},Y_{ms}) = \frac{p(Y_{mc},Y_{mr},Y_{ms}\,|\mathbf{H},\tau,\alpha,\beta,X_1)}{\sum_{Y_{mc}} p(Y_{mc},Y_{mr},Y_{ms}\,|\mathbf{H},\tau,\alpha,\beta,X_1)}$$

$$p(Y_{mc},Y_{mr},Y_{ms}\,|\mathbf{H},\tau,\alpha,\beta,X_1) = p(Y_{ms}\,|\mathbf{H},\tau,\alpha,\beta,X_{1,}Y_{mc})p(Y_{mr}\,|\mathbf{H},\tau,\alpha,\beta,X_{1,}Y_{mc})p(Y_{mc}\,|\mathbf{H},\tau,\alpha,\beta,X_1)$$

$$p(Y_{mc},Y_{mr},Y_{ms}\,|\mathbf{H},\tau,\alpha,\beta,X_1) = p(Y_{ms}\,|\mathbf{X_0},\alpha,X_{1,}Y_{mc})p(Y_{mr}\,|\mathbf{X_0},\tau,Y_{mc})p(Y_{mc}\,|\mathbf{X_0},\beta,X_1)$$

$$p(Y_{mc}\,|\mathbf{H},\tau,\alpha,\beta,X_1,Y_{mr},Y_{ms}) \propto p(Y_{ms}\,|\mathbf{X_0},\alpha,X_{1,}Y_{mc})p(Y_{mr}\,|\mathbf{X_0},\tau,Y_{mc})p(Y_{mc}\,|\mathbf{X_0},\beta,X_1)$$

$$p(Y_{mc}\,|\mathbf{H},\tau,\alpha,\beta,X_1,Y_{mr},Y_{ms}) \propto \frac{\Gamma(1+X_1-Y_{mc})}{\Gamma(1+X_1-Y_{mc}-Y_{ms})}(1-\alpha)^{X_1-Y_{mc}-Y_{ms}}\alpha^{Y_{ms}} \times p(Y_{mr}\,|\mathbf{X_0},\tau,Y_{mc})p(Y_{mc}\,|\mathbf{X_0},\beta,X_1)$$

$$p(Y_{mr}\,|\mathbf{X_0},\tau,Y_{mc})p(Y_{mc}\,|\mathbf{X_0},\beta,X_1) \propto \frac{\Gamma(1+Y_{mc})}{\Gamma(1+Y_{mc}-Y_{mr})}(1-\tau)^{Y_{mc}-Y_{mr}}\tau^{Y_{mr}}\frac{(1-\beta)^{X_1-Y_{mc}}(\beta)^{Y_{mc}}\Gamma(1+X_1)}{\Gamma(1+Y_{mc})\Gamma(1+X_1-Y_{mc})}$$

$$p(Y_{mc}\,|\mathbf{H},\tau,\alpha,\beta,X_1,Y_{mr},Y_{ms}) = \frac{\frac{[(1-\beta)(1-\alpha)]^{-Y_{mc}}}{\Gamma(1+X_1-Y_{mc}-Y_{ms})}\frac{[(1-\tau)(\beta)]^{Y_{mc}}}{\Gamma(1+Y_{mc}-Y_{mr})}}{\sum_{Y_{mc}=Y_{mr}}^{Y_{mc}=X_1-Y_{ms}}\frac{[(1-\beta)(1-\alpha)]^{-Y_{mc}}}{\Gamma(1+X_1-Y_{mc}-Y_{ms})}\frac{[(1-\tau)(\beta)]^{Y_{mc}}}{\Gamma(1+Y_{mc}-Y_{mr})}}$$

Conditionnelle complète pour la variable latente Y_{ms}

En observant sur la figure 4 comment la grandeur Y_{ms} est «prise en sandwich» entre des tirages binomiaux, on établit ses bornes de variation :

$$X_1 - Y_{mc} = Y_{mf} \geqslant Y_{ms} \geqslant X_5$$

Le tableau 3 nous dit que seuls $Y_{mc}, Y_{uc}, X_1, \alpha, \pi, X_5$ interviennent dans la conditionnelle complète de Y_{ms}. Un examen attentif de la figure 4 révèle que Y_{mc}, Y_{uc}, X_1 interviennent par la médiation du noeud déterministe $Y_{mf} = X_1 - Y_{mc}$ qui agit comme un parent (avec α) du noeud Y_{ms} sur lequel on focalise l'attention :

$$p(Y_{ms} \,|(Y_{ms})^-) = p(Y_{ms} \,|Y_{mc}, Y_{uc}, X_1, \alpha, \pi, X_5) = p(Y_{ms} \,|Y_{mf}, \alpha, \pi, X_5)$$

L'équation (9) s'applique avec la binomiale $p(Y_{ms} \,|Y_{mf}, \alpha)$ jouant le rôle de prior et la binomiale $p(X_5 \,|Y_{ms}, \pi)$ jouant le rôle de vraisemblance :

$$p(Y_{ms} \,|(Y_{ms})^-) \propto p(Y_{ms} \,|Y_{mf}, \alpha) \times p(X_5 \,|Y_{ms}, \pi)$$

Ne retenant dans ces expressions que les termes où intervient Y_{ms} on trouve :

$$p(Y_{ms} \,|(Y_{ms})^-) \propto \left[\frac{\alpha\,(1-\pi)}{(1-\alpha)}\right]^{Y_{ms}} \frac{1}{\Gamma(1+Y_{mf}-Y_{ms})\Gamma(1+Y_{ms}-X_5)}$$

$$p(Y_{ms} \,|(Y_{ms})^-) = \frac{\left[\frac{\alpha(1-\pi)}{(1-\alpha)}\right]^{Y_{ms}} \frac{1}{\Gamma(1+Y_{mf}-Y_{ms})\Gamma(1+Y_{ms}-X_5)}}{\sum_{y=X_5}^{Y_{mf}} \left[\frac{\alpha(1-\pi)}{(1-\alpha)}\right]^{y} \frac{1}{\Gamma(1+Y_{mf}-y)\Gamma(1+y-X_5)}}$$

Conditionnelle complète pour la variable latente Y_{mr}

Sur la figure 4, on voit que la grandeur Y_{mr} est prise en «sandwich binomial» entre X_2 et Y_{mc} tandis que la grandeur $Y_{ur} = X_4 - Y_{mr}$ est "coincée" de la même manière entre X_3 et Y_{uc} . On en déduit les inégalités conditionnelles :

$$Max(X_4 - Y_{uc}, X_2) \leqslant Y_{mr} \leqslant Min(Y_{mc}, X_4 - X_3)$$

La table 3 dit que seuls $Y_{mc}, Y_{uc}, X_4, X_2, X_3, \tau, \delta$ interviennent dans la conditionnelle complète de Y_{mr}.

$$p(Y_{mr} \,|(Y_{mr})^-) = p(Y_{mr} \,|(Y_{mc}, Y_{uc}, X_4, X_2, X_3, \tau, \delta))$$

L'examen de la figure 4 et la mise en pratique de l'équation (9) montrent que les binomiales $p(Y_{mr} \,|Y_{mc}, \tau)$ et $p((X_4 - Y_{mr}) \,|Y_{uc}, \tau)$ jouent le rôle de prior

tandis que les binomiales $p(X_2\,|Y_{mr},\delta\,)$ et $p(X_3\,|(X_4-Y_{mr}),\delta\,)$ jouent le rôle de la vraisemblance.

$$p(Y_{mr}\,\Big|(Y_{mr})^{-}\,) \propto p(Y_{mr}\,|Y_{mc},\tau\,)p((X_4-Y_{mr})\,|Y_{uc},\tau\,)p(X_2\,|Y_{mr},\delta\,)$$
$$p(X_3\,|(X_4-Y_{mr}),\delta\,)$$

En faisant apparaître ces opérations comme une fonction de Y_{mr}, on trouve :

$$p(Y_{mr}\,\Big|(Y_{mr})^{-}\,) \propto \frac{[\Gamma(1+Y_{mr}-X_2)\Gamma(1+X_4-X_3-Y_{mr})]^{-1}}{\Gamma(1+Y_{mc}-Y_{mr})\Gamma(1+Y_{uc}-X_4+Y_{mr})}$$

$$p(Y_{mr}\,\Big|(Y_{mr})^{-}\,) = \frac{\frac{[\Gamma(1+Y_{mr}-X_2)\Gamma(1+X_4-X_3-Y_{mr})]^{-1}}{\Gamma(1+Y_{mc}-Y_{mr})\Gamma(1+Y_{uc}-X_4+Y_{mr})}}{\sum\limits_{y=Max(X_4-Y_{uc},X_2)}^{Min(Y_{mc},X_4-X_3)} \frac{[\Gamma(1+y-X_2)\Gamma(1+X_4-X_3-y)]^{-1}}{\Gamma(1+Y_{mc}-y)\Gamma(1+Y_{uc}-X_4+y)}}$$