Utiliser un modèle bayésien à effets aléatoires partagés pour analyser les variations spatiales du risque de maladies à partir de sources d'information multiples.

Sophie Ancelet, Juanjo Abellan, Sylvia Richardson, Victor del Rio Vilas, Colin Birch

Utiliser un modèle bayésien à effets aléatoires partagés pour analyser les variations spatiales du risque de maladies à partir de sources d'information multiples.

Sophie Ancelet (a)(b) & Juanjo Abellan (b)(c) & Sylvia Richardson (b) & Victor Del Rio Vilas (d) & Colin Birch (d)

(a) INSERM-INED U822, Equipe Epidémiologie de la reproduction et du développement de l'enfant, Hôpital du Kremlin Bicêtre, France
(b) Imperial College, Department of Epidemiology and Public Health, St Mary's Campus, Londres, UK
(c) Biomedical Research Centre Network for Epidemiology and Public Health, Espagne
(d) Veterinary Laboratories Agency, UK

Résumé

Ces dernières années, la modélisation hiérarchique bayésienne a connu un essor considérable en épidémiologie géographique. Les principaux modèles développés sont axés autour de la description spatiale et spatio-temporelle des variations du risque d'une ou plusieurs maladies à partir d'une source de recensement unique des cas. Parallèlement, l'analyse de sources de données multiples se développe dans les études épidémiologiques en vue notamment d'améliorer la fiabilité des diagnostics. Dans ce contexte, nous présentons un modèle bayésien à effets aléatoires partagés pour l'analyse jointe des variations spatiales du risque d'une ou plusieurs maladies à partir d'indicateurs de santé multiples. Nous considérons le cas particulier où ces indicateurs sont mesurés à la même échelle géographique et fournissent des mesures du même risque d'intérêt sous-jacent à partir de sources différentes. Le modèle hiérarchique bayésien proposé partage le risque de maladie, non expliqué par des facteurs d'expositions individuels, en un effet spatialement structuré commun aux différentes sources d'information et un effet non-structuré spécifique qui contrôle les éventuelles dissimilarités entre sources. Nous avons appliqué ce modèle à l'analyse jointe des variations spatiales du risque de tremblante du mouton au Pays de Galles à partir de deux sources de surveillance des cas d'infection. Nous avons également estimé l'effet de plusieurs facteurs d'exposition susceptibles d'expliquer ces variations. Enfin, nous avons utilisé ce cas d'étude pour illustrer les avantages d'une analyse jointe par rapport à une analyse séparée de sources de données multiples.

Key-Words : champs de Markov cachés, épidémiologie géographique, inférence bayésienne, modèles à effets aléatoires partagés, sources de données multiples, statistique spatiale

Recently, bayesian hierarchical models have flourished in geographical epidemiology. The main developed models have focused on describing the spatial and spatio-temporal variations of one or several diseases risk from an unique surveillance source. At the same time, the analysis of data from multiple sources has become increasingly common in epidemiologic studies to improve the validity of diagnoses. In this context, we present a bayesian shared component models to the joint analysis of spatial variations of one of several diseases risk from multiple health outcomes. We focus on the particular situation where these health outcomes are observed at the same geographical scale and provide measures of the same underlying risk surface from different sources. Our bayesian hierarchical model splits the risk of disease not explained by individual exposure factors into a shared spatial component and a specific unstructured component that controls the possible differential between the multiple sources. We applied this model to jointly analyse the spatial variations of the risk of scrapie infection in Wales from two surveillance sources. We also assessed the effect of several exposure factors which could explain this variability. Finally, we used this case study to illustrate the benefits of a joint analysis compared to a separate analyse of multiple sources.

Key-Words : bayesian inference, geographical epidemiology, hidden Markov Random Fields, multiple sources, shared component models, spatial statistics

1 Introduction

Ces dernières années, la modélisation hiérarchique bayésienne a connu un essor considérable en épidémiologie géographique. Elle est reconnue pour fournir des outils flexibles et efficaces face aux difficultés techniques induites par certaines caractéristiques des indicateurs de santé : sur-dispersion dûe à un faible nombre de cas de maladie, forte corrélation spatiale etc. Les travaux les plus récents se sont principalement concentrés autour de la modélisation spatiale ou spatio-temporelle des variations du risque d'une ou plusieurs maladies à partir d'une source de recensement unique des cas (Best (2005), Richardson (2006)). Parallèlement, la modélisation de sources de données multiples se développe dans les études épidémiologiques (Horton (2004)) et notamment, dans le contexte Bayésien (Jackson (2008)). En effet, une seule base d'information peut s'avérer insuffisante pour obtenir des diagnostics fiables au vu des nombreuses sources d'incertitude induites par les méthodes de recensement des cas (données manquantes, biais de sélection et/ou d'information, erreurs de mesure, facteurs de confusion non-mesurés,...). Ceci est notamment le cas lorsque la maladie d'intérêt est rare et/ou difficilement diagnosticable et/ou que le recueil des données conduit à un repérage préférentiel des cas ou des témoins. Dans ce contexte, nous nous intéressons au problème de la modélisation jointe des variations spatiales du risque de maladies à partir d'indicateurs de santé multiples. Nous considérons le cas particulier où ces indicateurs sont mesurés à la même échelle géographique et fournissent des mesures du même risque d'intérêt sous-jacent à partir de sources différentes.

La modélisation et l'inférence statistique de sources de données multiples est un problème méthodologique ouvert qui nécessite le développement de structures aléatoires complexes. Il s'agit de lier plusieurs sources d'information dans une structure globale cohérente et qui tienne compte explicitement de l'existence possible de dépendances entre les différentes variables réponse. Il s'agit également d'apporter des solutions aux difficultés de modélisation induites par le traitement de sources d'information multiples issues de protocoles observationnels différents : occurrence de données manquantes, biais de nature différente ... Un des objectifs commun à l'utilisation de telles structures est de tirer profit du gain de puissance obtenu en combinant plusieurs sources de données.

2 Une approche possible : les modèles à effets aléatoires partagés

Nous présentons un modèle à effets aléatoires partagés pour l'analyse jointe des variations spatiales du risque de maladies à partir d'indicateurs de santé multiples. Cette classe de modèles hiérarchiques permet d'introduire des liens de corrélations indirects entre les multiples indicateurs observés, avec un ou plusieurs effet(s) aléatoire(s) commun(s). En pratique, ces modèles sont utilisés dans les études de cohorte soit pour modéliser l'association entre des mesures longitudinales et le temps d'occurrence d'un événement (Vonesh (2006), Williamson (2008)) soit pour modéliser une variable réponse longitudinale en présence de données manquantes à caractère informatif (Hogan (1997), Albert (2002)). Leur utilisation se développe également en épidémiologie géographique pour analyser les variations spatiales ou spatio-temporelles du risque associé à plusieurs maladies partageant des facteurs de risque communs (Knorr-Held (2001), Knorr-Held (2005), Richardson (2006)). L'intérêt de cet exposé est donc d'élargir la gamme d'utilisation possible de cette classe de modèles à la modélisation de sources d'information multiples.

Nous nous limitons au cas où deux sources de données sont disponibles mais le modèle pourrait être élargi pour combiner plus de deux sources. Nous considérons des indicateurs de santé binaires (e.g., sain/malade) géo-référencés. Suivant les idées de Knorr-Held et al.(2005), notre modèle partage le risque de maladie, non expliqué par des facteurs d'expositions individuels, en un effet spatialement structuré commun aux deux sources d'information et un effet résiduel spécifique à l'une d'entre elles. L'effet aléatoire spatial crée un lien de dépendance indirect entre les sources d'information considérées et joue le rôle de substitut pour tous les facteurs d'exposition spatialement structurés mais non-mesurés qui peuvent expliquer la "vraie" répartition spatiale du risque de maladies. L'effet spécifique joue le rôle de substitut pour tous les facteurs résiduels non-mesurés susceptibles d'expliquer des dissimilarités géographiques dans la répartition des risques entre les deux sources d'observation. L'inférence du modèle proposé a été effectuée sous le paradigme bayésien avec des algorithmes Monte-Carlo par Chaînes de Markov (MCMC) implémentés sous le logiciel OpenBUGS (Spiegelhalter (2007)) couplé au package R2WinBUGS de R.

Par rapport à une approche qui spécifierait directement la loi jointe des réponses multivariées, le modèle proposé a plusieurs avantages. Il permet d'estimer et de cartographier séparément la surface de risque commune aux différentes sources d'information et l'hétérogénéité résiduelle du risque spécifique à l'une ou l'autre d'entre elles. Il fournit également un estimateur du pouvoir explicatif de l'effet aléatoire partagé considéré sur les variations de risques associées à chaque source d'information.

3 Présentation du cas d'étude

Nous avons appliqué ce modèle à l'analyse des variations spatiales du risque de tremblante du mouton dans 1034 comtés administratifs du Pays de Galles sur la période 2002-2006. La tremblante du mouton est une maladie neurodégénérative fatale qui affecte les moutons et autres petits ruminants. Elle sévit de manière permanente au Royaume-Uni. Cette maladie est d'autant plus préoccupante que l'hypothèse d'un lien indirect entre la tremblante du mouton et la maladie de Creutzfeld-Jacob chez l'homme (dont les symptômes sont assez voisins) est scientifiquement posée.

Les objectifs principaux de cette étude sont (1) de pouvoir générer des hypothèses sur l'étiologie de la tremblante du mouton atypique, une forme de tremblante qui sévit depuis 2002 au Royaume-Uni et dont les signes cliniques et facteurs de risque sont peu connus des vétérinaires (2) de valider l'hypothèse selon laquelle cette forme de tremblante ne serait pas infectieuse (Benestad (2008) en comparant, à partir de deux sources de recensement des cas, les variations spatiales du risque de tremblante atypique à celles d'une autre forme de tremblante, dite classique, plus connue des vétérinaires.

Pour chaque forme de tremblante, nous considérons deux sources de recensement des cas. La première (nommée AS pour "Abattoir Survey") est basée sur un examen vétérinaire de moutons échantillonnés aléatoirement lors de leur arrivée à l'abattoir. La seconde (nommée SFD pour "Scrapie Notification Database") est basée sur une liste supposée "exhaustive" des cas de tremblante identifiés par examens vétérinaires après signalement des fermiers. Très peu de cas de tremblante sont reportés dans chaque source d'observation.

Pour chacune des deux sources de données disponibles (SFD et AS), un indicateur binaire indique, pour chaque ferme du dénominateur de base choisi, si au moins un mouton a été déclaré infecté sur la période 2002-2006. Les données relatives aux fermes dont aucun mouton n'a été échantillonné à l'abattoir sur cette période sont considérées manquantes.

4 Application

Nous avons mené l'inférence du modèle à effets aléatoires partagés proposé à partir des deux sources de données disponibles (AS et SFD) pour chaque forme de tremblante (atypique et classique). Nous sommes parvenus à identifier une forme de structure spatiale

commune au risque de tremblante classique entre SFD et AS, essentiellement interprétable comme la "vraie" répartition de ce risque au Pays de Galles. Ce signal spatial reste très faible pour la tremblante atypique par rapport à la tremblante classique confortant l'hypothèse selon laquelle la forme atypique ne serait pas infectieuse. Pour les deux formes de tremblante, le modèle indique que les sources d'information se ressemblent significativement indiquant une sensibilité identique dans la détection des fermes infectées. Toutefois, certains facteurs d'exposition semblent affecter différemment les variations du risque de tremblante entre SFD et AS. L'étude a également montré que, parmi les différents facteurs de risque envisagés, seule la taille du troupeau semble avoir un impact significatif commun aux risques de tremblante classique et atypique. Enfin, les résultats obtenus montrent que la répartition géographique du risque de tremblante au Pays de Galles sur la période 2002-2006 est différente entre les formes classique et atypique.

Nous avons appliqué une version univariée du modèle à effets aléatoires partagés proposé à chaque source de données SFD et AS séparément. Ceci nous a permis de mettre en evidence les avantages d'une analyse jointe des sources AS et SFD par rapport à une analyse séparée. L'analyse jointe a permis d'améliorer la convergence des algorithmes MCMC. De l'amélioration de puissance obtenue en combinant plusieurs sources d'information a découlé une plus grande robustesse au choix des lois *a priori* et une nette amélioration des précisions d'estimations des risques relatifs estimés. Enfin, les cartes indiquant les variations spatiales du risque de tremblante au Pays de Galles sont globalement mieux définies avec une analyse jointe.

Bibliographie

[1] Best, NG., Richardson, S., Thomson, A. (2005) A comparison of Bayesian spatial models for disease mapping *Statistical Methods in Medical Research*, 14, 35-59

[2] Richardson, S., Abellan, J.J., Best, NG. (2006) Bayesian spatio-temporal analysis of joint patterns of male and female lung cancer risks in Yorkshire (UK) *Statistical Methods in Medical Research*, 15, 385-407

[3] Horton, N.J. et Fitzmaurice, G.M. (2004) Regression analysis of multiple source and multiple informant data from complex survey samples *Statistics in medicine*, 23, 2911-2933

[4] Jackson, C.H., Best, N.G., Richardson, S. (2008) Bayesian graphical models for regression on multiple data sets with different variables *Biostatistics*, 9(4), 1-17

[5] Vonesh, E.F., et al. (2006) Shared parameter models for the joint analysis of longitudinal data and event times. *Statistics in medicine* 25 :143-163

[6] Williamson, P.R., et al. (2008) Joint modelling of longitudinal and competing risks data. *Statistics in medicine* 16 :259-272

[7] Hogan, J.W. et Laird, N.M. (1997) Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in medicine* 16 :259-272

[8] Albert, P.S. et al. (2002) A latent autoregressive model for longitudinal binary data subject to informative missingness *Biometrics* 58 : 631-642

[9] Knorr-Held, L. et Best, NG. (2001) A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society Series A. (Statistics in Society)* 164(1), 73-85
[10] Knorr-Held, L. et al. (2005) Towards joint disease mapping. *Statistical Methods in Medical Research*, 14, 61-82
[11] Spiegelhalter, A. et al. (2007) OpenBUGS User Manual Version 3.0.2. *MRC Biostatistics Unit, Cambridge*
[12] Benestad, S.L. et al. (2008) Atypical/Nor98 scrapie : properties of the agent, genetics, and epidemiology *The Veterinary Record* 39(4) :19-33