# On the difficulty to clearly identify and delineate 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 Epidémiologie Animale, F-63122 Saint-Genès-Champanelle, France*
[b]*INRIA, Équipe Mistis, F-38334 Saint-Ismier, France*
[c]*INRA, UR 875 Biométrie et Intelligence Artificielle, F-31326 Castanet Tolosan, France*

## 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, the 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

[*]Corresponding Author. Unité d'Epidémiologie Animale, Département de santé animale Centre INRA de Clermont-Ferrand-Theix, 63122 Saint-Genès-Champanelle, France. tel : (+33) 4.73.62.40.65, fax : (+33) 4.73.62.45.48

*Email 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)

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)[1] 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.

*Keywords:* Classification, Disease Mapping, Epidemiology, Generalized Potts Model, Spatial clustering, Hidden Markov Random Field,

## 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

[1] Abbreviations used in the article: Bovin Spongiform Encephalopathy (BSE); Model of Besag, York and Mollié (BYM model); Conditionally Auto-Regressive (CAR); Expectation-Maximization algorithm (EM algorithm); Monte-Carlo EM algorithm (MCEM algorithm)

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 Figure 1(a) and (b)). In our BSE example, as in most of applications, the different zones at risk we want to determinate do not correspond to an underlying reality, but are only needed by the epidemiologists for ease of interpretation, and to determine 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.

INSERT FIG 1 ABOUT HERE

Figure 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. (available in color online)

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 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 $i$ ($i \in S = \{1, \dots, n\}$) 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 $\theta_i$, 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 $\rho_i$ for the BSE example in Figure 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, York and Mollié (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 Figure 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.

INSERT FIG 2 ABOUT HERE

Figure 2: Standard disease mapping model (BYM) applied to BSE data: a continuous estimation of the risks. (available in color online)

## 2.1. *Standard disease mapping models*

The expected number of cases $e_i$ (for $i \in S$) are assumed to be known and constant during the study period. Since disease mapping is generally applied to rare and non contagious diseases such as cancer, the number of observed cases are usually modeled by Poisson distributions: $Y_i \sim Poisson(e_i r_i)$ for $i \in S$, where $r_i$ is the unknown relative disease risk (see Besag et al. (1991)). This first level accounts for the local variability, *i.e.* the intra-area variability. A second level accounts for the spatial dependence between areas through a log-linear mixed model. The spatial variation in log-relative risk $u_i = \log(r_i)$ is modeled with a Gaussian *intrinsic* Conditionally Auto-Regressive (CAR) Markov random field prior $(U_i)_{i\in S}$ with distribution $U_i|U_j, j \neq i \sim \mathcal{N}(\bar{U}_i, \sigma^2/m_i)$, where $\bar{U}_i = \sum_{j\neq i} w_{ij}U_j/m_i$. The weight matrix $\{w_{ij} : i, j \in S\}$ generally corresponds to a 0-1 neighboring in which $w_{ij} = 1$ for neighbors and 0 otherwise (usually areas that share a common boundary are defined as neighbors). In this case, $m_i = \sum_{j=1}^{n} w_{ij}$ is the number of neighbors of area $i$. The amount of smoothing in the random effects $U_i$, is controlled by the unknown precision parameter $\tau = 1/\sigma^2$: a small value induces little smoothing, while an infinite value forces all $u_i$'s (so all the risks) to be equal. The log-linear mixed model can also incorporate, at this second level, the effect of covariates and a random effect accounting for unstructured heterogeneity, see *e.g.* Pascutto et al. (2000) or Lawson et al. (2000). At a third level, a Gamma non-informative prior is usually adopted for $\tau$. Estimation of the posterior distribution of the relative risks is achieved by sampling from posterior distributions using Markov chain Monte Carlo methods.

The BYM model has been widely used and extended for disease mapping. In particular, the intrinsic CAR being an improper prior, two variants have been proposed (see MacNab (2010)): the proper CAR prior (see *e.g.* Lagazio et al. (2003)) and the Leroux et al. (2000) CAR prior (see also *e.g.* Ugarte et al. (2009)). Recent developments in risk mapping concern spatio-temporal mapping (see *e.g.* Knorr-Held and Richardson (2003) or Robertson et al. (2010)) and multivariate disease mapping (see *e.g.* Knorr-Held et al. (2002) or MacNab (2010)).

### *2.2. Risk partition models*

Grouped representations have the advantage of providing clearly delimited areas for different risk levels. This is helpful for decision-makers to interpret the risk structure and determine protection measures such as culling, movement restriction, mass vaccination, *etc.* These areas at risk can be considered as clusters (see Knorr-Held and Rasser (2000)), but we prefer the interpretation as risk classes, since geographically separated areas can have similar risks and be grouped in the same risk class. Using the traditional BYM model, risk classes are commonly determined with an additional *post-processing classification step* (see *e.g.* Abrial et al. (2005) or Allepuz et al. (2007)), as in Figure 2(b) for the BSE example. Including the risk partition in the risk estimation procedure, the goal is to assign, to each geographical unit, one of $K$ possible risk levels (*e.g.* the absolute risk levels) in $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_K\}$. These risk levels are themselves unknown and need to be estimated. Therefore, the data is naturally divided into observed variables $\mathbf{Y} = \{Y_1, \dots, Y_n\}$ (numbers of cases) and unobserved or hidden variables $\mathbf{Z} = \{Z_1, \dots, Z_n\}$, so that $Z_i = k$ when region $i$ is assigned to risk level $\theta_k$. As already mentioned, in the case of a non contagious and rare disease, a common assumption (see Section 2.1) is that for an area $i \in S$, $Y_i$ comes from a $Poisson(n_i \theta_{Z_i})$ with the following class dependent distributions: $P(Y_i = y_i | Z_i = k, \boldsymbol{\theta}) = \exp(-n_i \ \theta_k)(n_i \ \theta_k)^{y_i} / y_i!$. Concerning the hidden field $\mathbf{Z}$, which encodes the spatial correlation characterizing disease maps, the dependencies between neighboring $Z_i$'s are modeled by assuming that the joint distribution of $\mathbf{Z}$ is a discrete Markov random field on the graph connecting different locations (as is usual in Hidden Markov random fields, we restrict the neighborhood to pair-wise interactions to facilitate computation and interpretation):

$$P(\mathbf{Z} = \mathbf{z}|\boldsymbol{\alpha}, I\!\!B) \propto \exp\left(\sum_{i \in S} \alpha_{z_i} + \sum_{i \in S} \sum_{j \in V_i} I\!\!B_{z_i, z_j}\right), \tag{1}$$

where $V_i$ is the set of neighbors of $i$. Here adjacent regions $i$ and $j$ are defined as neighbors. Parameter $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]^T$ is a $K$-dimensional vector which acts as a weight for the different values of $z_i$. $I\!\!B$ is a $K \times K$ symmetric matrix which encodes spatial interactions between the different classes.

If, in addition to a null $\boldsymbol{\alpha}$, $I\!\!B = b\, I_K$ where $b$ is a real positive value and $I_K$ the $K \times K$ identity matrix, we get the *Potts model* commonly used for image segmentation. Note that this standard Potts model used by Green and Richardson (2002) and Alfó et al. (2009) for classified disease mapping is often suited for clustering tasks since it tends to favor neighbors that are in the same class (*i.e* have the same risk level). However this model penalizes two neighbors that have different risk levels with the same penalty whatever the difference between these risk levels. Nevertheless, for disease mapping, it is unlikely to observe a very low risk area just next to a very high risk area. Thus, $I\!\!B$ matrices that are different from $b\, I_K$ have been proposed to encode higher penalties when the risk levels are very different, and produce smooth gradation of the risks. Assuming an ordering of the classes, in the sense that $\theta_k < \theta_{k+1}$, new $I\!\!B$s' penalize pairs of classes $k$ and $l$ at neighboring sites according to the distance between the two classes: the closer the two classes, the higher the probability of observing this configuration. The *grad*-1 $I\!\!B$ proposed in Charras-Garrido et al. (2011) is based on the absolute distance: $I\!\!B_{k,l} = b(1 - |k - l|/(K - 1))$. The *grad-2-neg* $I\!\!B$ is based on the squared distance: $I\!\!B_{k,l} = b(1-(k-l)^2/(K-1))$ and is a variant of the *grad*-2 described in Charras-Garrido et al. (2011). For $K \geq 3$, it has the particularity that some terms in the $I\!\!B$ matrix are negative. In these two cases, the probability of observing two given classes in neighboring areas always decreases with the distance between these two classes. The *semi-grad* $I\!\!B$ proposed in Azizi et al. (2011) consists of a matrix with three non zero diagonals: $I\!\!B_{k,k} = b$ for all $k = 1, \ldots, K$, $I\!\!B_{k,l} = b/2$ when $|k - l| = 1$ and $I\!\!B_{k,l} = 0$ otherwise. This last one can be seen either as a simplification of the *grad*-1 model in which all pairs of risk classes far apart are equally weighted, or as a variant of the Potts model where the zero penalty is pushed a unit further. In these four matrices, parameter $b$ can be interpreted as the strength of interaction between neighbors. The higher $b$, the more weight is given to the neighbors' influence. Here, the correlation only depends on the distance between the risk classes of two adjacent units, but other $I\!\!B$ definitions are quite easy to construct and to interpret depending on the targeted application.

For an epidemiologist, the first output of interest is the risk map, *i.e.* the values of the risk class assignments $\mathbf{Z}$, followed by the values of the risk $\boldsymbol{\theta}$. To recover $\mathbf{z}$, a Maximum Posterior Marginal (MPM) principle consisting of assigning each region $i$ to the class $k$ that maximizes $P(Z_i = k|\mathbf{y}, \boldsymbol{\Psi})$ with $\boldsymbol{\Psi} = (\boldsymbol{\theta}, \boldsymbol{\alpha}, b)$ is considered. In order to perform these maximizations, we have to estimate the parameters $\boldsymbol{\Psi}$ which are usually unknown: the risks $\boldsymbol{\theta}$, the classes weights $\boldsymbol{\alpha}$ and the smoothing strength $b$. For estimation, the Expectation-Maximization (EM) algorithm is applied (Dempster et al. (1977)) using two approximations: the Monte-Carlo EM (MCEM) algorithm proposed by Wei and Tanner (1990) as in Charras-Garrido et al. (2011) and a Mean Field approximation of EM presented in Celeux et al. (2003) as in

Azizi et al. (2011).

## 3. Clustering methods for point data

Our data are aggregated and a natural choice is to apply disease mapping models. Although, for comparison purposes, we can also consider our data as point data and apply clustering methods. Let $X_1, \ldots, X_M$ be random variables that denote the spatial coordinates of $M$ observed events. The objective of clustering methods for point data is to identify, if they exist, the zones in which the concentration of events is abnormally high, usually named clusters. To assess the significance of a supposed cluster, the observed concentration is usually compared with the concentration observed under the null hypothesis $H_0$ that the events are sampled independently from the underlying population density, generally a Poisson distribution in epidemiology.

As the monitoring tools get improved and the computational capacities increase, the methods for individual data become more and more applicable. Existing spatial methods are often derived from one-dimensional cluster detection methods, which are mainly applied to temporal point processes. For example, the techniques introduced by Kelsall and Diggle (1995a,b) are based on a kernel intensity estimation of the events process. The method of Besag and Newell (1991) is based on the $k$ nearest neighbor and the $p$-value of observing $k$ or more cases within the neighboring area is computed by a Poisson probability given the population at risk in the area. The analysis is repeated for different values of $k$ and a cluster is detected if its statistical significance (with a correction of the $p$-value) persisted over three values of $k$. As noticed for example by d'Aignaux et al. (2002), this method may provide false clusters because of multiple testing.

As in the temporal setting, the most popular approach is the scan statistic adapted to the spatial setting by Kulldorff (1997). It relies on the generalized likelihood-ratio test statistic of $H_0$ against a piecewise constant density alternative. To apply this method, one needs to set the family of the possible clusters, for example all the discs centered on a point of a predefined grid. The radius of each circle is generally set to vary continuously from zero to an upper limit (*e.g.* less than 50% of the total area). This predefined shape of the cluster can be an important limitation since, in the real world, an excess of events may be recorded along a river for example. An alternative has been proposed recently: Kulldorff et al. (2006) investigated a wide family of elliptic windows with predetermined shape, angle and center. The ultimate solution would be to consider all the convex envelopes including any subset of the events locations. However, this becomes computationally infeasible when the number of events is large. The statistical significance of the largest likelihood (for positive clusters or the lowest likelihood for negative clusters) is assessed by determining its distribution under the null hypothesis through Monte Carlo simulation.

## 4. Results on the BSE data set

BSE is a non contagious neurodegenerative disease in cattle. It is transmitted by meat and bone meal. Since there is no direct transmission and no vector, the spatial analysis are important to understand and explain the geographical localization of the cases. The cases in our data set occurred between July 1, 2001 and December 31, 2005, although at that time the meat and bone meal had been already forbidden for cattle in France.

We first compare the results obtained with the risk partition model, for the four $I\!B$ matrices presented in Section 2.2, and two estimation methods: MCEM and Mean-Field EM. We use approximations of the Bayesian Information Criterion (BIC) to determine the number $K$ of classes: that proposed by Stanford and Raftery (2002) when using the MCEM algorithm, and that proposed by Forbes and Peyrard (2003) when applying Mean Field EM. The selected number $K$, the corresponding BIC values and the $\boldsymbol{\alpha}$ and $b$ estimations are presented in Table 1. For the MCEM algorithm, $K = 2$ is chosen for all $I\!B$ matrices. Note that when $K = 2$ all the models are identical and $I\!B = bI_K$. The resulting map is shown in Figure 3(a). For the Mean Field EM algorithm, $K = 3$ is chosen for all $I\!B$ matrices. The resulting maps are shown in Figure 3(b) for *grad*-1 and *semi-grad*, which are identical in this case, in Figure 3(c) for *grad*-2-*neg* and in Figure 3(d) for Potts. To facilitate the maps comparison in Figure 3, the estimated risk values, which differ from one map to another, are associated to colors going continuously from green (lowest risks) to red (highest risks) through yellow (medium risks).

The $b$ values (Table 1) are higher for Mean Field EM with lower (*i.e* better) BIC values too. Since for $K \geq 3$, the *grad*-2-*neg* $I\!B$ matrix has negative terms, for large $b$ values, neighboring locations belonging to the same risk class are strongly favored while they are strongly penalized if they correspond to distant risk levels. In terms of risk mapping, the more likely maps should then minimize the common borders between the lowest and the highest risk classes. However, this is not what is observed in Figure 3(c) because this interaction effect is compensated by the fact that the middle

| Algorithm | MCEM | Mean Field EM | | |
|---|---|---|---|---|
| $\mathbb{B}$ model | all models | *grad*-1 | *grad*-2-*neg* | Potts |
| $K$ | 2 | 3 | 3 | 3 |
| Parameters | 4 | 6 | 6 | 6 |
| Loglikelihood | $-787$ | $-788$ | $-788$ | $-785$ |
| BIC | 1592 | 1589 | 1589 | 1586 |
| $\boldsymbol{\alpha}$ | $(0, 0)$ | $(0, 0.3, 1)$ | $(0, -147.3, -0.8)$ | $(0, 1.2, 1.16)$ |
| $b$ | 0.68 | 2.32 | 15.4 | 8.15 |

Table 1: BSE data set: Number of free parameters, Selected number of classes $K$ using BIC, log-likelihood and BIC values, $\boldsymbol{\alpha}$ and $b$ estimations using the different $\mathbb{B}$ models and implementations.

risk level has a much smaller $\boldsymbol{\alpha}$ weight than the two others. Indeed, when $K = 3$, the middle risk level is the only alternative to prevent the penalized borders to occur. The interpretation of the parameter values for the *grad*-2-*neg* model when $K \geq 3$ is more complex and does not correspond to the same intuition as the other Potts variants which model only positive interactions.

INSERT FIG 3 ABOUT HERE

Figure 3: BSE data set: Different estimated risk maps with the risk partition model; (a) all models using MCEM; (b) *grad*-1 & *semi-grad*, (c) *grad*-2-*neg*, and (d) Potts model using Mean Field EM. To facilitate the maps comparison, the estimated risk values, which differ from one map to another, are associated to colors going continuously from green (lowest risks) to red (highest risks) through yellow (medium risks). (available in color online)

The maps presented in Figure 3 globally retrieve the three known zones at risk located in the Brittany (West), the Center, the Alps (East) and the South-West, corresponding to the regions expected by the experts and highlighted in previous works (see Abrial et al. (2005)). Indeed, it is suspected that the BSE risk can be explained by a cross-contamination with an ingredient used in poultry or pig feed, and in these regions there is a high density of monogastric species (Abrial et al. (2005)), e.g. pigs and poultry, and meat and bone meal were used to feed these species Paul et al. (2007). With all models and MCEM (Figure 3(a)), and with *grad*-1 (*i.e.* also *semi-grad* since $K = 3$) and Mean Field EM (Figure 3(b)), the different zones at risk are particularly well separated. The maps obtained with *grad*-2-*neg* and Mean Field EM (Figure 3(c)), and with Potts and Mean Field EM (Figure 3(d)) are similar regarding the lowest risk region going from the North-East to the South-Center. Potts and Mean Field EM (Figure 3(d)) present misleading risks in the South-East, where there is little population and no observed cases, as well as in Corsica. This island, with only a few units has no neighborhood with the rest of France, which limits spatial regularization. The *grad*-2-*neg* and Mean Field EM (Figure 3(c)) present a low risk in the South-East, and, in contrast to the other maps (except the BYM model, see Figure 2), an unimportant risk in the North. With the BYM model (see Figure 2), the known zones at risk are also retrieved, but a counterintuitive risk is recovered in some regions with low population and no observed cases, in particular in the South-East. Also this model produces estimated risk values so close to each other that BIC suggests only one class. These close risk values and their smoothness explain why in all the presented maps there are few risk classes and why the classifications are quite different since it is difficult to determine a cutting point. In particular in Figure 2(b), where we choose $K = 3$ to be consistent with most of the maps of Figure 3, the two highest risks are almost equal and the corresponding classes are difficult to distinguish.

Although globally coherent, these different maps does not delineate exactly the same zones. From a modeling viewpoint, classifying risk values into a finite number of levels and homogeneous areas may not correspond to the underlying reality in which risk values are more likely to vary smoothly across the different geographical units. However, as already mentioned, the need for such a classification is expressed by epidemiologists as it helps interpretation and decision-making. The main issue in rare disease cases is that counts are relatively small with a large number of zero values. It may follow that spatial information is only mildly supported by the observed data. Estimated risks are very small and close to each other, making the separation into different regions difficult and somewhat unstable. Each method actually define its own risk classes, which feature specific boundaries. Indeed the risk values associated to each class are different from one model to another. In this case, we suspect the BYM model to produce smooth maps that essentially reflect the CAR prior while the risk partition models hesitate between various solutions (as illustrated in Figure 3) that are not so different in terms of likelihood or quality of fit to the observed data. Aware of this issue, in the case of the Mean Field EM implementation, a number of initializations (see Azizi et al. (2011)) have been carried out and the highest likelihood result, displayed in Figure 3(b), have been selected. The MCEM implementation should be less sensitive to initialization but convergence is more difficult to diagnose. The map of Figure 3(b) can therefore be considered as our reference result. As already observed, important zones are clearly and accurately detected there. In the other maps of Figures 3, the same features are recovered with a different precision. The two classes of Figure 3(a) roughly correspond to the fusion of two of the regions in (b) with more discontinuous but geographically close borders. For the other maps, important borders are missing either in the South-East or between South and North in the West. For the Potts model maps (Figure 3 (c)), the issue is not only missing borders but also irrelevant additional ones which seem to be prevented by the risk gradation modeling.

We then apply the spatial scan statistic (see Figure 4) detailed in Section 3 for circular clusters and for ellipsoidal clusters. With this method, among the four known regions at risk, only the Brittany (West) is retrieved as a cluster. The Center, the Alps (East) and the South-West, are not detected as positive clusters. Moreover, these zones considered as at risk are partly included in the negative clusters detected, *i.e.* highlighted as having a low BSE risk. This may be related to the fixed shapes of the clusters. With the same cause, we can also remark that each detected cluster include large zones that does not belong to the region under study, corresponding either to the sea or to adjacent countries. It is not satisfying to include such regions with no data in the idea to propose zones where to apply protection measures. Moreover, we can see on left panel of Figure 4 that positive and negative clusters can be superimposed, which is very counterintuitive. With this method, we can only identify risk hot spots but we have no measure of the risks associated to these zones, although it can also be an important information.

INSERT FIG 4 ABOUT HERE

Figure 4: BSE data set: clustering with the Kulldorff's scan statistics for circular clusters on right panel and ellipsoidal clusters on left panel.

## 5. Discussion

For practical purposes we are interested in the delineation of different zones at risk. As illustrated with the BSE real data set, the general pattern and the known zones at risk are globally retrieved by most of the tested methods. In particular, the risk partition models provide maps where risk zones are more clearly delimited, especially with *semi-grad* (or *grad*-1) and Mean Field EM. The models are flexible in that they can be easily adapted to different epidemiological situations and all parameters are easy to interpret. In particular, the interpretation of the $\mathbb{B}$ matrix in terms of neighborhood interaction allows users to design their own spatial smoothing. The BYM model produces maps that are too smooth: for a very rare disease in very heterogeneous population, such as in our BSE data set, it tends to estimate similar risks that are difficult to classify. Excessively high risks are also estimated in regions with very small populations. The risk partition models appear less sensitive to heterogeneous or very small populations. The clustering method for point data identify less zones than the other methods. Their shapes are fixed and are thus less interesting in the aim to precisely delineate regions where to apply control procedures. Moreover, this method does not associate a risk value to each highlighted region.

Risk zones facilitate interpretation of disease maps for epidemiologists, and they are needed with the aim to determine where to apply control procedures. Thus, such a classification is difficult to estimate since it does not correspond to an underlying reality of boundaries with a jump of the risk value. Thus, each method actually define its own risk classes, which feature specific boundaries. This probably explain most of differences that can be noticed among the estimated maps which illustrate our paper. However, even if differences are noticed in the exact delimitation of the regions and the corresponding estimated risk values, all classified risk maps globally highlight the same zones at risk. The maps obtained by the risk partition model appear to more correspond to the practical requirements, with more clear zones, and would preferably be used in this context.

## References

Abrial, D., Calavas, D., Jarrige, N., Ducrot, C., 2005. Poultry, pig and the risk of BSE following the feed ban in France - A spatial analysis. Vet. Res. 36, 615–628.

Alfó, M., Nieddu, L., Vicari, D., 2009. Finite mixture models for mapping spatially dependent disease counts. Biometrical J. 52, 84–97.

Allepuz, A., Lopez-Quilez, A., Forte, A., Fernandez, G., Casal, J., 2007. Spatial analysis of bovine spongiform encephalopathy in Galicia, Spain (2000-2005). Prev. Vet. Med. 79, 174–185.

Anderson, R., Donnelly, C., Ferguson, N., Woolhouse, M., Watt, C., Udy, H., MaWhinney, S., Dunstan, S., Southwood, T., Wilesmith, J., Ryan, J., Hoinville, L., Hillerton, J., Austin, A., Wells, G., 1996. Transmission dynamics and epidemiology of BSE in British cattle. Nature 382, 779–788.

Azizi, L., Forbes, F., Doyle, S., Charras-Garrido, M., Abrial, D., 2011. Spatial risk mapping for rare disease with hidden Markov fields and variational EM. research report inria-00577793. INRIA.

Besag, J., Newell, J., 1991. The detection of clusters in rare diseases. J. R. Statist. Soc. A 154, 143–155.

Besag, J., York, J., Mollié, A., 1991. Bayesian image restoration, with two applications in spatial statistics. Ann. I. Stat. Math. 43, 1–59.

Celeux, G., Forbes, F., Peyrard, N., 2003. EM procedures using mean field-like approximations for Markov model-based image segmentation. Pattern Recogn. 36, 131–144.

Charras-Garrido, M., Abrial, D., Peyrard, N., de Goër, J., Dachian, S., 2011. Classification method for disease risk mapping based on discrete hidden Markov random fields. Biostat. 13, 241–255.

d'Aignaux, H., Cousens, S.N. anhd Delasnerie-Lauprtre, N., Brandel, J., Salomon, D., Laplanche, J., Hauw, J., Alprovitch, A., 2002. Analysis of the geographical distribution of sporadic Creutzfeldt-Jakob disease in France between 1992 and 1998. Int. J. Epidemiol. 31, 490–495.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. B Met. 39, 1–38.

Ducrot, C., Arnold, M., de Koeijer, A., Heim, D., Calavas, D., 2008. Review on the epidemiology and dynamics of BSE epidemics. Vet. Res. 39:15.

Forbes, F., Peyrard, N., 2003. Hidden Markov model selection based on mean field like approximations. IEEE T. Pattern Anal. 25, 1089–1101.

Green, P., Richardson, S., 2002. Hidden Markov models and disease mapping. J. Am. Stat. Assoc. 97, 1055–1070.

Kelsall, J., Diggle, P., 1995a. Kernel estimation of relative risk. Bernoulli 1, 3–16.

Kelsall, J., Diggle, P., 1995b. Non-parametric estimation of spatial variation in relative risk. Stat. Med. 14, 2335–2342.

Knorr-Held, L., Rasser, G., 2000. Bayesian detection of clusters and discontinuities in disease maps. Biometrics 56, 13–21.

Knorr-Held, L., Rasser, G., Becker, N., 2002. Disease mapping of stage-specific cancer incidence data. Biometrics 58, 492–501.

Knorr-Held, L., Richardson, S., 2003. A hierarchical model for space-time surveillance data on meningococcal disease incidence. J. R. Stat. Soc. B 52, 169–183.

Kulldorff, M., 1997. A spatial scan statistic. Commun. Stat.-Theor. M. 26, 1481–1496.

Kulldorff, M., Huang, L., Pickle, L., Duczmal, L., 2006. An elliptic spatial scan statistic. Stat. Med. 25, 3929–3943.

Lagazio, C., Biggeri, A., Dreassi, E., 2003. Age-period-cohort models and disease mapping. Environmetrics 14, 475–490.

Lawson, A., Biggeri, A., Boehning, D., Lesaffre, E., Viel, J., Clark, A., Schlattmann, P., Divino, F., 2000. Disease mapping models: an empirical evaluation. Stat. Med. 19, 2217–2241.

Leroux, B., Lei, X., Breslow, N., 2000. Estimation of disease rates in small areas: A new mixed model for spatial dependence, in: Halloran, M., Berry, D. (Eds.), Statistical models in epidemiology, the environment and clinical trials. Springer-Verlag, pp. 135–78.

MacNab, Y., 2010. On gaussian Markov random fields and bayesian disease mapping. Stat. Methods Med. Res. 20, 49–68.

Mollié, A., 1999. Disease mapping and risk assessment for public health. Wiley. chapter Bayesian and Empirical Bayes approaches to disease mapping. pp. 15–29.

Pascutto, C., Wakefield, J., Best, N., Richardson, S., Bernardinelli, L., Staines, A., Elliott, P., 2000. Statistical issues in the analysis of disease mapping data. Stat. Med. 19, 2493–2519.

Paul, M., Abrial, D., Jarrige, N., Rican, S., Garrido, M., Calavas, D., Ducrot, C., 2007. Bovine spongiform encephalopathy and spatial analysis of the feed industry. Emerg. Infect. Dis. 13, 867–872.

Robertson, C., Nelson, T., MacNab, Y., Lawson, A., 2010. Review of methods for space-time disease surveillance. Spat. Spatio-temporal Epidemiol. 1, 105–116.

Stanford, D., Raftery, A., 2002. Approximate Bayes factors for image segmentation: The pseudolikelihood information criterion (PLIC). IEEE T. Pattern Anal. 24, 1517–1520.

Ugarte, M., Goicoa, T., Militino, A., 2009. Empirical bayes and fully bayes procedures to detect high-risk areas in disease mapping. Comput. Stat. Data An. 53, 2938–2949.

Wei, G., Tanner, M., 1990. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. J. Am. Stat. Assoc. 85, 699–704.