DIC et autres critères de sélection en cartographie du risque

Ioana Molnar [1] & Sylvain Coly [2] & Myriam Charras-Garrido [3] & David Abrial [4] & Anne-Françoise Yao-Lafourcade [5]

[1,2,3,4] *Centre INRA Auvergne/Rhône-Alpes, Unité d'Épidémiologie Animale, Route de Theix, 63122 Saint-Genès-Champanelle, France*
[2,5] *Université Blaise Pascal – Laboratoire de Mathématiques, UMR 6620 – CNRS, Campus des Cézeaux, B.P. 80026, 63171 Aubière cedex, France*
[1] *ioana.molnar@clermont.inra.fr,* [2] *sylvain.coly@clermont.inra.fr,*
[3] *myriam.charras-garrido@clermont.inra.fr,* [4] *david.abrial@clermont.inra.fr*
[5] *anne-francoise.yao@math.univ-bpclermont.fr*

Résumé. L'un des objectifs de la cartographie du risque appliquée en épidémiologie est l'étude et la comparaison de plusieurs modèles pour apporter un soutien aux différentes hypothèses concernant une maladie donnée. Il s'agit notamment du choix des variables explicatives, de leurs interactions, de leur influence sur le risque modélisé ou des hypothèses sur leurs fonctions de répartition. Ayant d'habitude un grand nombre de modèles implémentés pour un problème donné, on a besoin d'une méthode pour comparer leurs performances et sélectionner le meilleur modèle parmi tous ceux considérés. Idéalement, on voudrait identifier celui qui représente le meilleur compromis entre l'ajustement aux données, le lissage des cartes obtenues, la sensibilité du modèle aux perturbations, sa généralité, sa simplicité et sa facilité d'estimation. On voudrait être capable de comparer des modèles imbriqués aussi que des modèles assez éloignés les uns des autres. A cause de ces divers objectifs, il n'existe pas de procédure de sélection de modèle qui soit universellement valable. Le sujet de notre étude concerne le problème d'identification d'une méthode de sélection de modèle adaptée au cadre spatio-temporel de la cartographie du risque des maladies contagieuses. Notre objectif est d'investiguer la performance du critère le plus utilisé dans la cartographie du risque, le critère DIC, et de proposer des solutions mieux adaptées à notre cadre.

Mots-clés. critère DIC, cartographie du risque, méthodes bayésiennes, modélisation spatio-temporelle, sélection de modèle, données épidémiologiques.

Abstract. One of the missions of disease risk mapping applied in epidemiology is the study and comparison of several models to provide support to different theories concerning a specific disease. These include the choice of explanatory variables, of their interactions, of their influence on the risk, or of the assumptions about their distribution functions. Having usually a large number of models implemented for one problem, we need a method to compare their performance and select the best model among those considered. Ideally,

we would like to identify which one represents the best compromise between the adjustment to the data, the smoothing of the resulting maps, the robustness to perturbations, the generality of the model, its simplicity and its ease of estimation. We would want to be able to compare nested models and rather distinct models alike. Because of these different goals, there is no model selection procedure that is universally valid. The subject of our study concerns the problem of identifying a model selection method adapted to the spatio-temporal risk mapping of contagious diseases. Our goal is to investigate the performance of the most widely used criterion in risk mapping, the Deviance Information Criterion, and propose solutions that better suit our context.

Keywords. Deviance Information Criterion, disease risk mapping, Bayesian methods, spatio-temporal modeling, model selection, epidemiological data.

1 Sélection de modèle en cartographie du risque

La cartographie du risque est utilisée pour l'estimation et l'analyse spatiale du risque sous-jacent à l'incidence observée d'une maladie. Grâce à son potentiel d'identification des motifs, la cartographie peut aider à mieux comprendre le comportement de la maladie. En effet, la modélisation statistique des données de comptage d'une maladie donne lieu au lissage des données, à la détection des agrégats et à la détermination, au sein d'un ensemble donnée des co-variables susceptibles d'être influentes, de celles qui seraient les plus pertinentes. En outre, cet outil peut fournir des informations sous une forme aisément exploitable, constituant ainsi un soutien pratique aux mesures de santé publique. Enfin, il peut également être utilisé pour la prédiction à court-terme, participant ainsi à la prévention des épidémies.

La sélection de modèles est un élément intrinsèque au processus de modélisation. Dans le contexte des modèles bayésiens hiérarchiques, ce problème demeure ouvert, car il n'y a aucun outil formel de sélection de modèle complètement approuvé. Naturellement, la méthode de sélection que l'on choisit d'utiliser devrait dépendre prioritairement des objectifs prévus de l'analyse, fixés préalablement. Par exemple, on pourrait rechercher un modèle qui ait le meilleur pouvoir explicatif, ou la meilleure précision prédictive, on pourrait donner la priorité à la parcimonie ou au lissage, à l'identification des zones à hauts risques, ou encore, on pourrait vouloir optimiser simultanément plusieurs de ces objectifs. D'autre part, l'étendue de l'applicabilité et la facilité d'utilisation sont des caractéristiques non-négligeables que l'on a souvent besoin de prendre en compte quand on adopte un critère de sélection de modèle spécifique, en particulier lorsque les modèles ajustés sont nombreux, complexes ou très divers. Plus que dans la statistique bayésienne en général, dans la cartographie du risque, qui vise à évaluer, expliquer et (dans une moindre mesure) prédire, non pas la survenue des cas, mais un risque latent, le problème de l'évaluation et de la comparaison des modèles est d'autant plus compliqué.

2 Commentaires sur le critère DIC

La méthode de comparaison la plus largement utilisée en cartographie du risque est le Deviance Information Criterion (DIC), introduit par Spiegelhalter et al. (2002) : DIC = $\overline{D(\theta, y)} + p_D$. Il est construit d'une manière classique pour un critère d'information, sous la forme d'une mesure du non-ajustement du modèle, $\overline{D(\theta, y)}$, pénalisée par un terme additif, p_D, mesurant la complexité du modèle. On rappelle que $D(\cdot, y) = -2 \log p(y \mid \cdot)$ (à une constante près) est la déviance bayésienne, avec $p(y \mid \cdot)$ la fonction de vraisemblance qui définie le modèle, y les données observées, θ le vecteur des paramètres, et que

$$p_D = \overline{D(\theta, y)} - D(\overline{\theta}, y). \tag{1}$$

La justification théorique du DIC est valable seulement asymptotiquement, dans le cadre des modèles non-hiérarchiques et lorsque la vraisemblance du modèle appartient à la famille des distributions exponentielles (Zhu et Carlin (2000), van der Linde (2005)). En revanche, le DIC a prouvé une efficacité remarquable dans des cas en-dehors de ce cadre spécifique et restreint. Néanmoins, les failles de ce critère sont également largement connues (voir, par exemple, Speigelhalter et al. (2014) et Celeux et al. (2006)). On évoque ici que les inconvénients du DIC qui sont les plus visibles et les plus inquiétants dans le contexte de la cartographie du risque.

Le premier défaut est que la formule (1) peut conduire à $p_D < 0$. En fait, cela semble être un problème récurrent, commun à de nombreuses études, et l'interprétation d'une valeur négative pour le nombre effectif de paramètres n'est pas du tout évidente. Quelques explications possibles pour l'apparition de ce phénomène ont été proposées – comme le fait que les distributions *a priori* sont trop asymétriques, ou bimodales, ou qu'il existe un conflit entre la vraisemblance et les aprioris, ou tout simplement que les modèles sont défectueux. Bien que ces raisons puissent être vraies, la fréquence des situations où p_D est négatif, ou, le fait qu'il puisse même passer, par une petite perturbation, d'une valeur positive à une valeur négative, est décourageant et ne peut pas être ignoré.

Une deuxième faiblesse du DIC est sa tendance à privilégier les modèles sur-adaptés. Ces modèles sont indésirables parce qu'ils manquent de robustesse et s'opposent au principe de parcimonie. En cartographie du risque, cela présente en outre l'inconvénient de produire des cartes très bruitées qui sont inutiles du point de vue pratique, car elles apportent très peu d'information et sont peu interprétables.

En dépit des nombreuses critiques que le DIC a reçues au cours du temps, sa grande popularité en analyse bayésienne en général est bien installée dans le cas particulier de la cartographie du risque, où il est systématiquement utilisé. Il semble que la principale raison de sa présence presque obligatoire dans la boîte à outils de la sélection de modèle dans de telles études est sa facilité de calcul et également l'absence d'un sérieux concurrent. En effet, à l'exception de quelques variantes du DIC développées par différents auteurs (et qui n'ont pas réussi à prendre sa place), il reste le seul critère d'information facilement

implémentable représentant un compromis entre la complexité et la qualité d'ajustement du modèle applicable aux modèles bayésiens hiérarchiques complexes.

3 Présentation des modèles considérés

Notre étude s'appuie sur un ensemble de données représentant des cas de tuberculose bovine identifiés en France, annuellement, entre 2001 et 2010 (un *cas* signifiant une ferme d'élevage de bovins contenant des animaux atteints par la tuberculose). Sur une carte de France partitionnée en 448 hexagones, le nombre total des cas survenus dans une année est recensé, ainsi que le nombre total d'exploitations présentes dans chaque région, supposé constant sur l'ensemble des périodes concernées. Les données sont de type spatio-temporel, chaque valeur est associée à l'un des hexagones indexés par $i = 1, \ldots, 448$ et à l'une des périodes indexées par $j = 1, \ldots, 10$. A partir de ces données réelles, on cherche à identifier le risque sous-jacent à la manifestation observée de la maladie. Pour cela, on modélise les cas survenus à l'endroit i au moment j par des variables aléatoires discrètes et positives, notées Y_i^j.

On considère 63 modèles hiérarchiques bayésiens, construits sur trois niveaux. Le premier niveau concerne la distribution paramétrique des données de comptage. Le deuxième niveau est dédié à la structure spatio-temporelle du risque relatif sous-jacent à la maladie et les distributions *a priori* des variables explicatives. Enfin, le dernier niveau est réservé à la description complète des hyperparamètres utilisés au deuxième niveau; on utilise un choix classique : la distribution Gamma.

Au premier niveau de nos modèles, on considère deux distributions de probabilité distinctes décrivant les données de comptage (Y_i^j) – la loi de Poisson et la loi binomiale négative (sous la forme de mélange Poisson-Gamma). D'une part, la distribution de Poisson est le choix habituel en cartographie du risque pour la modélisation du nombre d'occurrences d'une maladie rare. D'autre part, la distribution binomiale négative pourrait être plus adaptée à la modélisation d'une maladie infectieuse, car elle peut prendre en compte la surdispersion locale associée au phénomène de contagion.

Le paramètre d'intérêt est la moyenne λ_{ij} de la distribution de Poisson et le focus est mis sur les risques relatifs (r_i^j), quantité définie par le rapport entre la moyenne λ_{ij} et l'attendu (fixé) qui prend en compte la taille de la population et le nombre total des cas survenus dans chaque période temporelle. Au deuxième niveau des modèles, on considère le modèle log-linéaire pour le risque relatif avec quatre effets aléatoires

$$\ln(r_i^j) = a \cdot U_i^j + b \cdot T_i^j + c \cdot V_i^j + d \cdot \varepsilon_i^j, \tag{2}$$

d'une manière assez similaire aux modèles habituellement utilisés en cartographie du risque. Les composantes U_i^j et T_i^j décrivent, respectivement, l'influence spatiale et temporelle, et la variable V_i^j décrit leur interaction spatio-temporelle. Le terme résiduel ε_i^j prend en compte l'hétérogénéité individuelle, et il est modélisé par un bruit blanc gaussien.

L'utilisation des poids est inspirée par Cressie et Stern (1999) et quantifie l'importance de chaque composante dans la spécification du log-risque relatif.

En vue de modéliser l'agrégation des cas (et donc une partie de la surdispersion) on utilise pour les composantes spatiales et/ou temporelles les processus de type CAR (Conditionnels Auto-Régressifs). Le voisinage spatial est défini par la relation d'adjacence et le voisinage temporel par la relation d'antériorité et de postériorité.

4 Discussion sur la performance des différents critères

Dans notre étude, nous nous sommes principalement intéressés à la structure du risque latent, qui se manifeste comme motifs visibles et dont l'origine est l'influence des composantes spatio-temporelles. Par conséquent, nous voulons choisir un modèle qui révèle des tendances temporelles et des régions dont les valeurs de risque soient semblables à leurs voisins. En outre, nous cherchons à obtenir des cartes qui permettent une interprétation facile des résultats, donc la qualité du lissage de la carte qui en résulte est d'une grande importance. Ces critères spécifiques ne sont pas incorporés dans une formule de sélection d'un modèle, en particulier à cause de la grande diversité des études concernant la cartographie du risque. Cependant, il est indispensable de chercher un bon critère de sélection, même s'il restera imparfait. Idéalement, un tel critère serait un outil souple et facile à calculer qui filtre les modèles non-convenables et qui classe ceux qui sont pertinents.

Dans notre étude, nous utilisons quelques mesures classiques facilement calculables dans un double but : la comparaison des modèles et la comparaison des critères. Plus précisément, d'une part on compare les modèles en évaluant différents aspects, comme la qualité d'ajustement ou le lissage. D'autre part, on compare les résultats obtenus avec ceux obtenus *via* le critère DIC, et ainsi nous évaluons leurs performances et leur cohérence. Les mesures de la famille de l'EQM (Erreur Quadratique Moyenne) et le coefficient de Spearman (appliqués tous deux sur le risque relatif, et non pas sur les données) ne peuvent être utilisés que dans des études de simulation car ils nécessitent la connaissance du véritable risque relatif sous-jacent. Bien qu'ils mesurent l'ajustement du modèle, ils n'ont pas la capacité de mesurer son pouvoir de prédiction, et ignorent complètement la notion de parcimonie. Ils peuvent cependant être utilisés pour évaluer la performance du DIC par une étude comparative. En effet, trop de divergences entre les modèles favorisés par le DIC et ceux qui sont favorisés par l'EQM ou le ρ de Spearman peuvent indiquer une absence de pertinence du DIC dans le choix du modèle pour notre étude.

Afin d'évaluer le caractère lisse des cartes obtenues, il est raisonnable d'utiliser une mesure de l'auto-corrélation entre les valeurs du risque relatif obtenues pour des régions adjacentes spatialement et temporellement, comme l'indice de Moran. Il est dépourvu de toute sorte d'idée d'ajustement du modèle, mais entraîne la mesure du lissage, perception

que le DIC n'est pas capable d'avoir. En fait, le DIC et l'indice de Moran peuvent être vus comme totalement complémentaires l'un à l'autre.

Les objectifs spécifiques que nous nous sommes fixés nous ont conduits naturellement à considérer l'utilisation combinée de plusieurs méthodes de comparaison différentes. Il nous semble pertinent d'utiliser la puissance de plusieurs méthodes de sélection de modèles dans le but de créer un critère plus efficace. Nous nous limitons à l'utilisation simultanée de deux méthodes, qui sont toutes deux applicables dans le cas où le vrai risque sous-jacent est inconnu, donc sur des données réelles. Nous proposons deux approches différentes. La première approche, que nous appelons critère de couplage, consiste à sélectionner les modèles en les classant selon deux critères successifs. C'est une méthode utile dans le cas où le critère souhaité n'est pas capable d'établir une hiérarchie entre plusieurs modèles distincts. Il arrive que, sous un certain aspect, des modèles montrent une performance similaire, tandis qu'ils sont très différents sous une autre perspective. La méthode du couplage peut aussi être utilisée lorsqu'on veut filtrer les modèles selon un certain critère, pour enlever ceux qui ne sont pas du tout satisfaisants, et qu'on se propose de faire une sélection finale selon un deuxième critère. La seconde méthode proposée, que nous appelons le critère d'information de déviance lisse, est une nouvelle variante du DIC. Dans la définition du critère, nous remplaçons la pénalisation additive par nombre effectif des paramètres pour une pénalisation multiplicative selon le manque de lissage des cartes. Les résultats que nous avons obtenus mettent en évidence l'intérêt et la pertinence de ces nouveaux critères.

Bibliographie

[1] Celeux, G., Forbes, F., Robert, C. P., Titterington, D. M. (2006), Deviance information criteria for missing data models, *Bayesian Analysis*, 1 (4), 651–673.

[2] van der Linde, A. (2005), DIC in variable selection, *Statistica Neerlandica*, 59 (1), 45–56.

[3] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., van der Linde, A. (2002), Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64 (4), 583–639.

[4] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., van der Linde, A. (2014), The deviance information criterion: 12 years on, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 76 (3), 485–493.

[5] Stern, H., Cressie, N. (1999), Inference for extremes in disease mapping, *Disease mapping and risk assessement for public health*, John Wiley & Sons, 61–82.

[6] Zhu, L. et Carlin, B. P. (2000), Comparing hierarchical models for spatio-temporally misaligned data using the deviance information criterion, *Statistics in Medicine*, 19 (17-18), 2265–2278.