Spatiotemporal disease mapping applied to infectious diseases

Sylvain Coly\textsuperscript{a,b*}, Myriam Charras-Garrido\textsuperscript{a}, David Abrial\textsuperscript{a}, Anne-Françoise Yao-Lafourcade\textsuperscript{b}

\textsuperscript{a}Centre INRA de Clermont-Ferrand/Theix – Unité d’Épidémiologie Animale, Route de Theix, Saint Genès Champanelle 63122, France
\textsuperscript{b}Université Blaise Pascal – Laboratoire de Mathématiques UMR 6620, 24 Avenue des Landais, Aubière 63171, France

Abstract

Disease mapping aims to determine the underlying disease risk scattered from health data. This methodology enables to represent this disease risk by a gradation of colours on a map. 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. This study highlighted the relevance of using the negative binomial distribution to model such data. It also showed the need to take into account both spatial and temporal dimensions in this type of epidemiological study.

Keywords: Bayesian Inference; Disease Mapping; Epidemiology; Infectious Disease; Spatiotemporal.

1. Introduction

Disease mapping aims to highlight the underlying structure of scattered spatial health data\textsuperscript{1}. It is relevant to apply such a methodology when the underlying risk is meant to be spatially, temporally and/or spatiotemporally structured. It allows representing the structure of the risk in the form of smoothed colored maps\textsuperscript{2}. These methods were introduced by Besag in the seventies. They were initially used to study the distribution of different cancers in the United States\textsuperscript{3}. They have also been applied in European countries and then adapted to other diseases. Their scope has greatly expanded during the last decades. This approach can be used to treat a wide range of issues in various fields (archaeology, natural disaster reduction, climatology, epidemiology, ecology …) mainly thanks to significant methodological changes. The three levels Bayesian hierarchical models keep being used, however their
different components have deeply evolved. In this class of models, the first level models the data with a count distribution, the second one defines the structure of the mean of this distribution, and the third one sets all the model parameters. Our main aim was to determine if the disease mapping framework suits infectious diseases and how the contagion influences the results. In fact we expected that infectious pathologies would imply some overdispersion of data and a different structure of the underlying risk. Thus such an issue could need specific methodological choices.

2. Background and Methods

2.1. Bayesian framework and prior information

The Bayesian framework is known for its relevance to deal with very few data and take into account the prior knowledge concerning the phenomenon of interest. Most of the studies in the literature put their attention on an only disease. Thus its characteristics can be used to build a well fitted model. In particular population and environmental variables can be incorporated into the model. Thus if the pathology differently affects certain population groups, its incidence in the studied area can depend on the structure and the repartition of the population. In this case some authors standardize the population in order to define the structure of the remaining heterogeneity4, 5, 6, 7. If we want to test the impact of environmental factors, they can be taken into account at the second level5, 8, 9. In the literature lots of studies use the disease mapping approach to determine which ones are the most relevant to explain the structure of the risk2. The hidden structure of data can be seen as the expression of unknown (population or environmental) cofactors.

2.2. Spatial dependencies and extension to the temporal dimension

All the factors explaining the outcome of disease cases can not be identified. Thus, in every disease mapping study, spatial dependencies still exist between data of neighbouring regions2. One of the main aims of disease mapping is to determine the most relevant processes and functions to model these spatial correlations. Many approaches have been proposed to consider these spatial dependencies: different conditional autoregressive processes10, 11, 12, 13, spatial trends14 or classification of the studied area15. The spatial framework of disease mapping has naturally extended to the spatiotemporal context14,16. This implies significant methodological improvements, especially to model the structure of the risk and to interpret the results. Different methods were considered at the second level of the Bayesian hierarchical model: a linear trend14, 17, polynomial functions18, random walks19, 20, 21, spatiotemporal conditional autoregressive processes6, 20, splines22...
this case, a Gaussian term whom expectation is the mean of the neighbouring values is added in the decomposition of the logarithm of the relative risk. Some information concerning the pathology of interest, such as for instance the evolution trend and the spreading of the disease, can be taken into using linear, polynomial and spline functions. A Gaussian white noise is usually added to treat the individual heterogeneity. In our study, we do not focus on an only phenomenon, thus we do not want to consider a prior knowledge in our models. In the context of infectious diseases, primary cases may imply secondary ones. These secondary cases can occur in the same geographical area and at the same period, or in a close neighbourhood. Thus we assume that the contagion may imply a reinforcement of the structure of the risk, we have tested the relevance of every possible combination of spatial, temporal, and spatiotemporal Conditional Autoregressive processes, and of Gaussian white noise. We are interested in the weights introduces by Cressie to quantify the contribution of each component in the structure of the data26.

2.5. Third level of the model

The third level sets all the parameters of both the first and the second levels. The aim is to make them as poorly influent in the resulting probabilities. Non-informative distributions such as for instance the infinite uniform distribution are generally not used as they imply convergence defects. Thus poorly informative distributions are considered in the literature. In most of the studies, the Gamma distribution is used to model precision parameters (inverse of variance parameters). We performed our analyses using this distribution too.

2.6. Model selection method

In most of the disease mapping studies, and in all the spatiotemporal ones, the Deviance Information Criterion (DIC) is used to select the most relevant27, 28. The DIC is a generalization of the Bayesian Information Criterion and the Akaike information Criterion. It is a compromise between fit and parsimony. Indeed the fit of the model is penalized by the effective number of parameters. The lower is the value of DIC, the better is the associated model. We used this criterion in our study to determine the most appropriate model associated with each data set.

2.7. Computational aspects

The OpenBUGS software was used to perform the Bayesian inference. Jobs were parallelized on different processors thanks to HTCondor engine. Processing results was performed by programs in C. The R software was used for the preparation of the analyses and for the exploitation of results.

3. Results

We modeled simulated data and real data (bovine tuberculosis cases). The study of simulated outbreaks can identify the most relevant models under different experimental conditions. The analysis of real data allows verifying the congruence of the method with the pre-existing knowledge concerning bovine tuberculosis.

For both types of analyses, the Metropolitan France (excluding Corsica) is the geographical area of interest. The territory is divided into 448 small hexagons (40km) in which cases are simulated (simulated data) or counted (real data). The considered period is between 2001 and 2010. The studied population consists in cattle farms.

3.1. Simulated outbreaks

In order to characterize the method and to evaluate the most efficient models, we analyzed datasets which contain simulated outbreaks. More precisely, we first simulated a uniform relative risk in the whole studied area. We then introduced three high-risk areas: the first is constant over both time and space, the second one moves linearly over time from North-West to South-East and the last one increases and then decreases in intensity over time. We also considered two different risk levels. We finally modeled the cases by two different count distributions: the negative binomial distribution (with overdispersion, which often occurs in the context of infectious diseases) and the Poisson distribution (without overdispersion).
The calculations showed that the models with a negative binomial distribution at the first level fit overdispersed data well. Such models also appeared as relevant when the risk is highly contrasted throughout the studied area, even for Poisson simulations. Also the highest ranked models (according to the DIC) included both spatial and temporal dimensions in their risk structure. The Gaussian noise was in some cases relevant to describe a part of the heterogeneity, for both binomial negative and Poisson simulations and analyses. The weight parameters suggested in Cressie’s publications dramatically penalized the models; they are not relevant to fit the simulated data. It may be due to how data were simulated; we wanted to give an equivalent importance to temporal and spatial dimensions, so weight coefficients might not be relevant in this context.

We also noticed that constant outbreaks were more easily detected. Moving outbreaks were more hardly identified. It may be due to the repartition of the population. Outbreaks whose intensity increases and then decreases were recognized but disappeared when they became too weak.

3.2. Bovine tuberculosis cases

The data concern bovine tuberculosis cases which have been diagnosed in France between 2001 and 2010. A « case » is defined as a farm affected by bovine tuberculosis. All the cases are counted in each hexagon and for each year. Models defined as the best (according to the DIC) have congruent characteristics. First they use the negative binomial distribution at the first level. It is coherent with the high overdispersion observed in the data. It may be due to the very high frequency of null values, and the occurrence of high values. Such a repartition is congruent with the contagion; indeed a case and the following cases can occur in the same hexagon and at the same period. Furthermore, the most suitable models have at least two components in the structure of the risk (spatial, temporal and/or spatiotemporal). It shows the relevance to consider both the spatial and temporal dimensions. Furthermore the weights are considered relevant. They give a much higher importance to spatial and spatiotemporal components (> 1 for both) than to time processes (<0.1). The negative binomial distribution and the CAR processes considered succeeded in explaining most of the heterogeneity of the data; indeed, the models with a Gaussian white noise tend to be penalized.

Areas identified as being at risk are consistent with the epidemiological knowledge on the subject. The center of France was concerned by this pathology, especially in the neighborhood of two departments: the Cote-d'Or and the Dordogne. It is congruent with the conclusions of the institutes in charge of the sanitary surveillance. The Auvergne was surrounded by areas at risk; nevertheless this region is not really affected by bovine tuberculosis. It may be explained by the predominance of the extensive agriculture. The Bretagne was nearly unscathed. It is an important point as this region is the biggest producer of cattle in the country. The South East of France is identified as at risk by our study, although this region is not known as particularly concerned by the bovine tuberculosis. This phenomenon can be explained by the occurrence of some cases in the Camargue, by the low population of cattle farms, or perhaps by a side effect.

4. Conclusion

Disease mapping applied in a spatiotemporal framework has succeeded in highlighting the areas which present a high relative risk concerning bovine tuberculosis. Our analysis is congruent with the literature dealing with this the repartition of bovine tuberculosis in France between 2001 and 2010. Indeed the Dordogne and the Côte d’Or are the regions which are the most concerned by this pathology.

Both real and simulated data confirmed the relevance of the negative binomial distribution at the first level of the hierarchical model, in the case of highly contrasted risk values throughout the territory and in the case of overdispersed data. The Poisson distribution better fits which are weakly variable and not overdispersed. In the context of infectious diseases, cases data are overdispersed and present highly contrasted levels of risk, thus the binomial negative distribution appears as a relevant way to investigate contagious pathologies.
The approach of Cressie, who introduced weights to quantify the impact of each term in the structure of the risk, appeared relevant to model bovine tuberculosis data but not to analyze simulated outbreaks. It may be due to the structure of simulated data. Thus further investigations would give significant clues to understand this phenomenon.

This study also showed the interest of considering the temporal dimension in risk mapping. In fact all the models identified as the most efficient (according to the DIC) for each data set include both the spatial and temporal dimensions. In the case of contagious diseases, the spatial, temporal and spatiotemporal components of the risk play a predominant role to explain the observed heterogeneity. The disease mapping approach was classically dedicated to pathologies which are influenced by environmental factors. This study showed that Bayesian disease mapping is also relevant to model contagious data. However the assumption of low incidence remains necessary.

Acknowledgements

The authors would like to thank the computer scientists of the animal epidemiology unit of the INRA (Agronomy French Research Institute) center based in Clermont-Ferrand/Theix for their assistance in the implementation of the calculations and for the network maintenance.

The authors thank Christian Ducrot for the information he provided concerning bovine tuberculosis.

References