Classification automatique par champs de Markov cachés pour la cartographie du risque en épidémiologie

David Abrial, Myriam Charras-Garrido

Classification automatique par champs de Markov cachés pour la cartographie du risque en épidémiologie

David Abrial & Myriam Charras-Garrido

INRA
Unité d'Epidémiologie Animale
Département de Santé Animale
Centre de recherche de Clermont-Ferrand-Theix
63122 Saint-Genès-Champanelle

RÉSUMÉ

La cartographie du risque en épidémiologie permet de mettre en évidence la situation géographique de zones à faible ou fort risque de contamination, ainsi qu'une mesure des "différences de risque" entre ces régions. Actuellement, les modèles de cartographie pour données groupées utilisés par les épidémiologistes sont basés sur des approches de type Bayésien hiérarchique se focalisant sur l'estimation du risque pour chaque unité géographique et appliquant un lissage spatial de type auto-régressif Gaussien. La classification des risques, nécessaire pour le tracé des cartes, est effectuée dans un deuxième temps. Par analogie avec les méthodes utilisées en analyse d'images, nous proposons une nouvelle méthode de cartographie du risque basée sur l'approche par champs aléatoires de Markov cachés. Le champ caché correspond à une classification du risque qui fait ainsi directement partie des paramètres à estimer. L'originalité de la cartographie du risque consiste notamment à modéliser le champ observé par une loi de Poisson, et non plus par une loi normale comme c'est généralement le cas en analyse d'images. De plus, dans ce contexte, l'ordre des classes, en lien avec leur situation géographique et leur interprétation en terme de gravité du risque, est très important. Les classiques fonctions de lien spatial, par exemple de type Potts, ne prennent pas en compte cette interprétation. Nous proposons donc de nouvelles fonctions de lien spatial permettant de tenir compte de l'ordre des classes. Afin d'estimer les paramètres du modèle et déterminer les classes de risque, nous utilisons l'algorithme EM, en particulier sa variante MCEM.

ABSTRACT

Risk mapping in epidemiology will reveal the location of areas with low or high risk of contamination, as well as a measure of risk differences between these regions. Current risk mapping models for pooled data used by epidemiologists are based on hierarchical Bayesian approaches focusing on the estimated risk for each geographical unit and applying a Gaussian auto-regressive spatial smoothing. The risk classification, necessary to

draw maps, is performed in a second time. By analogy with the methods used in image analysis, we propose a new method of risk mapping based on hidden Markov random fields approach. The hidden field corresponds to a classification of the risk, that is directly part of the parameters to be estimated. The originality of risk mapping is in particular to model the observed field by a Poisson distribution, rather than a normal distribution as it is usually the case in image analysis. Moreover, in this context, the order of the classes, together with their location and their interpretation in terms of risk level, is very important. The usual functions of spatial link, like Potts type, do not take into account this interpretation. Therefore, we propose new functions of spatial link that take into account the order of the classes. To estimate the parameters of the model and determine the risk classes, we use the EM algorithm, in particular its Monte Carlo variant MCEM.

MOTS CLÉS

distribution de Gibbs, algorithme EM, méthodes MCMC, lissage spatial, loi de Poisson, analyse d'images, carte de risque.

La cartographie du risque en épidémiologie permet de mettre en évidence l'hétérogénéïté spatiale, les zones géographiques les plus à risque et l'influence de certains facteurs, ce qui aide à mieux comprendre les mécanismes des maladies.

Nous nous plaçons dans un cadre où les données sont agrégées par unités géographiques. Les observations concernent des effectifs de cas (nombre de malades, de morts, etc.) et les effectifs de populations cibles (données démographiques), disponibles par unités géographiques (par exemple données groupées par cantons). A partir de ces informations, il s'agit alors d'estimer le risque de contamination pour chaque unité géographique et surtout de comparer ce risque entre les unités.

Afin d'atténuer les contrastes entre unités souvent observés avec une estimation "brute" du risque (nombre de cas/population dans chaque unité), notamment dans le cas d'une maladie rare, l'estimation du risque doit inclure un lissage spatial. Dans le but de comparer les différents risques, une symbologie de couleurs est nécessaire pour tracer des cartes colorées afin de visualiser les différentes zones de risque. On s'intéresse en particulier à la situation géographique des zones à faible ou fort risque, mais aussi aux "différences de risque" entre ces régions : par exemple, que le risque soit multiplié (respectivement divisé) par deux ou dix d'une région à l'autre est une information capitale pour les autorités sanitaires qui doivent en fonction de cela décider (ou non) la prise de mesures locales ou nationales. De plus, la situation claire des zones à risque permet aux épidémiologistes de se faire une idée des facteurs les plus influents (qui auront la même répartition géographique). Ces facteurs peuvent ensuite être intégrés dans la modélisation. Nous ne présenterons ici que le cas où seules les différences de populations entre unités géographiques sont prises en compte.

Actuellement les modèles de cartographie utilisés par les épidémiologistes sont basés sur des approches de type Bayésien hiérarchique proposées par Besag *et al.* (1991) et Mollié (1999). Ces méthodes se focalisent sur l'estimation du risque pour chaque unité géographique en appliquant un **lissage spatial de type auto-régressif Gaussien**. La classification des risques, nécessaire pour le tracé des cartes, est effectuée *a posteriori* par l'utilisateur. Les épidémiologistes procèdent notamment de manière empirique, mais la mise en classe peut également être effectuée par une méthode statistique de classification.

Par analogie avec les méthodes utilisées en analyse d'images (voir Guyon (1993), Winkler (1995), Chalmond (2000)), nous proposons une nouvelle méthode de cartographie du risque basée sur un autre type de modèles graphiques probabilistes : l'approche par **champs aléatoires de Markov cachés**, voir Chandler (1987). Le champ caché correspond à une classification du risque qui fait partie des paramètres à estimer. Ainsi, avec cette méthode, la classification fait partie intégrante de la procédure d'estimation et n'est plus effectuée dans un deuxième temps comme c'est le cas avec les modèles actuels. La construction des modèles de type champ aléatoire se fait "naturellement " en suivant les connaissances et les hypothèses épidémiologiques. De ce fait, tous les paramètres de ces modèles sont biologiquement interprétables.

Cette nouvelle application des champs de Markov cachés à la cartographie du risque a nécessité l'adaptation de la méthode. Tout d'abord, en cartographie du risque l'organisation spatiale des unités géographiques est généralement différente de l'analyse d'images classique. Au lieu de points (les pixels) répartis sur une grille carrée régulière, en général les données sont naturellement groupées en unités admistratives. Le nombre de voisins est alors très variable et la répartition des unités n'est absolument pas régulière. Cela rend plus complèxe la définition du voisinage, et les calculs sont plus longs lorsque le nombre de voisins est important (par exemple jusqu'à 13 voisins pour les cantons en France). On peut également choisir de trasformer les données en les groupant sur une grille régulière choisie. Nous avons par exemple utilisé un maillage de la France en 1264 hexagones, les unités ayant alors généralement 6 voisins.

Dans le contexte d'un modèle par champs de Markov caché, sachant les variables cachées (les classes) $x = (x_i)_i$, la loi de la variable aléatoire observée Y_i (ici le nombre de cas) ne dépend pas des observations y_{-i} faites dans les autres unités géographiques. De plus, voir Besag *et al.* (1991) et Guyon (1993), on suppose que la loi de Y_i dépend uniquement de la classe x_i de cette même unité i :

$$P(Y_i = y_i | y_{-i}, x, \lambda) = P(Y_i = y_i | x_i, \lambda) = P(Y_i = y_i | \lambda_{x_i}). \quad (1)$$

L'originalité principale de la cartographie du risque consiste à modéliser le champ observé par une **loi de Poisson**, et non plus par une loi normale comme c'est généralement le cas en analyse d'images. Plus précisément, pour chaque unité géographique i, le nombre

de cas observés y_i suit une loi de Poisson dont la moyenne s'exprime comme le risque λ_{x_i} correspondant à la classe x_i de l'unité multiplié par l'effectif de population n_i de l'unité : $y_i \sim \mathcal{P}(n_i\lambda_{x_i})$. La loi du nombre de cas observés s'écrit donc :

$$P(Y_i = y_i|\lambda_{x_i}) = \exp(-n_i\lambda_{x_i})\frac{(n_i\lambda_{x_i})^{y_i}}{y_i!} \tag{2}$$

L'introduction de cette loi discrète qui ne prend qu'assez peu de valeurs différentes (surtout lorsque l'on a une maladie rare et/ou une faible population, et donc une moyenne faible pour la loi de Poisson) rend l'estimation plus délicate et les erreurs de confusions de classes plus fréquentes.

Introduisons à présent la modélisation du champ de Markov caché discret X des classes de risque. D'après Besag (1974), la loi d'un champ de Markov peut s'exprimer comme une loi de Gibbs :

$$P(X = x) = \gamma \exp[-H(x|\alpha, \beta)], \tag{3}$$

avec

$$H(x|\alpha, \beta) = \sum_{i=1}^{n} \varphi_1(x_i|\alpha) - \beta \sum_{<i,j>} \varphi_2(x_i, x_j) + \dots \tag{4}$$

H, appelé l'**énergie interne** du champ x, est exprimée comme une somme de **fonctions potentielles** φ définies sur des cliques. γ est la constante de normalisation :

$$\gamma^{-1} = \sum_{x} \exp[-H(x)].$$

Le voisinage que nous considérons est d'ordre 1 (V_i est constitué des unités adjacentes à l'unité i), et nous nous limitons uniquement aux fonctions potentielles d'ordre 1 et 2.

La fonction potentielle d'ordre 1, notée φ_1, permet de contrôler les **proportions des différentes classes** de risques sur la carte. Nous avons choisi de l'exprimer comme $\varphi_1(x_i|\alpha) = \alpha_{x_i}$. Elle peut être également négligée en prenant $\varphi_1(x_i|\alpha) = 0$.

L'information *a priori* modélisée par φ_1 est contrebalancée par la fonction potentielle d'ordre 2, φ_2, qui prend en compte les corrélations spatiales deux-à-deux. Il s'agit d'une **fonction de lien spatial**. Le paramètre β agit comme une balance entre ces deux termes. Il est lié à l'**importance du lissage** et donc à l'homogénéïté des zones de risque.

Contrairement aux applications classiques en analyse d'images, dans le contexte de la cartographie du risque l'ordre des classes, en lien avec leur situation géographique et leur interprétation en terme de gravité du risque, est très important. Les classiques fonctions de lien spatial, notamment celles de type Potts (qui ne tiennent compte que de l'égalité ou la différence de variables cachées voisines), ne prennent pas en compte cette interprétation et peuvent aboutir à des cartes incohérentes. Par exemple, les classes à plus fort et plus faible risque (les deux classes extrêmes) peuvent se retrouver côte à côte, alors

que l'on s'attend plutôt à une gradation des risques et à des transitions progressives entre les classes. Nous proposons donc de nouvelles fonctions de lien spatial qui permettent de prendre en compte l'ordre des classes et d'éviter que deux classes trop différentes se retrouvent côte à côte.

Afin d'estimer les paramètres du modèle (par la méthode du maximum de vraisemblance) et déterminer les classes de risque, nous utilisons l'algorithme EM (Expectation-Maximisation). Proposé par Dempster *et al.* (1977) l'algorithme EM est un algorithme itératif pour le calcul de l'estimateur du maximum de vraisemblance des paramètres d'un modèle dans le cas de données manquantes (ou cachées, ici les classes). Plus précisément, d'une part, la vraisemblance des données cachées x et des données complètes (x, y) est approchée par la pseudo-vraisemblance proposée par Besag (1975) et (1986). Et d'autre part, l'espérance conditionnelle de la vraisemblance des données complètes (x, y) par rapport aux données observées est approchée par la moyenne empirique de simulations, voir Wei et Tanner (1990), c'est-à-dire que nous utilisons la variante Monte-Carlo EM (appelée MCEM) de l'algorithme.

Cette nouvelle méthodologie est testée sur des exemples simulés, puis sur le jeu de données réelles de l'encéphalopathie spongiforme bovine (ESB) en France. L'ESB est une maladie très rare, et se place donc certainement aux limites de la méthode. De plus, cette épidémie a déjà été cartographiée avec l'approche bayésienne par Abrial *et al.* (2004), ce qui nous donne un point de comparaison.

Bibliographie

[1] Abrial D., Calavas D., Jarrige N., et Ducrot C. (2004) Spatial heterogeneity of the risk of BSE in France following the ban of meat and bone meal in cattle feed. *Preventive Veterinary Medicine*, 67, 69-82.
[2] Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *Jounal of Royal Statistical Society B.* 2, 192-236.
[2] Besag, J. (1975) Statistical analysis of non latice data. *The Statistician*, 24, 179-195.
[4] Besag, J. (1986) On the statistical analysis of dirty pictures. *Jounal of Royal Statistical Society B.* 48(3), 259-302.
[5] Besag, J., York, J., et Mollié, A. (1991) Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1), 1-59.
[6] Chalmond, B. (2000) *Éléments de modélisation pour l'analyse d'images.* Mathématiques et Applications 33, Springer-Verlag.
[7] Chandler, D. (1987) *Introduction to Modern Statistical Mechanics.* Oxford University Press.

[8] Dempster, A.P., Laird, N.M. et Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm.*Jounal of Royal Statistical Society B.* 39(1), 1-38.
[9] Guyon, X. (1993) *Champs aléatoires sur un réseau : modélisation, statistique et application.* Masson.
[10] Mollié, A. (1999) Bayesian and Empirical Bayes approaches to disease mapping, dans *Disease Mapping and Risk Assessment* Lawson A., Biggeri A. et Bohning D., Wiley, 15-29.
[11] Wei, G.C.G. et Tanner, M.A. (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of American Statistical Association.* 85, 411
[12] Winkler, G. (1995) *Image analysis, random fields and dynamic Monte Carlo methods: a mathematical introduction.* Springer.