Application de la cartographie du risque aux données contagieuses

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

[1,2,3] *Centre INRA de Clermont-Ferrand/Theix*
Unité d'Épidémiologie Animale
Route de Theix
63122 Saint-Genès-Champanelle
[1] *sylvain.coly@clermont.inra.fr*
[2] *myriam.charras-garrido@clermont.inra.fr*
[3] *david.abrial@clermont.inra.fr*

[1,4] *Université Blaise Pascal - Laboratoire de Mathématiques*
UMR 6620 - CNRS
Campus des Cézeaux
BP 80026
63171 Aubière CEDEX
France
[4] *Anne-francoise.Yao@math.univ-bpclermont.fr*

Résumé. La cartographie du risque appréhende la répartition du risque associé à une pathologie et le représente sous la forme de carte suivant un dégradé de couleurs. Depuis son introduction par Besag, la cartographie du risque a connu de nombreuses améliorations et variantes méthodologiques, pour couvrir un spectre de problématiques de plus en plus large. Cette approche est usuellement appliquée à des maladies non-contagieuses ; dans ce cas les dépendances spatiales (voire spatio-temporelles) sont liés à des facteurs environnementaux et populationnels. Notre objectif est d'appliquer la cartographie du risque à des maladies infectieuses, pour lesquelles un cas primaire peut engendrer des cas secondaires. La contagion peut être source de surdispersion et de renforcement des structures spatiales et temporelles. Nous avons testé 60 modèles aux lois de comptage et aux structures de risque différentes sur des données simulées (agrégats de cas) et sur des données réelles (tuberculose bovine). Le mode de sélection de modèle est le critère DIC (Deviance Information Criterion). Cette étude montre la pertinence d'utiliser la loi binomiale négative par rapport à la loi de Poisson dans le cas de données surdispersées et/ou présentant des niveaux de risque contrastés. Elle conclut à la nécessité de prendre en compte les dimensions spatiale et temporelle dans ce type d'étude épidémiologique. Elle permet par ailleurs d'appréhender la répartition du risque de la tuberculose bovine en France, ainsi que sa structure. Ces conclusions ouvrent des perspectives sur différents

sujets méthodologiques tels que la recherche des modèles les plus adaptés ou la sélection de modèle.

Mots-clés. Cartographie du Risque, Épidémiologie, Inférence bayésienne, Maladie Contagieuse, Spatio-Temporel.

Abstract. Disease mapping aims to determinate the underlying disease risk scattered from health data and to represent it by a gradation of colours on a map. Since their introduction by Besag, disease mapping methods have undergone many methodological improvements and variations, to treat a wider and wider range of issues. This approach is usually dedicated to non-infectious diseases; in this case the spatial or (spatio-temporal) dependencies are due to factors coming from the environment or from population. Our aim is to apply disease mapping to infectious diseases; when a primary case can result in secondary cases, by direct or vector transmission. Contagion can lead to overdispersion and strengthen spatial and temporal structures. So we tested 60 models which have different probability distribution and different parameters for the structure of risk, on both simulated outbreaks and on bovine tuberculosis data. These various models have been compared for each dataset using the DIC criterion (Deviance Information Criterion). This study highlighted the relevance of using the negative binomial distribution compared to the Poisson distribution in the case of over-dispersed data and/or with highly contrasting levels of risk. It also shows the need to take into account both spatial and temporal dimensions in this type of epidemiological study. It also provides information on the distribution of risk for bovine tuberculosis in France, as well as on its structure. These findings open prospects on many issues such as the importance of spatial and temporal scales for the estimation of structural parameters, or the research of the most relevant parameters to explain the risk and the most suitable distribution to treat this type of problem.

Keywords. Bayesian Inference, Disease Mapping, Epidemiology, Infectious Disease, Spatiotemporal.

1 Contexte et méthodes

La cartographie du risque permet d'appréhender la structure sous-jacente à des données de santé spatiales dispersées [1]. On cherche en fait à représenter le risque inhérent au phénomène étudié sous la forme de cartes lissées, par un dégradé de couleur [2]. Au moment de son introduction par Besag, ce type de méthode a été développé pour réaliser des atlas des différents types de cancer aux États-Unis [3]. Son champ d'application s'est fortement élargi depuis. Cette diversification des problèmes traités par cette approche est induite par d'importants changements méthodologiques. Si les modèles bayésiens hiérarchiques à trois niveaux sont toujours utilisés, leurs différentes composantes ont beaucoup évolué. Dans cette classe de modèles, le premier niveau permet de modéliser

les données par une loi de comptage, le second définit la structure de la moyenne de cette loi, et le troisième conditionne l'ensemble des paramètres du modèle.

Le cadre bayésien est propice à la prise en compte de connaissances *a priori* sur la pathologie d'intérêt. Dans la mesure où la plupart des études se penche sur une unique maladie, ses caractéristiques peuvent être mises à profit pour construire des modèles bien ajustés. L'influence de covariables environnementales peut être testée [4,5], on pourra les introduire au second niveau du modèle pour voir quels facteurs sont les meilleurs substituts de la structure spatiale du risque [2]. Cette structure spatiale elle-même peut être perçue comme l'expression de covariables inconnues.

Tous les facteurs associés à une maladie ne peuvent être identifiés. Ainsi, des dépendances spatiales subsistent dans les données de régions voisines [2]. Un des soucis majeurs de la cartographie du risque est la recherche de processus ou fonctions modélisant au mieux ces dépendances. Un grand nombre d'approches a été envisagé pour prendre en compte les corrélations spatiales : différents processus conditionnels auto-régressifs [6,7], une tendance spatiale [8] ou classification de la région d'étude [9]. Le cadre spatial de la cartographie du risque s'est naturellement étendu au spatio-temporel [8,10], ce qui sous-entend des améliorations méthodologiques importantes, autant dans la structure du modèle que dans la façon même d'appréhender ces problèmes plus complexes. Différentes méthodes ont été considérées pour la modéliser, dans le second niveau du modèle hiérarchique : tendance linéaire [8], fonctions polynomiales [11], marches aléatoires [12,13], processus conditionnels auto-régressifs spatiaux (voir spatio-temporels) [4,12], splines [14], ... Dans notre étude, notre point de vue n'est pas centré sur une unique pathologie, nous ne souhaitons donc pas inclure de connaissance *a priori* à nos modèles. Dans la mesure où l'on suppose que la contagion peut se traduire par une augmentation de la structuration des données, nous testons donc la pertinence de toutes les combinaisons possibles de processus CAR spatial, temporel, spatio-temporel et de bruit blanc gaussien. Nous nous intéressons également à l'intérêt de leur attribuer des coefficients de pondération sur le modèle des travaux de Cressie [15].

Comme suggéré par les trois paragraphes précédents, la plupart des réflexions méthodologiques sur la cartographie du risque concerne le second niveau du modèle hiérarchique bayésien. Néanmoins la loi de distribution de premier niveau est également l'objet de tests. La plus utilisée est la loi de Poisson, bien qu'au départ la loi binomiale ait été tout aussi employée. D'autres lois ont été considérées pour répondre à des problématiques spécifiques. Ainsi par exemple des lois telles que la loi binomiale négative [16] ont été employées dans l'analyse de données surdispersées. En effet, elles sont définies par deux paramètres, ce qui leur confère davantage de souplesse que la loi de Poisson. Pour notre étude, nous testons la distribution de Poisson (la plus utilisée en cartographie du risque) et la loi binomiale négative *a priori* davantage en mesure de prendre en compte la surdis-

persion des données.

2 Résultats

Simulations d'épidémie. De façon à caractériser la méthode et évaluer les modèles les plus performants, on analyse des jeux de données contenant des épidémies simulées. Plus précisément, on simule un niveau de risque moyen dans lequel se situent trois zones à haut risque, l'une fixe, l'autre se déplaçant linéairement au cours du temps, et la dernière augmentant puis diminuant en intensité au cours du temps. On considère par ailleurs deux niveaux d'intensité de risque différents. On simule ensuite les cas suivant deux lois de comptage différentes (binomiale négative avec surdispersion et loi de Poisson sans surdispersion).

Cette étude a permis de constater que les modèles avec la loi binomiale négative comme distribution de premier niveau avaient de meilleures performances dans le cas de données surdispersées et/ou quand le risque est bien contrasté sur l'ensemble de la zone d'étude. Par ailleurs les modèles les mieux classés incluent les dimensions spatiale et temporelle dans leur structure de risque. Par ailleurs, les bruits gaussiens sont parfois considérés comme pertinents. Enfin les paramètres de pondération pénalisent largement les modèles qui les comportent.

On remarque par ailleurs que les épidémies fixes dans le temps et géographiquement sont beaucoup mieux identifiées. Lorsqu'elles se déplacent, elles sont tantôt identifiées, tantôt non. Par ailleurs les épidémies dont l'intensité augmente puis diminue vont être reconnues, puis disparaître à partir d'un certain pallier d'intensité. En outre, un signal est beaucoup mieux identifié lorsqu'il intervient dans une région fortement peuplée que faiblement peuplée.

Cas de tuberculose bovine. Les données de tuberculose bovine couvrent les années de 2000 à 2010, elles sont accessibles sur l'ensemble de la France. Elles sont constituées des exploitations concernées par la tuberculose bovine. On comptabilise les "cas" à l'échelle de petits hexagones de 60 km de côté, et pour des périodes d'un an. Les données relatives à l'année 2000 ne sont pas conservées, les données de cas étant beaucoup plus faibles et laissant supposer un nombre important de valeurs manquantes.

Les modèles définis comme étant les meilleurs au sens du DIC présentent des caractéristiques similaires. Tout d'abord ils emploient la loi binomiale négative comme distribution de comptage au premier niveau, ce qui est cohérent avec la forte surdispersion observée dans les données, induite par la très forte fréquence de valeurs nulles, et la survenue de valeurs

importantes par rapport à la moyenne globale. Par ailleurs, les modèles bien classés comportent au moins deux composantes de structure du risque (spatiale, temporelle et/ou spatio-temporelle) qui prennent donc en compte les dimensions spatiale et temporelle. Par ailleurs les coefficients de pondération sont considérés comme pertinents, et donnent une importance très supérieure aux composantes spatiales et spatio-temporelles (>1 pour les deux) par rapport au processus temporel (<0.1). Les lois de comptage considérées et les processus incorporés à la décomposition du risque expliquent la plus grande part de l'hétérogénéité des données, en effet les modèles qui comportent un bruit blanc gaussien ont tendance à être pénalisés .

Les zones identifiées comme étant à risque sont cohérentes avec les connaissances épidémiologiques sur le sujet. Le centre de la France est relativement touché, en particulier aux alentours des départements de la Côte d'Or et de la Dordogne. Ce constat est logique avec le relevé de cas au niveau départemental qui ciblent ces deux secteurs comme les plus critiques en France vis-à-vis de cette pathologie. L'Auvergne, bien que située entre ces deux départements, n'est que très peu concernée par le risque de tuberculose bovine. La Bretagne, principale terre d'élevage française est quasi-indemne, ce qui présente un enjeu économique fort. Le Sud-Est de la France est identifié comme à risque, bien que non-reconnu comme tel. Ce phénomène peut s'expliquer par la survenue de quelques cas en Camargue, dans un contexte de population faible en exploitations d'élevage bovin, peut-être sur fond de problème de bord.

3 Conclusion

La cartographie du risque, appliquée dans un cadre spatio-temporel, a permis de retrouver les connaissances que l'on avait sur la répartition et la structure du risque de tuberculose bovine en France dans les années 2000. Cette étude, que ce soit sur données réelles ou simulées a confirmé la pertinence de considérer la loi binomiale négative comme distribution de comptage au premier niveau du modèle, mais aussi lorsque les risques sont très contrastés sur l'ensemble du territoire, même sans surdispersion. On a également mis en évidence l'intérêt de l'ajout de la dimension temporelle dans la cartographie du risque, puisqu'elle apparaît comme indispensable dans tous les modèles identifiés comme les plus performants pour chaque jeu de données. Enfin, notre étude a mis en évidence la pertinence de l'application de la méthode à des jeux de données contagieuses, la structuration du risque ayant alors une place prépondérante dans l'explication de l'hétérogénéité observée dans les données.

Références

[11] Assunção R. M., Reis I. A. and Oliveira C. D. L. (2001). Diffusion and prediction of Leishmaniasis in a large metropolitan area in Brazil with a Bayesian space-time model. *Statistics in Medicine*, 20(15):2319-2335.
[8] Bernardinelli L., Clayton D., Pascutto C., Montomoli C., Ghislandi M. and Songini, M. (1995). Bayesian analysis of space-time variation in disease risk.
[6] Besag J., York J., and Mollié A. (1991). *Ann. Inst. Statist. Math.*, 43(1):1-59.
[1] Best N., Richardson S. and Thomson A. (2005). A comparison of bayesian spatial models for disease mapping. *Statistical methods in medical research*, 14(1):35-59.
[15] Cressie N. (1993). Statistics for spatial data. *Wiley series in probability and mathematical statistics: Applied probability and statistics*. J. Wiley.
[13] Hossain M. M. and Lawson A. B. (2010). Space-time Bayesian small area disease risk models: development and evaluation with a focus on cluster detection. *Environmental and ecological statistics*, 17(1):73-95.
[7] Knorr-Held L. and Besag J. (1998). Modelling risk from a disease in time and space. *Statistics in medicine*, 17(18):2045-2060.
[2] Lawson A., Biggeri A. B., Boehning D., Lesaffre E., Viel J.-F., Clark A., Schlattmann P. and F. D. (2000). Disease mapping models: an empirical evaluation. *Statistics in medicine*, 19:2217-2241.
[10] MacNab Y. C. (2003). Hierarchical Bayesian modeling of spatially correlated health service outcome and utilization rates. *Biometrics*, 59:305-316.
[14] MacNab Y. C. and Dean C. B. (2002). Spatio-temporal modelling of rates for the construction of disease maps. *Statistics in medicine*, 358(April 2001):347-358.
[3] Mason T., MacKay F., Hoover R., Blot W. and Fraumeni J. (1975). Atlas of Cancer Mortality for US counties, 1950-1969. Technical report, Department of Health, Education, and Welfare Publication, Washington.
[4] Ocana-Riola R. (2007). The misuse of count data aggregated over time for disease mapping. *Statistics in medicine*, 26:4489-4504.
[12] Richardson S., Abellan J. J. and Best N. (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.
[16] Tsutakawa R. K. (1988). Mixed model for analyzing geographic vari- ability in mortality rates. *Journal of the American Statistical Association*, 83:37-42.
[9] Wikle C. K. and Anderson C. J. (2003). Climatological Analysis of Tornado Report Counts Using a Hierarchical Bayesian Spatio-Temporal Model. *Journal of Geophysical Research*, 108.
[5] Xia H. and Carlin B. P. (1998). Spatio-temporal models with errors in covariates: mapping Ohio lung cancer mortality. *Statistics in medicine*, 17:2025-2043.