International Journal of Applied Earth Observation and Geoinformation 22 (2013) 99–105

Contents lists available at SciVerse ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag

On the difficulty to delimit disease risk hot spots

M. Charras-Garrido[a,*], L. Azizi[a,b], F. Forbes[b], S. Doyle[b], N. Peyrard[c], D. Abrial[a]

[a] INRA, UR 346, F-63122 Saint-Genès-Champanelle, France
[b] INRIA, Équipe Mistis, F-38334 Saint-Ismier, France
[c] INRA, UR 875, F-31326 Castanet Tolosan, France

ARTICLE INFO

Article history:
Received 11 July 2011
Accepted 11 April 2012

Keywords:
Classification
Disease mapping
Epidemiology
Generalized Potts model
Spatial clustering
Hidden Markov random field

ABSTRACT

Representing the health state of a region is a helpful tool to highlight spatial heterogeneity and localize high risk areas. For ease of interpretation and to determine where to apply control procedures, we need to clearly identify and delineate homogeneous regions in terms of disease risk, and in particular disease risk hot spots. However, even if practical purposes require the delineation of different risk classes, such a classification does not correspond to a reality and is thus difficult to estimate. Working with grouped data, a first natural choice is to apply disease mapping models. We apply a usual disease mapping model, producing continuous estimations of the risks that requires a post-processing classification step to obtain clearly delimited risk zones. We also apply a risk partition model that build a classification of the risk levels in a one step procedure. Working with point data, we will focus on the scan statistic clustering method. We illustrate our article with a real example concerning the bovin spongiform encephalopathy (BSE) an animal disease whose zones at risk are well known by the epidemiologists. We show that in this difficult case of a rare disease and a very heterogeneous population, the different methods provide risk zones that are globally coherent. But, related to the dichotomy between the need and the reality, the exact delimitation of the risk zones, as well as the corresponding estimated risks are quite different.

1. Introduction

Efficient disease control requires correct understanding of the determinants and dynamics of the disease. The first questions to ask are: Where are the high risk populations located? Are these locations structured in space? If so, how? Therefore, the analysis of the geographical variations of a disease and their cartographical representation is an important step in epidemiology. Representing the health state of a region offers interesting insights into the mechanism underlying the spread of a disease. It allows to highlight spatial heterogeneity, localize high risk areas (*i.e.* important contaminations) and identify potential sources of a disease. To go further and help to determine protection measures, we need to clearly identify and delineate homogeneous regions in terms of disease risk, and in particular disease risk hot spots.

In this article, we will illustrate and comment our purpose with the example of bovine spongiform encephalopathy (BSE) in France between July 2001 and December 2005. This sudden, non contagious and unexpected disease (see Anderson et al., 1996; Ducrot et al., 2008) threatened bovine production in Europe and has been intensively studied (for spatial analyses, see *e.g.* Abrial et al., 2005; Allepuz et al., 2007 or Paul et al., 2007). To guarantee confidentiality, the exact localization of the cases are not available. Thus, the territory of France is divided into $n = 1264$ hexagons of 23 km width, in which cases and population are counted (see Fig. 1(a) and (b)). In our BSE example, as in most of applications, the different zones at risk we want to determine do not correspond to an underlying reality, but are only needed by the epidemiologists for ease of interpretation, and to decide where to apply control procedures. As we will illustrate it with our BSE example, this difference between our requirements and the reality may imply estimation difficulties. Moreover, this example has been chosen to compare the behavior of the different methods in a challenging scenario with very low risk values, small numbers of observed cases and population sizes that increase the estimation difficulties.

Since we work with grouped data, a natural choice is to apply disease mapping models, such as presented in Section 2. In Section 2.1, we present one of the most commonly used disease mapping models, producing a continuous estimation of the risk that does not clearly delimit zones of different disease risk. However, we can apply a post-processing classification step that delineate different

Abbreviations: BSE, bovin spongiform encephalopathy; BYM model, model of Besag, York and Mollié; CAR, conditionally auto-regressive; EM algorithm, expectation–maximization algorithm; MCEM algorithm, Monte-Carlo EM algorithm.

* Corresponding author. Tel.: +33 4 73 62 40 65; fax: +33 4 73 62 45 48.
E-mail addresses: Myriam.Charras-Garrido@clermont.inra.fr (M. Charras-Garrido), lamiae.azizi@inria.fr (L. Azizi), florence.forbes@inria.fr (F. Forbes), senan.doyle@inria.fr (S. Doyle), Nathalie.Peyrard@toulouse.inra.fr (N. Peyrard), David.Abrial@clermont.inra.fr (D. Abrial).

0303-2434/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jag.2012.04.005

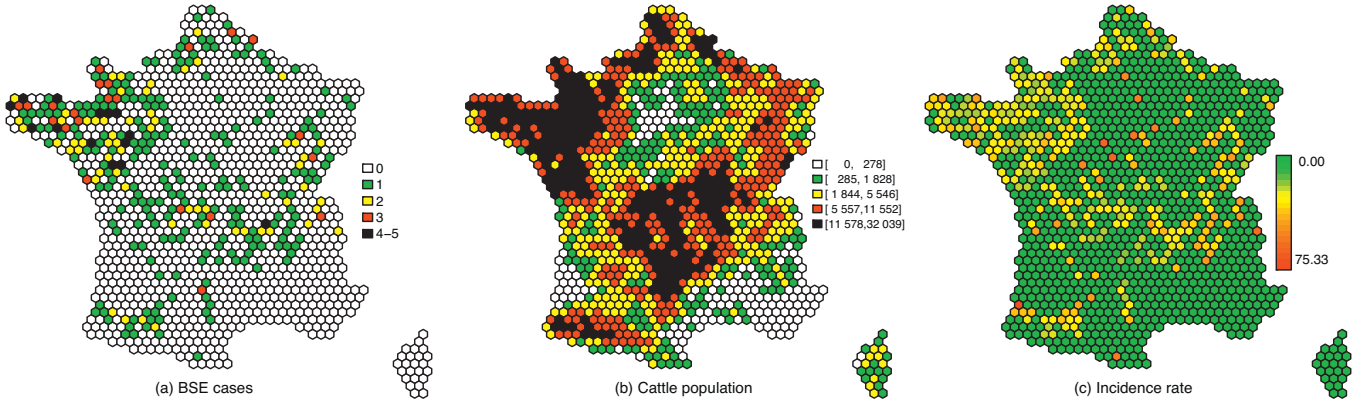

Fig. 1. Real data set: BSE in France. (a) Number of cases for the study period, (b) cattle population map, and (c) simple estimation of the risk: standardized incidence rate.

zones in the map. The risk partition model presented in Section 2.2 build a classification of the risk levels in a one step procedure. Although our data are aggregated, we can also consider them as point data. Section 3 present one of the most used clustering method for point data based on the scan statistic. In Section 4, we illustrate the performance of these different methods in determining hot spots for the BSE risk in France. A discussion ends the paper in Section 5.

2. Disease mapping models

As in our example, epidemiological data are frequently aggregated count data: for each unit i ($i \in S = \{1, \ldots, n\}$) observed cases of a given disease are counted (y_i) and compared to the population size (n_i) in this area. We denote by Y_i the random variable associated with y_i. A natural simple estimation of the risk is the common maximum likelihood estimate computed independently in each unit: the incidence rate. The absolute epidemiological risk θ_i, the probability that an individual in $i \in S$ is contaminated by the disease, is estimated by the raw incidence rate $p_i = y_i/n_i$. The relative risk r_i measures the departure of the local risk from the empirical mean risk over the whole spatial area. It is estimated by the standardized incidence rate $\rho_i = y_i/e_i$, where $e_i = n_i p$ is the expected number of cases for an homogeneous risk $p = (\sum_{i=1}^{n} y_i)/(\sum_{i=1}^{n} n_i)$. These estimations (see ρ_i for the BSE example in Fig. 1(c)) produce noisy maps difficult to interpret with over dispersion (isolated high risk values) and very extreme values of the risk (many have either null values or estimated risks that are more than 70 times higher than the mean overall risk). It is therefore clear that spatial dependencies have to be taken into account when analyzing such location dependent data, in order to produce smoothed maps.

Most statistical methods for risk mapping of aggregated data dedicated to non contagious diseases, are based on a Poisson log-linear mixed model (see *e.g.* Mollié, 1999; Pascutto et al., 2000 or Lawson et al., 2000). The model proposed by Besag et al. (1991) (or BYM model) presented in Section 2.1 is one of the most popular approaches, but inference results in a *real-valued estimation* of the risk at each location, see Fig. 2(a).

One of the main reported limitations (*e.g.* by Green and Richardson (2002)) is that local discontinuities in the risk field are not modeled leading to potentially over-smoothed risk maps. Also, in some cases, as in animal epidemiology (see *e.g.* Abrial et al., 2005), a coarser spatial representation of risk is needed in which locations with similar risk values are grouped. Section 2.2 is then devoted to risk partition models.

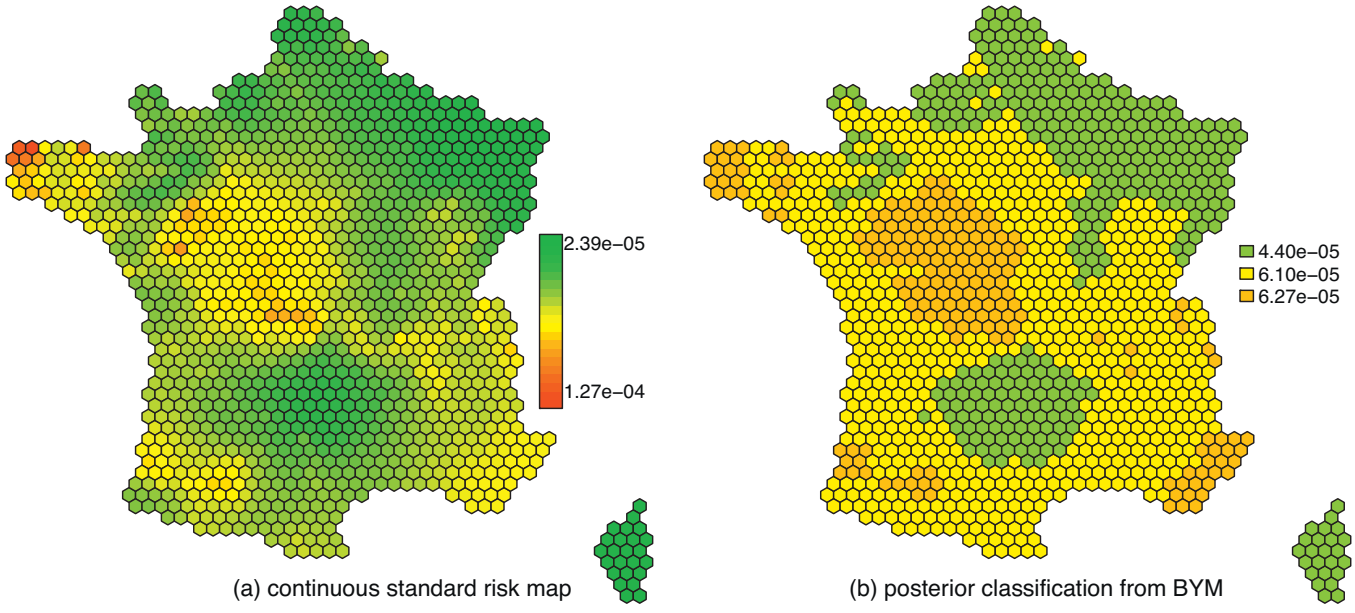

Fig. 2. Standard disease mapping model (BYM) applied to BSE data: (a) a continuous estimation of the risks and (b) a posterior classification.