Approche variationnelle pour la cartographie spatio-temporelle du risque en épidémiologie à l'aide de champs de Markov cachés

David Abrial, Myriam Charras-Garrido, Florence Forbes

Approche variationnelle pour la cartographie spatio-temporelle du risque en épidémiologie à l'aide de champs de Markov cachés

David Abrial[2], Lamiae Azizi[1,2], Myriam Charras-Garrido[2] & Florence Forbes[1]

[1] *INRIA Rhône-Alpes & Laboratoire Jean Kuntzmann, Equipe Mistis,Inovallée, 655 av. de l'Europe, Montbonnot, 38334 Saint-Ismier cedex*

[2] *INRA, Unité d'Epidémiologie Animale, Centre de recherche de Clermont-Ferrand-Theix, 63122 Saint-Genès-Champanelle*

Résumé

L'analyse spatio-temporelle d'une épidémie permet aux épidémiologistes de comprendre son étiologie et fournit des suggestions pour planifier de nouvelles études pour examiner les causes sous-jacentes. Cette analyse donne lieu à une estimation du risque épidémiologique dans différentes unités géographiques et produit ainsi des cartes de risque permettant la détection de différences de niveau de risque à différents pas de temps. Nous proposons d'adapter une méthode par champs de Markov cachés discrets (issue de l'analyse d'images) dans le cadre spatial, pour permettre une classification intrinsèque des risques en vue du tracé des cartes de niveaux de risque et de l'étendre par la suite à un contexte spatio-temporel. Afin d'estimer les paramètres du modèle et de définir les classes, l'algorithme EM-champ moyen est utilisé.

Mots clés : Champs de Markov cachés, analyse d'images, algorithme EM, épidémiologie.

Abstract

The analysis of variation of risk for a given disease in space and time may give important clues about its aetiology and provide useful suggestions for planning further studies to investigate the underlying causes. This analysis lead to the estimation of the risk for each area and allow to detect the variation of rate in space and time. We propose to adapt a spatial method based on hidden markov random field approach (method used in image analysis) which lead to a ˜natural˝classification of the risk. We will extend this method to study the space-time variation on disease risk. To estimate the parameters of the model and define the risk classes, we use the EM-mean field algorithm.

Keywords : Hidden markov field, image analysis, EM algorithm, epidemiology.

L'objectif de la modélisation spatio-temporelle du risque de contamination en épidémiologie est de comprendre les mécanismes de la propagation de l'épidémie en

prenant en compte les dimensions spatiales et temporelles. Notre cadre d'étude est la France et les données disponibles sont des données agrégées par unités géographiques (le nombre de cas par canton par exemple). La France administrative est découpée en 3705 cantons (avec un nombre de voisins variable) ce qui rend la définition d'une notion de voisinage pour les champs de Markov assez complexe. Ainsi, la France a été découpée en 1240 hexagones pour palier l'hétérogénéité spatiale des cantons, unités administratives d'origine des données. Nos données sont le nombre de cas et l'effectif de la population par unité géographique.

Soit $y = (y_1, \ldots, y_n)$ le nombre de cas observés et $e = (e_1, \ldots, e_n)$ le nombre de cas attendus si la population était homogène. En chaque unité géographique, $\forall i \in (1, \ldots, n)$, y_i suit une loi de Poisson dont la moyenne est $e_i r_i$ avec $r = (r_1, \ldots, r_n)$ le risque inconnu. La loi du nombre de cas observés s'écrit comme :

$$P(Y_i = y_i | r_i) = exp(-e_i r_i) \frac{(e_i r_i)^{y_i}}{y_i!}.$$

L'estimation "naturelle"par maximum de vraisemblance du risque $\hat{r_i} = \frac{y_i}{e_i}$ conduit à des estimations non satisfaisantes, notamment dans le cas d'une maladie rare, et ne prend pas en compte la structure spatiale des unités géographiques.

Les modèles souvent utilisés par les épidémiologistes sont basés sur des approches de type Bayésien hiérarchique proposées par Besag et al. (1991). Ces modèles dans un cadre continu et spatial, suppose que pour chaque hexagone, on s'intéresse au nombre de cas et on fait l'hypothèse que cet effectif suit une loi de Poisson. Les paramètres de ces lois sont décomposés en plusieurs composantes faisant chacune l'objet d'information a priori (modèle Bayésien). On suppose, premièrement, que les risques de contamination ne sont pas indépendants d'un hexagone à l'autre. Aussi, une composante prenant en compte (*a priori*) une relation de structure spatiale de voisinage local a été ajoutée. Cette structure de voisinage est complétée par une composante spatiale (dite globale) sans structure particulière $log(r_i) = u_i + v_i$. Ces modèles ne permettent que l'estimation du risque par unité géographique et la classification est effectuée par les épidémiologistes *a posteriori* d'une manière empirique.

On reproche à ce genre de méthodes de produire des cartes de risque très lisses et de ne pas permettre de détecter les discontinuités. Pour palier ce genre de problème, d'autres modèles de type Bayésiens hiérarchiques discrets ont été proposés. Green et Richardson (2002) proposent de modéliser le risque comme issu d'un mélange $\forall r_i \in r_1, \ldots, r_k \sim \mathcal{M}(\pi_1, \ldots, \pi_k)$, d'introduire une variable d'allocation z (prenant des valeurs dans $1, \ldots, k$), qui prend en compte la structure spatiale et qui est relié au risque par la relation $r_i = r_{z_i}$ et $z_i \in 1, \ldots, K$(indépendants) $\sim$ Finite Poisson Mixture. Cette variable est modélisée comme étant un champ de Markov caché. Green et Richardson utilisent ensuite pour l'inférence un cadre bayésien et une

résolution par algorithme à sauts réversibles, très couteux en calculs.

Nous proposons d'utiliser comme Green et Richardson des champs de Markov cachés discrets mais en adoptant pour leur estimation une approche alternative basée sur l'algorithme EM [3] et l'approximation par champ moyen. Outre un moindre cout de calcul, cette approche a l'avantage de fournir une interprétation plus facile pour les épidémiologistes. Dans notre cas, le champ caché correspond à une classification du risque qui fait partie des paramètres à estimer. L'avantage de cette méthode est que la classification est faite en même temps que l'estimation.

Dans le cadre des champs aléatoires de Markov cachés, les variables cachées z_i (les classes) sont liées aux données observées y_i (nombre de cas) par les lois des observations conditionnelles :

$$P(Y_i = y_i|y_{-i}, z, r) = P(Y_i = y_i|z_i, r) = P(Y_i = y_i|r_{z_i}).$$

Connaissant les variables cachées $z = (z_i)$, la distribution de la variable Y_i ne dépend pas des observations y_{-i} effectuées dans les autres unités géographiques.

La loi d'un champ de Markov peut s'exprimer comme une loi de Gibbs :

$$P(Z = z) = W^{-1}exp(-H(z)).$$

H l'énergie du champ z, est exprimée comme une somme de fonctions potentiels $V_c(z)$ sur les cliques. W est la constante de normalisation. Nous nous limitons aux fonctions potentielles d'ordre 1 et 2, cela correspond à une énergie de la forme :

$$H(z) = \sum_i (V_i(z_i) + V_{i,j}(z_i, z_j))$$

avec les potentiels sur les singletons $V_i(z_i)$ qui permettent de modéliser la probabilité d'occurence de la classe z_i au site i considéré individuellement et les potentiels sur les paires $V_{i,j}(z_i, z_j)$ permettent de modéliser la dépendance entre les classes Z_i et Z_j en des sites i et j voisins.

Les classiques fonctions potentiels sur les paires, telles celles de type Potts (qui ne prennent en compte que l'égalité ou non des variables cachées voisines) ne tiennent pas compte de la notion de gradation des risques, à laquelle on s'attend, entre les classes. Nous proposons alors des fonctions de potentiels qui permettent d'eviter que deux classes extrêmes se retrouvent côte à côte.

Nous utilisons l'algorithme EM-champ moyen [2] pour l'estimation des paramètres du modèle. le principe de cet algorithme repose sur l'idée de remplacer le modèle de champ de Markov caché de loi $P(y, z|\psi)$, avec ψ les paramètres du modèle, par une approximation de type champ moyen définie par :

$$P(y, z|\psi) = \prod_{i \in I} P_{\tilde{z}^y}(y_i, z_i|\psi) = \prod_{i \in I} P(z_i|\tilde{z}^y_{N_i}, \theta) f(y_i|r_{z_i})$$

Trois choix sont naturels pour le champ des voisins $\tilde{z}^y$, conduisant aux algorithmes en champ moyen (mean fied), en champ modal (modal field) et en champ simulé (simulated field). Pour illustrer cette méthodologie, nous présentons des résultats d'application sur des données épidémiologiques simulées.

Nous proposons par la suite d'étendre ce modèle spatial au cas spatio-temporel avec une incorporation de la dimension temporelle.

Bibliographie

[1] 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.
[2] Celeux, G., Forbes, F., Peyrard, N. (2003) *EM procedures using mean field-like approximations for Markov model-based image segmentation*, Pattern Recognit. 36(1), 131-144
[3] Dempster, A.P., Laird, N.M. et Rubin, D.B. (1977) *Maximum likelihood from incomplete data via the EM algorithm*, Journal of Royal Statistical Society B.39(1), 1-38.
[4] Green, P.J. et Richardson, S. (2002) *Hidden Markov models and disease mapping*, Journal of the American Statistical Association, 97, 1055-1070.
[5] Mollié, A. (1999) *Bayesian and Empirical Bayes approaches to disease mapping*, Wiley, 15-29.