UTILISATION D'UNE FONCTION DE PERTE QUADRATIQUE PONDÉRÉE DANS LES APPROCHES BAYÉSIENNES : APPLICATION À LA CARTOGRAPHIE D'INDICATEUR DE SANTÉ

L. FORTUNATO [1 2], A. LIVERNEAUX[1 2], D. HÉMON[1 2]
C. GUIHENNEUC-JOUYAUX[1 2 3]

RÉSUMÉ

Position du problème

Dans le cadre des modèles écologiques, nous comparons différents estimateurs bayésiens des risques relatifs en utilisant un modèle de cartographie hiérarchique bayésien.

Méthodes

Les estimateurs ponctuels bayésiens dépendent du choix de la fonction de perte. La plus « classique » est la fonction de perte quadratique et sa minimisation amène à l'estimation par moyenne *a posteriori*. Une discussion est menée sur les avantages de cette approche, notamment en envisageant différents objectifs comme l'intérêt particulier porté sur les valeurs extrêmes des risques relatifs lors de cartographie. Ces objectifs guident le choix de la fonction de perte. Trois fonctions sont envisagées : la fonction de perte quadratique, la fonction de perte en valeur absolue et une fonction de perte quadratique pondérée avec des poids de type exponentiel. La dernière fonction correspond à un choix moins fréquent avec une pondération importante sur les risques relatifs extrêmes.

Résultats

Une étude par simulations a été réalisée en considérant trois modèles de risques relatifs. Les deux premiers correspondent à des surfaces de risques discontinues alors que le dernier cas présente une décroissance du risque sur le territoire selon un gradient géographique. Les performances des estimateurs sont évaluées par l'erreur quadratique et le biais relatif ainsi que la capacité à détecter les zones à risques élevés. Les résultats montrent que de manière générale, les différents estimateurs des risques relatifs sont convergents avec de bonnes performances. Néanmoins, cette étude permet de montrer que dans le cas de risques relatifs discontinus, le choix d'une fonction de perte de type exponentiel permet d'obtenir une amélioration souvent significative de l'estimation des risques relatifs.

1. INSERM, U754, 16 av Paul Vaillant-Couturier, F-94807 Villejuif cedex, France.

2. Université Paris Sud, IFR69, 16 av Paul Vaillant-Couturier, F-94807 Villejuif cedex, France.

3. CNRS UMR 8145, UFR Biomédicale, Université Paris 5, 45 rue des Saints-Pères, F-75006 Paris, France, email : chantal.guihenneuc@univ-paris5.fr

Conclusion

Le choix d'une fonction de perte et l'étude de l'estimateur qui en découle constituent une démarche rarement utilisée et pourtant prometteuse dans les approches bayésiennes. Ce choix peut être guidé par l'objectif recherché comme ici, la mise en valeur des risques relatifs extrêmes. Cette étude a montré de légères différences entre les estimateurs, prouvant dans ce sens leur robustesse mais avec une amélioration souvent significative de l'estimateur pondéré. D'autres configurations de risques relatifs et/ou de nombre de cas attendus pourraient peut-être mettre à jour des différences plus importantes.

Mots-clés : Modèle bayésien hiérarchique, Fonctions de perte, Risques relatifs, Modèle écologique de Poisson.

ABSTRACT

Background

In the context of ecological model, we compare different Bayes estimates of relative risks by disease mapping model.

Methods

Bayesian estimates are based on the choice of loss function. A usual choice is the quadratic loss function and its minimization leads to the posterior mean. In this study, a discussion is done on the advantages of this choice, in particular by considering different objectives as focussing on extreme relative risks in disease mapping. These objectives determine the choice of the loss function. Three different loss functions are considered : the quadratic loss, the absolute loss and a weigthed quadratic loss function with weights having an exponential shape. The last one corresponds to an unusual choice with important weights on extreme relative risks.

Results

A study based on simulations is done with three different cases of relative risks in France : two first cases correspond to discontinuous risk surfaces while the last case presents continuous risk decrease following a geographic gradient. Estimates performances are evaluated by mean quadratic error and bias as well as ability of units detection with high relative risks. Results show that the different Bayesian estimates are coherent with good performances. Nevertheless, often significative improvements of estimates based on exponential loss function are found in the case of discontinuous risk surface.

Conclusion

The choice of loss function and the study of the associated estimate are rarely done and nevertheless interesting in Bayesian approach. This choice is based on a specific objective as, in geographical analyse, a particular attention on extreme relative risks. In this study, slight differences between the three considered loss functions are found showing in this sense their robustness but with often significant improvements of weighted estimate. Different configurations of relative risks and/or expected number of cases could give more important differences.

Keywords : Hierarchical Bayesian model, Loss functions, Relative risks, Ecological Poisson model.

1. Introduction

La disponibilité d'indicateurs de santé (morbidité ou mortalité) repérés géographiquement a favorisé le développement de la représentation cartographique de tels indicateurs ayant pour objectif la description et la compréhension de leurs variations spatiales, la suggestion d'hypothèses étiologiques, la surveillance de zones à hauts risques... La production d'atlas de mortalité (qui existe depuis longtemps) procure un outil visuel résumant globalement les variations des indicateurs de santé sur l'ensemble d'un domaine. Dans le cas de petits effectifs (nombres attendus faibles de cas ou de décès), une approche par un modèle de Poisson est alors souvent adoptée. Plus précisément, le nombre observé dans chaque zone géographique est supposé suivre une loi de Poisson dont la moyenne est le produit entre le nombre attendu dans cette zone (standardisé potentiellement sur la structure d'âge, sur le sexe...) et un paramètre spécifique à chaque zone, le risque relatif. Ainsi, la valeur du risque relatif mesure « l'éloignement » entre le nombre observé et le nombre attendu de décès. Une zone est alors considérée à haut risque si le risque relatif associé à cette zone est nettement plus grand que la valeur 1. La représentation cartographique des risques relatifs permet ainsi de visualiser les zones où le nombre observé de décès est différent en moyenne du nombre attendu. De plus, la proximité géographique de telles zones peut suggérer le partage de caractéristiques communes et donc guider la recherche de facteurs de risque environnementaux. Dans le cas de l'étude d'une pathologie rare, une représentation cartographique « naïve » peut être trompeuse sans distinction entre de vraies et apparentes zones à hauts risques. En effet, l'estimateur du maximum de vraisemblance des risques relatifs de chaque zone est le rapport du nombre observé sur le nombre attendu de la zone, la variance de cet estimateur étant inversement proportionnelle au nombre attendu. Le nombre attendu pouvant être faible, la variance est alors élevée [4 , 12]. Ainsi, une représentation graphique directe des risques relatifs estimés sous ce modèle peut produire l'apparence visuelle de zones géographiques à hauts risques de par cette imprécision. Ce problème est, en particulier, dû au fait de considérer les risques indépendamment d'une zone géographique à l'autre sans prendre en compte une structure globale inter-zone sur le domaine étudié. Les modèles hiérarchiques permettent d'analyser les variations géographiques d'événements rares sur des domaines irréguliers [3, 4, 9]. Le premier niveau modélise la variabilité locale (intra zone) en faisant l'hypothèse, comme précédemment, que le nombre d'événements observés dans chaque zone suit une loi de Poisson dont le paramètre dépend du risque relatif propre à la zone géographique. Le second niveau modélise la variabilité spatiale de ces risques relatifs entre les zones du domaine étudié. Cette structure hiérarchique est essentielle pour une bonne estimation des risques relatifs et permet l'estimation conjointe des variabilités locales et globales. Dans les approches bayésiennes, les paramètres sont considérés comme aléatoires, et l'objectif est de déterminer analytiquement si cela est possible, ou sinon approximativement, leur loi *a posteriori* qui est le résultat de la combinaison entre les lois a priori et la vraisemblance (l'information provenant des données). A partir de ces lois *a posteriori*, différentes inférences statisti-

ques peuvent être menées (moyenne, quantiles, variance ...). Dans le domaine de la cartographie, il est particulièrement important d'obtenir des estimations ponctuelles des paramètres (en l'occurrence des risques relatifs) afin de pouvoir faire une représentation fiable sur le domaine étudié de la variabilité spatiale de l'indicateur de santé. Nous comparons ici différents estimateurs ponctuels bayésiens. Ces derniers sont basés sur le choix d'une fonction de perte et la minimisation de la fonction de perte moyenne détermine l'estimateur associé. La fonction de perte usuelle est la fonction de perte quadratique et son estimateur associé, la moyenne *a posteriori*. D'autres fonctions de perte peuvent être envisagées notamment pour répondre à des objectifs spécifiques. En cartographie, on peut vouloir obtenir des estimateurs particulièrement robustes pour les valeurs extrêmes des risques relatifs. Il est intéressant alors d'envisager d'autres formes de fonctions de perte mettant une pénalité plus forte sur les valeurs élevées des risques relatifs [15] . L'objectif de notre travail est d'étudier les différences entre les estimateurs bayésiens des risques relatifs en fonction du choix de la fonction de perte dans différentes situations caractéristiques des risques relatifs (surface continue ou pas des vrais risques relatifs sous-jacents) pour des valeurs de risques relatifs cohérentes avec celles rencontrées lors de l'étude de pathologie rare, ici fixées au maximum à 1,7.

2. Matériel et méthodes

2.1. Estimation des risques relatifs

Notre étude porte sur le domaine irrégulier de la France Métropolitaine, Corse exclue, découpée en n zones géographiques, les départements ($n = 94$ du fait de l'exclusion de la Corse). Dans le cas d'une maladie rare et non contagieuse, le nombre observé de cas dans chaque zone i, X_i $(i = 1, ..., n)$ est modélisé par une loi de Poisson de paramètre $E_i\theta_i$, où θ_i est le risque relatif associé à la zone i et E_i est le nombre attendu de cas dans la zone i. Le nombre de cas attendus dans une zone pouvant être faible, nous utilisons un modèle hiérarchique bayésien. L'intérêt de la modélisation hiérarchique réside en une décomposition claire et structurée de la variabilité des nombres observés via les distributions conditionnelles, ceci permettant de guider l'interprétation des différents paramètres. Ainsi, dans ce modèle, la variabilité est décomposée en une variabilité locale intra-zone poissonnienne au premier niveau et en une variabilité inter-zone log-normale au second niveau du modèle. Le modèle est le suivant :

$$X_i|\theta_i \sim P(E_i\theta_i)$$
$$Log(\theta_i)|\beta, \varepsilon_i = \beta + \varepsilon_i$$

pour $i = 1, ..., n$ et $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \sim N(0_n, \Sigma)$ où Σ est une matrice de variance-covariance permettant d'introduire les autocorrélations résiduelles entre les zones. Le troisième niveau de ce modèle définit les lois *a priori* des paramètres introduits dans les niveaux précédents qui seront précisées ultérieurement.

Dans les analyses statistiques utilisant des méthodes bayésiennes, les inférences bayésiennes d'un ensemble de paramètres $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)^t$ sont basées sur la loi jointe *a posteriori*. Une estimation ponctuelle « optimale » demande la définition d'une fonction de perte adaptée, puis les paramètres sont estimés comme étant les valeurs minimisant l'espérance *a posteriori* de cette fonction de perte. Soit $L(\theta, \hat{\theta})$ une fonction de perte et X les données, nous cherchons $\hat{\theta}^* = (\hat{\theta}_1, \ldots, \hat{\theta}_n^*)$ tel que :

$$E\big(L(\theta, \hat{\theta}^*)|X\big) = \min_{\hat{\theta}} \; E\big(L(\theta, \hat{\theta})|X\big) = \min_{\hat{\theta}} \int L(\theta, \hat{\theta}) p(\theta|X) d\theta$$

2.1.1. Estimateurs classiques

La moyenne *a posteriori* et la médiane *a posteriori* sont deux estimateurs usuels « optimaux » obtenus en minimisant respectivement l'espérance *a posteriori* de la fonction de perte quadratique, $L_q(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ et l'espérance *a posteriori* de la fonction de perte valeur absolue, $L_a(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$. Cependant, les estimateurs obtenus par ce procédé sont souvent biaisés, et dans le cadre des analyses géographiques, il a été montré que les moyennes *a posteriori* entraînent une sous-estimation des écarts extrêmes [6, 7] : le plus grand risque relatif est sous-estimé, et le plus petit est surestimé.

2.1.2. Estimateurs pondérés

Wright D. L., Stern H. S., et Cressie N. [15] ont proposé une nouvelle famille de fonctions de perte quadratique *pondérée*, basée sur les rangs, ayant pour objectif de mieux estimer les valeurs extrêmes des paramètres. Cette fonction de perte, notée L_{p-q}, assigne un poids c_j à la j$^{\text{ème}}$ statistique d'ordre. Soit $c = (c_1, ..., c_n)^t$, le vecteur des poids choisis, la fonction de perte s'écrit de la manière suivante :

$$\begin{aligned} L_{p-q}(\theta, \hat{\theta}; c) &= \sum_{k=1}^{n} \sum_{j=1}^{n} c_j 1_{\{\theta_k = \theta_{(j)}\}} (\theta_k - \hat{\theta}_k)^2 = \sum_{k=1}^{n} c_{r(k)} (\theta_k - \hat{\theta}_k)^2 \qquad (1) \\ &= \sum_{k=1}^{n} L_{p-q,k}(\theta_k, \hat{\theta}; c) \end{aligned}$$

où $r(k) = \big\{j : \theta_k = \theta_{(j)}; j \in \{(1, \ldots, n\}\big\}$ identifie le rang de θ_k et $c_{r(k)} = \sum_{j=1}^{n} c_j 1_{\{\theta_k = \theta_{(j)}\}}$ identifie le poids attribué à ce rang. Nous remarquons que cette fonction de perte est en fait une somme de fonctions de perte. Il est donc possible et plus simple de ne considérer que le terme k de la sommation, $L_{p-q,k}(\theta_k, \hat{\theta}_k; c)$. L'estimateur optimal $\hat{\theta}_k^*$, selon cette fonction de perte, est,

pour $k = 1, ..., n$:

$$\hat{\theta}_k^* = \frac{\sum_{j=1}^{n} c_j \int 1_{\{\theta_k = \theta_{(j)}\}} \theta_k p(\theta|X) d\theta}{\sum_{j=1}^{n} c_j \int 1_{\{\theta_k = \theta_{(j)}\}} p(\theta|X) d\theta} \tag{2}$$

$$= \frac{\sum_{j=1}^{n} c_j E(\theta_k|\theta_k = \theta_{(j)}, X) P(\theta_k = \theta_{(j)}|X)}{\sum_{j=1}^{n} c_j P(\theta_k = \theta_{(j)}|X)}$$

En faisant varier k de 1 à n, dans l'expression (2), nous obtenons $\hat{\theta}^*$ qui est optimal selon la fonction de perte L_{p-q} donnée en (1).

Le choix des poids pour L_{p-q} est arbitraire. Une famille paramétrique est proposée correspondant à des fonctions « en forme de cuvette », qui porteront plus d'intérêt aux valeurs extrêmes comme c'est souvent le cas pour des applications dans la cartographie des maladies. De telles fonctions peuvent être exprimées sous la forme d'un mélange de fonctions exponentielles :

$c_j = \exp\left[a_1 \left\{j - \frac{n+1}{2}\right\}\right] + \exp\left[a_2 \left\{j - \frac{n+1}{2}\right\}\right]$ pour $j = 1, ..., n$, où a_1 est un nombre réel positif, et a_2 est négatif.

D'après l'expression (2) de l'estimateur $\hat{\theta}_k^*$, nous notons que les estimateurs optimaux sous L_{p-q} sont invariants par changement d'échelle des poids. Les poids peuvent ainsi être définis proportionnellement à une constante.

Dans le cadre de notre étude, l'objectif est de mieux estimer les valeurs maximales des risques relatifs, ceci afin d'identifier des zones à hauts risques pour guider la recherche de facteurs environnementaux. Nous avons choisi pour cela d'utiliser une fonction de poids croissante donnant ainsi plus d'importance aux valeurs élevées des risques relatifs. Une fonction de perte a été fixée, L'_{p-q}, basée sur le vecteur poids $c_j = \exp\left[\frac{j}{10}\right]$. Comme nous en avions fait la remarque un peu plus haut, le vecteur des poids reste invariant par changement d'échelle. Ce vecteur de poids correspond ainsi au mélange « limite » de fonctions exponentielles avec $a_1 = \frac{1}{10}$ et $a_2 \to -\infty$. Ce vecteur donne très peu de poids aux petits ordres statistiques mais met l'accent sur les plus élevés. La figure 1 représente cette fonction de poids (standardisée à une constante près) en fonction des rangs j ainsi que la fonction de poids constante égale à 1 utilisée dans la fonction de perte quadratique classique.

Nous voyons clairement dans cette figure que les poids attribués à la première moitié des paramètres (ordres plus petits que 47 pour $n = 94$) sont très faibles puis augmentent sur la seconde moitié des rangs, donnant ainsi plus d'importance aux valeurs élevées des paramètres. Il est essentiel de représenter

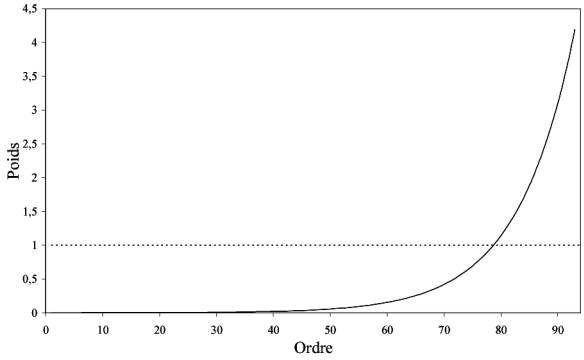

FIG 1. — Comparaison entre les poids associés à la fonction de perte quadratique L_q (- - -) et la fonction de perte quadratique poindérée L_{p-q} (—)

graphiquement les poids en fonction des rangs afin de bien faire un choix qui correspond à l'objectif recherché. Ce choix est arbitraire et nous avons vérifié qu'une légère variation ($a_1 = 1/11$ ou $a_1 = 1/9$) n'avait aucune implication sur les résultats (non montrés). Par contre, un changement important (comme par exemple, $a_1 = 1/2$) a des conséquences non négligeables sur les résultats. Ce dernier choix correspondrait à la volonté de mettre en valeur que les dernières plus grandes valeurs des risques relatifs. Pour ce travail, nous avons préféré présenter les résultats pour un choix moins excessif ($a_1 = 1/10$) où les pénalités sont croissantes dès la deuxième moitié des rangs.

De la fonction de perte L'_{p-q}, nous obtenons l'estimateur pondéré $\hat{\theta}_{p-q}$ défini suivant l'expression donnée en (2). Afin d'étudier les performances de cet estimateur des risques relatifs, nous le comparerons à la moyenne et à la médiane *a posteriori*.

2.1.3. Adaptation

Dans la pratique, les vraies valeurs des paramètres ne sont pas disponibles puisque notre but est justement de les estimer. Le vecteur des poids c basé sur les rangs de ces paramètres est donc impossible à obtenir. Afin de contourner ce problème, nous avons considéré une version adaptée de c, le vecteur c' en remplaçant les rangs des risques par les rangs des estimations des moyennes *a posteriori* des risques. Nous obtenons ainsi un vecteur poids évaluable dans la pratique.

La fonction de perte qui va nous permettre d'obtenir les estimateurs pondérés devient alors :

$$L_{p-q}(\theta, \hat{\theta}; c') = \sum_{k=1}^{n} \sum_{j=1}^{n} c'_j 1_{\{\tilde{\theta}_k - \tilde{\theta}_{(j)}\}} (\theta_k - \hat{\theta}_k)^2 = \sum_{k=1}^{n} c'_{\rho(k)} (\theta_k - \hat{\theta}_k)^2$$

où $\tilde{\theta}_k$ est l'estimation de la moyenne *a posteriori* du risque relatif de la zone k, $\rho(k)$ identifie son rang et $c'_{\rho(k)} = \sum_{j=1}^{n} c'_j 1_{\{\tilde{\theta}_k = \tilde{\theta}_{(j)}\}}$ identifie le poids attribué à ce rang.

2.2. Modèles de simulation

Les données dont nous disposons sont les cas attendus de leucémies, E_i, $i = 1, ..., n$ par département pour les enfants âgés de 0 à 14 ans, en France métropolitaine (sauf la Corse), pendant la période 1990-1998. Ces données proviennent du Registre National des Hémopathies Malignes de l'Enfant, mis en place par Clavel J. [2]. Nous avons utilisé trois séries de risques relatifs correspondant à trois situations géographiques contrastées [7] décrites ci-dessous. Les valeurs des risques relatifs ont été choisies relativement faibles, ceci correspondant à une situation plus réaliste lors de l'étude de pathologie rare. Pour chaque modèle de risque (et donc pour chaque jeu de risques relatifs), le nombre observé de cas X_i dans le département i est simulé selon la loi de Poisson : $X_i|\theta_i \sim P(E_i\theta_i)$.

2.2.1. Modèle de risque « Bloc 4 »

Ce modèle de risque correspond à une situation avec quatre groupes de zones géographiques possédant des risques élevés égaux à 1,5 et un arrière plan avec des risques relatifs à 0,7. En s'inspirant de la problématique épidémiologique concernant l'association entre leucémie de l'enfant et exposition au radon « domestique » [5], les quatre blocs choisis correspondent aux groupes de départements ayant une forte concentration en radon, les anciennes régions volcaniques telles que le Massif Central et les Vosges, ainsi que les régions très granitiques comme la Bretagne, et également dans les Pyrénées. Nous obtenons alors pour le modèle « Bloc 4 » la carte des risques (figure 2 (a)).

2.2.2.Modèle de risque « Nord-Sud »

Le cas « Nord-Sud » (figure 2 (b)) correspond à une division de la carte en deux parties, de telle façon que les départements du Sud aient des risques relatifs égaux à 1,2 et que ceux de la partie Nord soient égaux à 0,8. En comparaison avec le modèle de risque « Bloc 4 », ce modèle suppose à nouveau une discontinuité nette entre les valeurs de risque mais le nombre de zones géographiques est équivalent dans les deux blocs avec une contigüité « intra-bloc » parfaite.

2.2.3.Modèle de risque « Gradient »

Les risques sont simulés dans ce modèle en décroissant linéairement du Sud vers le Nord comme montré sur la figure 2 (c).

Cette situation est la plus défavorable pour les estimateurs $\overline{\theta}_{p-q}$ car la variation des risques relatifs est très progressive de type continue, il n'y a pas de différences brutales entre les valeurs.

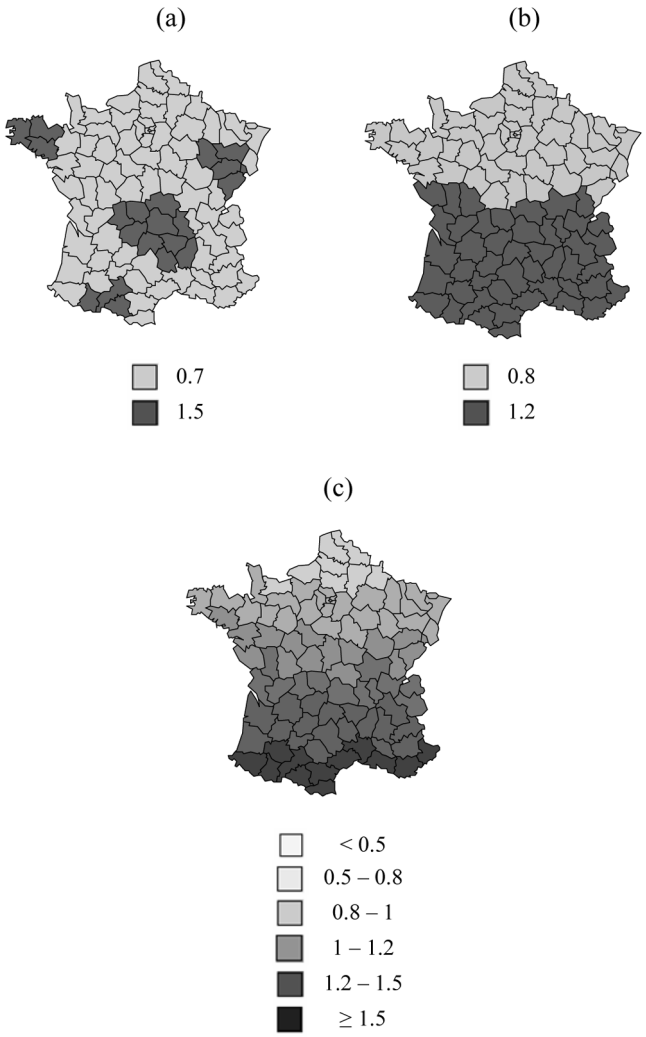

FIG 2. — Représentation cartographique des trois modèles de risques : (a) « Bloc 4 » (b) « Nord-Sud » (c) « Gradient ».

2.3. Modèles d'estimation

Les données consistent en n couples (E_i, X_i) sur la base desquelles les risques relatifs sont estimés. Le modèle d'estimation est un modèle hiérarchique bayésien de Poisson recommandé dans la littérature [3, 4, 9] où la variabilité résiduelle extra-poissonnienne est supposée avoir une structure spatiale. Cette approche est aujourd'hui souvent utilisée car reconnue pour améliorer la précision des estimateurs des risques relatifs. Remarquons que dans notre cas, ce modèle ne correspond pas au « vrai » modèle sous-jacent. Le modèle

d'estimation est le modèle hiérarchique suivant :

$$X_i|\theta_i \sim P(E_i\theta_i)$$
$$Log(\theta_i)|\beta, \varepsilon_i = \beta + \varepsilon_i$$

où le vecteur ε est modélisé par un processus gaussien suivant un modèle BYM proposé par [1]. Plus précisément, le modèle BYM est défini :

$$\varepsilon_i = u_i + v_i, \quad \text{pour } i = 1, ..., n$$

où v est un vecteur gaussien centré sans structure spatiale de variance σ_v^2 et u est un processus gaussien suivant un modèle CAR (Conditionnal AutoRegressive) intrinsèque de variance conditionnelle σ_u^2, u et v étant indépendants. La loi conditionnelle de u_i est définie par :

$$\lfloor u_i|u_j, i \neq j, \sigma_u^2 \rfloor \sim N(\overline{u}_i, \sigma_i^2)$$

avec $\overline{u}_i = \dfrac{1}{\sum\limits_j w_{ij}} \sum\limits_j u_j w_{ij}$ et $\sigma_i^2 = \dfrac{\sigma_u^2}{\sum\limits_j w_{ij}}$ où w_{ij} caractérise la proximité géographique entre les zones i et j (dans notre cas, w_{ij} vaut 1 si i et j partagent une frontière commune, 0 sinon). Les deux paramètres de variance (σ_v^2 et σ_u^2) permettent d'avoir un modèle décrivant des structures spatiales variées avec plus ou moins de force d'autocorrélation.

Nous avons choisi des lois *a priori* peu informatives au troisième niveau du modèle hiérarchique. Les lois *a priori* des variances sont des distributions Inverse Gamma de paramètres *0,5* et *0,0005* comme suggérés par [8]. Afin d'étudier la sensibilité aux choix des lois *a priori* sur les variances, nous avons comparé les résultats avec ceux obtenus quand les lois *a priori* des variances sont choisies beaucoup plus « plates » à savoir des lois Inverse Gamma de paramètres *0.0001* et *0.0001*. Aucune différence n'a été trouvée (résultats non montrés). Les différentes estimations ont été faites via un algorithme stochastique de la famille des algorithmes de Monte Carlo par Chaînes de Markov (MCMC) à l'aide du logiciel WinBUGS [13]. La convergence des algorithmes MCMC a été vérifiée par différents critères directement accessibles sous WinBUGS dont l'estimation de l'erreur de Monte-Carlo, cette quantité estimant, pour chaque paramètre, la différence entre la moyenne des valeurs échantillonnées (qui sont utilisées dans l'estimation des moyennes *a posteriori* des paramètres) et la vraie moyenne *a posteriori*. En pratique, si cette erreur est inférieure à 5 % de l'écart-type *a posteriori* estimé du paramètre, nous considérons que l'algorithme a convergé. Pour toutes les situations, les résultats sont basés sur 50 000 itérations de l'algorithme après avoir enlevé les 10 000 premières itérations pour la « période de chauffe ».

Afin de comparer les différentes fonctions de perte présentées précédemment, des critères spécifiques aux estimations des risques relatifs, les erreurs quadratiques et les biais relatifs (en pourcentage), ont été calculés *i.e.*

$EQ(\theta,\hat{\theta}) = \sum_{i=1}^{n}(\theta_i - \hat{\theta}_i)^2$ et $B(\theta,\hat{\theta}) = \frac{100}{n}\sum_{i=1}^{n}\frac{\hat{\theta}_i - \theta_i}{\theta_i}$ respectivement, où θ_i correspond au vrai risque relatif de la zone i et $\hat{\theta}_i$ son estimation. Ces critères ont été étudiés sur l'ensemble des zones géographiques mais aussi sur les zones géographiques présentant des valeurs de risques relatifs élevées *i.e.* pour le modèle « Bloc 4 », les quatre blocs de départements possédant des risques relatifs égaux à 1,5 (21 zones géographiques) ; pour le modèle « Nord-Sud », la partie Sud où les départements ont des risques relatifs fixés à 1,2 (46 zones géographiques) ; et pour le modèle « Gradient », les départements avec des risques relatifs supérieurs ou égaux à 1,25 (26 zones géographiques avec une moyenne des risques relatifs de 1,4). Un troisième critère correspondant à la sensibilité de l'estimateur, a été calculé uniquement sur les zones géographiques présentant des risques relatifs élevés. Si nous notons $\mathrm{D}_{\text{excès}}$ le domaine restreint aux zones présentant des risques relatifs importants comme défini juste ci-dessus, la capacité de l'estimateur à détecter des « vrais positifs » (notée VP) c'est-à-dire de réelles zones géographiques à risque relatif élevé est la proportion conditionnelle au domaine $\mathrm{D}_{\text{excès}}$ d'estimations supérieures à une valeur seuil θ_s. La définition de VP est donc :

$$VP(\theta,\hat{\theta}) = 100 * \frac{Card(A)}{Card(D_{excès})}$$

où $A = \{i; \hat{\theta}_i \geqslant \theta_s \text{ et } i \in D_{excès}\}$ et $Card(A)$ signifie le cardinal de l'ensemble A.

La valeur seuil θ_s a été fixée à *1,3, 1,1* et *1,2* pour respectivement le modèle « Bloc 4 », « Nord-Sud » et « Gradient » ceci afin de respecter les différences de valeurs entre les risques sur $\mathrm{D}_{\text{excès}}$. La valeur seuil choisie correspond à environ *85-90 %* de la valeur moyenne des vrais risques sur $\mathrm{D}_{\text{excès}}$. La capacité à détecter des « faux positifs » a également été étudiée. Sa définition est identique à celle de VP mais sur le domaine complémentaire à $\mathrm{D}_{\text{excès}}$ pour les modèles « Bloc 4 » et « Nord-Sud ». Pour le modèle « Gradient », les « faux positifs » sont calculés sur le domaine où les vrais risques relatifs sont inférieurs à 0,8.

3. Résultats

Dans cette section, nous étudions les critères de comparaison entre les différents estimateurs bayésiens des risques relatifs moyennés sur 100 réplications indépendantes et ceci, pour chaque modèle de risque. L'étude du tableau 1 (1ère colonne) nous montre que, de manière générale, les différents estimateurs donnent de bons résultats pour le modèle de risque « Bloc 4 ». L'estimateur $\overline{\theta}_{p-q}$ est cependant plus performant.

En effet, l'estimateur $\overline{\theta}_{p-q}$ minimise l'erreur quadratique moyennée sur les 100 réplications tout en possédant un biais moyen très faible (moins de *1 %* de biais). Pour des raisons de visibilité, la figure 3 représente pour chacune des 20 premières réplications, les valeurs de l'erreur quadratique, ce graphique étant représentatif de l'ensemble des 100 réplications. L'ensemble de ces valeurs

TABLEAU 1. — Comparaison entre les estimateurs des risques relatifs pour les trois modèles de risque moyennée sur 100 réplications.

	Bloc 4		Nord-Sud		Gradient	
	EQ[a] (s)[c]	B[b] (s)[d]	EQ[a] (s)[c]	B[b] (s)[d]	EQ[a] (s)[c]	B[b] (s)[d]
Moyenne *a posteriori*	2,61 (0,55)	1,42 (2,22)	0,68 (0,13)	0,41 (1,69)	0,80 (0,18)	0,21 (1,82)
Médiane *a posteriori*	2,69 (0,57)	0,43 (2,21)	0,69 (0,13)	0,02 (1,68)	0,81 (0,18)	−0,45 (1,81)
$\overline{\theta}_{p-q}$	2,33 (0,76)	-0,61 (2,41)	0,58 (0,20)	−1,20 (1,79)	0,89 (0,21)	−1,69 (1,92)

(a) Erreur quadratique moyenne sur les 100 réplications
(b) Biais relatif moyen (en pourcentage) sur les 100 réplications
(c) Ecart-type des 100 erreurs quadratiques
(d) Ecart-type des 100 biais relatifs

nous confirme, que l'estimateur $\overline{\theta}_{p-q}$, en trait plein sur la figure, fournit le plus souvent l'erreur quadratique la plus faible.

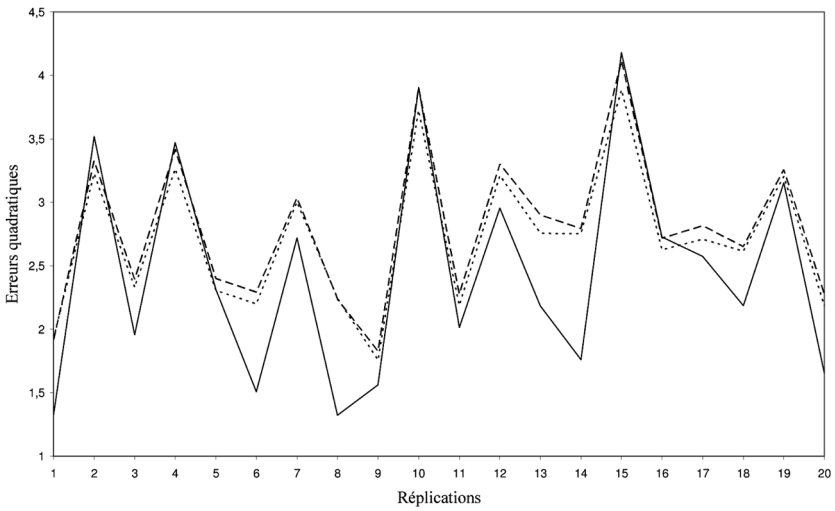

FIG 3. — Graphe des erreurs quadratiques pour le modèle « Bloc 4 » en fonction des réplications : la moyenne *a posteriori* (- - -) basée sur la fonction de perte quadratique, la médiane *a posteriori* (- - -) basée sur la fonction de perte en valeur absolue et l'estimateur pondéré $\overline{\theta}_{p-q}(-)$ basée sur la fonction de perte quadratique pondérée.

Une comparaison entre les moyennes des 100 valeurs des erreurs quadratiques obtenues par $\overline{\theta}_{p-q}$ et par chacun des deux autres estimateurs a été faite par un test de Wilcoxon. Les deux tests sont nettement significatifs en faveur d'erreurs quadratiques plus faibles pour $\overline{\theta}_{p-q}$, avec un degré de signification de $2\ 10^{-3}$ pour la comparaison entre $\overline{\theta}_{p-q}$ et la moyenne *a posteriori* et de $3\ 10^{-4}$ entre $\overline{\theta}_{p-q}$ et la médiane *a posteriori*. Les moyennes des 100 valeurs des biais relatifs pour chacun des trois estimateurs sont faibles, légèrement plus élevées pour la moyenne *a posteriori*.

Pour les résultats du modèle « Nord-Sud » (tableau 1, $2^{\text{ème}}$ colonne), nous remarquons qu'en moyenne sur les 100 réplications, l'estimateur $\overline{\theta}_{p-q}$ minimise à nouveau l'erreur quadratique moyenne. Cet estimateur sous-estime le vrai risque avec moins de *2* % de biais même si le biais est légèrement plus élevé en moyenne que pour les deux autres estimateurs. Pour ce cas également, l'erreur quadratique a été souvent la plus faible lors des 100 réplications. Les tests de comparaison de moyennes des erreurs quadratiques sont également significatifs avec un degré de signification de $2\ 10^{-5}$ pour la comparaison entre $\overline{\theta}_{p-q}$ et la moyenne *a posteriori* et de $2\ 10^{-5}$ entre $\overline{\theta}_{p-q}$ et la médiane *a posteriori*.

Pour le modèle de risque « Gradient », la moyenne et la médiane *a posteriori* permettent de réduire l'erreur quadratique (tableau 1, $3^{\text{ème}}$ colonne). Ces estimateurs présentent des biais moyennés faibles, moins de *1* %. L'estimateur $\overline{\theta}_{p-q}$ présente des performances légèrement moins bonnes mais donne encore des résultats sur ces critères convenables surtout concernant l'erreur quadratique bien que ce modèle de risque soit le moins favorable à cet estimateur.

Le tableau 2 fournit pour chacun des trois modèles de risque les résultats aux critères spécifiques uniquement sur les zones géographiques où les risques relatifs sont en excès ainsi que la proportion de « vrais positifs » VP c'est-à-dire de risques relatifs estimés supérieurs au seuil θ_s sur le domaine où les risques relatifs sont en excès.

En considérant les résultats sur les zones en excès du modèle de risque « Bloc 4 » (tableau 2, $1^{\text{ère}}$ colonne), les trois estimateurs présentent des résultats proches avec une erreur quadratique inférieure à 2 en moyenne. Les biais moyens sont assez importants, de l'ordre de 15 %. La proportion de vrais positifs est la plus importante pour l'estimateur pondéré, et ceci de manière relativement fréquente (64 fois sur les 100 réplications). Dans le cas « Nord-Sud » (tableau 2, $2^{\text{ème}}$ colonne), les trois estimateurs présentent à nouveau des résultats proches avec globalement de meilleures performances que pour le modèle de risque « Bloc 4 » (erreurs quadratiques et biais plus faibles). La proportion de vrais positifs est équivalente pour les trois estimateurs. Dans le cas du modèle « Gradient », nous observons que les résultats de la moyenne *a posteriori* et de la médiane *a posteriori* sont très proches sur les 100 réplications alors que $\overline{\theta}_{p-q}$ présente un biais et une erreur quadratique légèrement plus élevés. Les proportions de vrais positifs sont toutes proches et supérieures à 95 %. Les proportions de faux positifs ont été également calculées. Pour l'ensemble des configurations (les trois modèles de risque et les trois estimateurs), ces proportions ont toujours été nulles sur les 100 réplications. Enfin, les critères de comparaison (biais, erreurs quadratiques

et VP) sont légèrement plus fluctuants concernant l'estimateur pondéré, les écarts types moyennés sur les 100 réplications de ces critères sont toujours plus importants (tableaux 1 et 2) mais restent très proches.

TABLEAU 2. — Comparaison entre les estimateurs des risques relatifs sur les zones en excès pour les trois modèles de risque.

	Bloc 4			Nord-Sud			Gradient		
	EQ[a] (s)[d]	B[b] (s)[e]	VP[c] (s)[f]	EQ[a] (s)[d]	B[b] (s)[e]	VP[c] (s)[f]	EQ[a] (s)[d]	B[b] (s)[e]	VP[c] (s)[f]
Moyenne	1,77	13,89	47,1	0,45	3,72	74,2	0,43	0,76	96,3
a posteriori	(0,58)	(4,49)	(15,1)	(0,14)	(2,69)	(12,3)	(0,16)	(2,91)	(4,4)
Médiane	1,91	14,87	43,9	0,46	4,18	72,1	0,43	1,42	95,8
a posteriori	(0,61)	(4,43)	(14,6)	(0,15)	(2,66)	(12,6)	(0,16)	(2,89)	(4,8)
$\overline{\theta}_{p-q}$	1,81	15,19	53,1	0,41	5,48	73,5	0,46	5,63	96,3
	(0,79)	(5,40)	(19,6)	(0,19)	(2,91)	(17,3)	(0,21)	(3,05)	(5,1)

(a) Erreur quadratique moyenne sur les 100 réplications
(b) Biais relatif moyen (en pourcentage) sur les 100 réplications
(c) Proportion moyenne de vrais positifs (en pourcentage) sur les 100 réplications
(d) Ecart-type des 100 erreurs quadratiques
(e) Ecart-type des 100 biais relatifs
(f) Ecart-type des 100 proportions de vrais positifs

4. Discussion

Le choix d'une fonction de perte et l'étude de l'estimateur qui en découle constituent une démarche rarement utilisée et pourtant prometteuse dans les approches bayésiennes. Ce choix peut être guidé par l'objectif recherché comme ici, la mise en valeur des risques relatifs extrêmes. Les critères d'évaluation de qualité des estimateurs sont dans cette étude l'erreur quadratique, le biais relatif et la proportion de « vrais positifs ». Conditionnellement à ces trois critères, cette étude a montré des légères différences entre les estimateurs, prouvant dans ce sens leur robustesse. Néanmoins, pour les deux surfaces discontinues des risques relatifs (« Bloc 4 » et « Nord-Sud »), l'estimateur pondéré donne de plus faibles erreurs quadratiques que les estimateurs plus classiques, la moyenne et la médiane *a posteriori*, et ceci de manière significative sur l'ensemble du domaine géographique. Les biais sont soit plus faibles pour $\overline{\theta}_{p-q}$, soit équivalents aux autres estimateurs. Notons que les proportions de vrais positifs (critère VP) sont soit favorables à l'estimateur pondéré soit équivalentes entre les trois estimateurs. Ces différences pourraient peut-être être accentuées dans le cas d'autres configurations de risques relatifs et/ou de nombre de cas attendus avec notamment une plus grande variabilité, ceci ouvrant une perspective de travail. Pour le cas de la surface de risque de type

« Gradient », l'estimateur pondéré a, comme attendu, donné de moins bons résultats (biais et erreur quadratique) que les deux autres estimateurs mais garde des performances convenables. Notamment, les proportions de vrais positifs sont équivalentes entre les trois estimateurs. Le choix d'une autre fonction de perte exponentielle plus « brutale » c'est à dire basée sur des poids de la forme $c_j = \exp(j/2)$ (résultats non montrés) a donné lieu à des performances légèrement moins bonnes en terme d'erreur quadratique et de biais. En effet, ce choix correspond à des poids quasi nuls sur la totalité des risques relatifs excepté les dix risques relatifs les plus importants. Le coté arbitraire des poids reste à discuter et une orientation possible serait de guider ce choix en fonction des qualités de prédiction du modèle en s'inspirant des travaux basés sur le calcul des prédictions dans les modèles hiérarchiques bayésiens [11, 14].

Remerciements

Nous remercions les rapporteurs pour les intéressantes remarques et suggestions. Ce travail a bénéficié du support financier de l'AFSSE (RD2004004) et de l'INSERM-ATC (A03150LS).

Références

[1] Besag J., York J. & Mollie A. (1991). Bayesian image restoration, with two applications in spatial statistics. Ann. I. Stat. Math. 1991 ; 43 : 1-59.

[2] Clavel J., Goubin A., Auclerc M.F., Auvrignon A., Waterkeyn C., Patte C. *et al.* (2004). Incidence of childhood leukemia and non-Hodgkin's lymphoma in France – national registry of childhood leukemia and lymphoma, 1990-1999, Eur J Cancer Prev ; 13 : 97-103.

[3] Elliott P., Cuzick J., English D. and Stern R. (1992). Geographical and Environmental Epidemiology : Methods for Small-Area Studies, Oxford University Press, Oxford.

[4] Elliott P., Wakefield J., Best N.G. & Briggs D. (2000). Spatial Epidemiology, methods and applications. Oxford : Oxford University Press.

[5] Evrard A.S., Hemon D., Billon S., Laurier D., Jougla E., Tirmarche M. & Clavel J. (2005). Ecological association between indoor Radon concentration and childhood leukemia incidence in France, 1990-1998. Eur. J. Cancer Prev. ; 14 : 147-57.

[6] Ghosh M. (1992). Constrained Bayes estimation with applications. J. Am. Stat. Assoc. ; 87 : 533-540.

[7] Green P.J., Richardson S. (2002). Hidden Markov models and disease mapping. J. Am. Stat. Assoc. ; 97 : 1055-1070.

[8] Kelsall J.E. & Wakefield J. (1999). Discussion of « Bayesian models for spatially correlated disease and exposure data » by Best *et al.* (1999). In Bayesian Statistics 6, eds J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith. Oxford : Oxford University Press.

[9] Lawson A., Bigerri A., Bohning D., Lesaffre E., Viel J. F., Bertollini R. (1999). Disease Mapping and risk assessment for public health. John Wiley & Sons, UK.

[10] Louis T. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. J. Am. Stat. Assoc ; 79 : 393-398.

[11] Marshall E.C. and Spiegelhalter D.J. (2003). Approximate cross-validatory predictive checks in disease mapping models. Stat. Med. ; 22 : 1649-1660.

[12] Mollié A. (1996). Bayesian mapping of disease. In : Markov Chain Monte Carlo in Practice, Gilks W.R., Richardson S., Spiegelhalter D.J. Eds., Chapman & Hall ; Chapter 20, 359-380.

[13] Spiegelhalter D.J., Thomas A. & Best N.G. (2001). WinBUGS Version 1.4, User Manual.. Medical Research Council Biostatistics Unit, Cambridge, and Imperial College School of Medicine, London, UK.

[14] Stern H. S. et N. Cressie N. (2000). Posterior predictive model checks for disease mapping models. Stat. Med. ; 19 : 2377-2397.

[15] Wright D. L., Stern H. S., and Cressie N. (2003). Loss functions for estimation of extrema with an application to disease mapping. Can. J. Stat. ; 31 : 251-266.