A spatial hazard model for cluster detection on continuous indicators of disease: application to somatic cell score

Emilie Gay, Rachid Senoussi, Jacques Barnouin

Vet. Res. 38 (2007) 585–596
© INRA, EDP Sciences, 2007
DOI: 10.1051/vetres:2007018

Available online at:
www.vetres.org

Original article

A spatial hazard model for cluster detection on continuous indicators of disease: application to somatic cell score

Emilie GAY[a,b]*, Rachid SENOUSSI[a], Jacques BARNOUIN[b]

[a] INRA, UR546, Biostatistique et processus spatiaux, Avignon, 84000, France
[b] INRA, UR346, Épidémiologie animale, Saint-Genès-Champanelle, 63122, France

(Received 28 August 2006; accepted 12 February 2007)

Abstract – Methods for spatial cluster detection dealing with diseases quantified by continuous variables are few, whereas several diseases are better approached by continuous indicators. For example, subclinical mastitis of the dairy cow is evaluated using a continuous marker of udder inflammation, the somatic cell score (SCS). Consequently, this study proposed to analyze spatialized risk and cluster components of herd SCS through a new method based on a spatial hazard model. The dataset included annual SCS for 34 142 French dairy herds for the year 2000, and important SCS risk factors: mean parity, percentage of winter and spring calvings, and herd size. The model allowed the simultaneous estimation of the effects of known risk factors and of potential spatial clusters on SCS, and the mapping of the estimated clusters and their range. Mean parity and winter and spring calvings were significantly associated with subclinical mastitis risk. The model with the presence of 3 clusters was highly significant, and the 3 clusters were attractive, i.e. closeness to cluster center increased the occurrence of high SCS. The three localizations were the following: close to the city of Troyes in the northeast of France; around the city of Limoges in the center-west; and in the southwest close to the city of Tarbes. The semi-parametric method based on spatial hazard modeling applies to continuous variables, and takes account of both risk factors and potential heterogeneity of the background population. This tool allows a quantitative detection but assumes a spatially specified form for clusters.

spatial epidemiology / cluster detection / hazard function / mastitis / dairy herd

1. INTRODUCTION

Spatial aspects of health events are of growing concern in epidemiology. Whether for emerging or endemic disease, regional differences such as heterogeneity of the background population, climatic and landscape conditions, agricultural activities, local health policy and the occurrence of peculiar events such as cattle fairs can have a great influence on disease spread and control. The tools available to explore spatial patterns range from geostatistics to point process approaches. Among these, the issue of cluster detection [7, 30] is

* Corresponding author:
emilie.gay@u707.jussieu.fr

of major interest, since targeting possible causes for high disease concentration can assist in control and prevention.

The main techniques used in cluster detection rely on scan statistics [1, 21, 22]. The principle is to compare the observed number of cases inside a moving window to the expected number of cases under some distribution assumptions (e.g. Poisson, Bernoulli) [17]. Most of these methods can deal with specific additional information at the individual level and integrate some covariables. Spatial modeling is another way to explore spatial patterns, and allows for quantification of the effects of known disease risk factors, and then attempts to focus on unexplained spatial clustering [20]. Among the several approaches, one approach handles the concept of infectious potential, through Susceptible-Infected-Recovered (SIR) models [10], which can be linked to point-pattern methodology [12]. Other approaches use the classical framework of linear mixed models with risk factors as fixed effects, while spatial variations are included as a random effect [28, 31]. Some last methods make the intensity of case events depend on location of cluster centers [19].

Until recently, the methods available dealt only with binary variables, and cluster detection for diseases measured by continuous variables remained an unexplored field. Nevertheless, several diseases can be better approached through continuous biological indicators, when no internationally recognized threshold value is available, or when the predictive value of the indicator is linear, a frequent case for biological markers [4, 25].

Lately, several researchers tackled cluster detection for new types of variables, and especially continuous ones [15]. Huang et al. proposed a spatial scan statistic with an exponential survival distribution function, and extendable to other distributions like the gamma and log normal. Besides its potential adaptation to censored survival data, this spatial scan statistic allows adjustment for the covariate effects. They used a linear regression model for the logarithm of the survival data for this purpose and assumed the error term to follow an extreme value distribution. Actually, their model reduced to a full parametric proportional hazard model. A former approach by Patil and Taillie [24] used the notion of upper-level-sets. The ratio of the number of cases per expected number of cases was replaced by the ratio of continuous responses per the expected values, possibly adjusted to factors. The new version[1] of the software SaTScan™ allows performing cluster detection with this exponential model, designed for survival time data, and with a normal model, designed for continuous data[2].

In this paper, we chose a different approach to detect clusters of high risk of bovine subclinical mastitis. The diagnosis of this disease mainly relies on the determination of milk somatic cell score (SCS), a continuous variable internationally recognized as a good indicator for mastitis control [13]. Risk factors associated with SCS have been widely investigated [5, 6, 26], but in these studies the SCS spatial aspects were not taken into account, while SCS typically presents strong spatial variations [11, 23]. Differences in natural resources, farm structure and market conditions cause different regions of the same country to implement different dairy management systems, and call for the introduction of a spatial component in SCS data analysis.

The purpose of this paper was to propose a new method for spatial cluster

[1] Kulldorff M., Information Management Services, Inc. SaTScanTM v7.0: Software for the spatial and space-time scan statistics [on line] (2006) http://www.satscan.org/.

[2] Kulldorff M., SaTScan™ User Guide for version 7.0 [on line] (2006) http://www.satscan.org/.

detection on continuous variables, with an application to bovine subclinical mastitis. We quantitatively analyzed the spatialized risk of SCS, using a spatial hazard model to simultaneously estimate the effects on SCS of known risk factors and of potential spatial clusters.

2. MATERIALS AND METHODS

2.1. Data

The study population consisted of a cohort of French Holstein dairy herds enrolled in Dairy Herd Improvement Association (DHIA) in 2000. The dataset included 34 142 farms with at least 20 cows.

Data concerning mastitis were extracted from the national DHIA database, which contained monthly data for every healthy lactating cow. The outcome variable was the annual herd SCS (ASCS), which was computed as the arithmetic mean of all monthly cow SCS values during 2000. ASCS indicated the farm status for subclinical mastitis risk. The other variables of the dataset were mean parity, percentage of calvings during the winter and spring period, and herd size, which had been recognized as herd factors influencing SCS [3]. The geographic coordinates of the farmers' addresses were obtained via the French National Institute of Statistics and Economic Studies. The statistical unit was the herd-year.

2.2. Statistical analysis

Statistical procedures were conducted using the software R 2.0.1[3] (descriptive analysis, models and map-making) and SaTScan™ (spatial scan statistic)[1].

[3] R Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, [on line] (2007) http://www.R-project.org.

2.2.1. Descriptive analysis of variables and spatial patterns

After a descriptive analysis of the studied variables, we used mapping representations to explore the spatial patterns of the data. The interpolation technique of kernel smoothing [27] was performed to represent ASCS intensity. The presence of spatial correlation was assessed and quantified using a correlogram, which is the graph of empirical autocorrelations of SCS values with respect to distance between farms [9].

2.2.2. Spatial hazard model

We explored spatial patterns of ASCS from the point of view of survival analysis [14], considering ASCS values instead of usual lifetime. We followed the sequence of the spatial distribution of farms as ASCS increased. The hazard function $r(z)$ of a non negative random variable, like the probability density function, completely characterized its probability distribution, i.e. $r(z) = \frac{f(z)}{1-F(z)}$ where f is the probability density function, and F the cumulative distribution function $f(t) = r(t)\exp\left(-\int_0^t r(s)ds\right)$. In our example, $r(z)$ is defined as the probability that a farm ASCS belonged to a small interval $[z, z + \Delta z]$, given that the ASCS is known to be greater or equal to the value z. The map of the farms, whose ASCS were greater or equal to a given level z, hereafter called the z-level map, described the spatial structure of farms still "at risk" at level z. To compare to classical survival analysis, $r(z)$ represented the hazard of occurrence of the ASCS value of a farm, i.e. the probability of removing from the $(z + \Delta z)$-level map a farm present on the z-level map.

The hazard function depended on observable local explanatory variables and on the presence of potential clusters, according to the proportional hazard model type.

Conditionally to explanatory variables and clusters, the ASCS were independently distributed with a spatial hazard function r:

$$r(z,x,W^x) = r_0(z)\exp\left(\sum_{j=1}^{J}\beta_j W_j^x - \phi(\gamma,x)\right) \quad (1)$$

where $r_0(z)$ is the underlying hazard function at an ASCS value z, x the spatial coordinates of the farms, $W^x = (W_1^x, ..., W_J^x)$ the vector of risk factors specific of the farm at location x, β_j the coefficient for the j^{th} risk factor, and $\phi(\gamma, x)$ a potential spatial cluster effect specified hereafter. In this spatial hazard function, a unit variation of an explanatory variable W_j with a positive β_j coefficient would increase the hazard of occurrence of the ASCS value of a farm at any level z by a factor $\exp(\beta_j) > 1$, and thus would decrease the occurrence of higher ASCS levels. Three explanatory variables were included in the model, with a regression parameter β, as follows:

- 1 continuous variable for mean parity;
- 1 continuous variable for the percentage of winter and spring calvings;
- 1 binary variable for herd size: 0 codes for the herds with less than 50 cows, and 1 for the herds with 50 or more cows.

The cluster effect $\phi(\gamma, x)$ aimed to take into account the spatial aggregation of the farms sharing approximately the same ASCS values. Instead of including definite spatial zones for clusters, we introduced a soft version of such zones under the form of a smooth parameterized function. Mathematically speaking, one can always approach any point set as a limit of a smooth function by kernel techniques. Consequently, we specified the cluster function as a sum of spatial Gaussian kernels as follows:

$$\phi(\gamma,x) = \sum_{k=1}^{K}\frac{\alpha_k}{2\pi\rho_k^2}\exp\left(-\frac{\|x-c_k\|^2}{2\rho_k^2}\right) \quad (2)$$

where K is a fixed number of clusters defined by a set of parameters $\gamma = (\alpha, \rho, c)$ written as a $(K \times 4)$ matrix. The α_k parameter is the strength of the cluster k, ρ_k its positive range, and c_k its two geographic center coordinates. If a point went close to a cluster, the distance $\|x - c_k\|$ was low, the exponential increased to the maximum value 1, so the cluster effect tended to $\alpha_k/2\pi\rho_k^2$. Conversely, if a point went far from the cluster, the exponential tended to 0 and the cluster effect fainted with Gaussian rate. If α was positive, the hazard of occurrence of any ASCS value of a farm decreased by a factor $\left(\exp\left(-\alpha_k/2\pi\rho_k^2\right)\right)$, the cluster was "attractive" and increased the occurrence of higher ASCS levels. By contrast, a negative α meant a "repulsive focus" and decreased the occurrence of high ASCS levels. The cluster effect is actually a generalized regression model where the response value depends only on a vector parameter γ associated to observable covariates, which are the spatial coordinates of farms. Thus, even if clusters defined here could be interpreted as hidden fields or a type of frailty model, they were not.

Having ordered the farm indices i according to increasing ASCS values z_i, an adapted Cox conditional likelihood for the model was defined as follows:

$$L^* = \prod_{i=1}^{n}\left[\frac{\exp\left(\sum_{j=1}^{J}\beta_j W_j^{x_i} - \phi(\gamma,x_i)\right)}{\sum_{l\geq i}\exp\left(\sum_{j=1}^{J}\beta_j W_j^{x_l} - \phi(\gamma,x_l)\right)}\right]. \quad (3)$$

In regards to statistical estimation and test issues, the conditional likelihood L^* asymptotically behaves as a true likelihood function under regularity assumptions. The β coefficients and the vector γ were then estimated by maximization of L^*. For sake of simplicity, we had not further developed the formula (3) analytically to achieve maximization of the

conditional likelihood. We used the function "mle" of "stats4" package (part of the R base source) which relies on robust and well known approximation algorithms, and we chose the Nelder-Mead method. We used the Likelihood Ratio Statistic (LRS) to test whether the effects of the covariables were significant [14]. The LRS is the difference between deviances $D = -2\log(L^*)$ of two nested models $\mathcal{M}_i$ and $\mathcal{M}_j$. If $\mathcal{M}_j$ contained ν more parameters than $\mathcal{M}_i$:

$$LRS = D(\mathcal{M}_i) - D(\mathcal{M}_j) = -2\left(\log\left(L^*(\mathcal{M}_i)\right) - \log\left(L^*(\mathcal{M}_j)\right)\right) \sim \chi^2(\nu\,\mathrm{df}). \quad (4)$$

Starting with $K = 0$, we increased K progressively until the addition of a new cluster was not significant anymore.

To visualize the results obtained by the model, we mapped the estimated clusters.

The proportional hazards assumption was checked using the Schoenfeld residuals with function "cox.zph" of package "survival", as recommended in Hill et al. [14]. To confirm the findings of the survival model concerning the risk factors, we performed a more classical model, the multiple linear regression (R-function "lm"), and paid attention to the agreement of the results with the findings of the spatial hazard model.

2.2.3. Analysis of the model properties

Properties of this spatial hazard model were tested on simulated datasets. We analyzed the number of clusters detected and their position, and compared to the number and position of clusters simulated. We used a spatial domain of 1 by 1 with a homogeneous point process, and distributed marks according to the presence of 0 up to 3 attractive clusters in 200 simulated datasets.

2.2.4. Spatial scan statistic with normal model

The new version[1] of the software SaTScan™ allows performing cluster detection with a normal model, designed for continuous data. SaTScan™ can integrate some covariables but it is recommended to use external regression software to adjust for quantitative variables. We first performed a linear regression (R-function "lm") in order to take into account the risk factors of the disease (mean parity, percentage of calvings during the winter and spring period, and herd size). We then performed the scan statistic on the residuals of the linear model to focus on unexplained clustering and be able to compare to the results of the spatial hazard model. We then mapped the detected clusters.

3. RESULTS

3.1. Descriptive analysis

Mean (sd) ASCS was 3.12 (0.55), while mean (sd) herd size was 39 cows (16), mean (sd) parity was 2.4 (0.3), and mean (sd) percentage of winter and spring calvings was 38% (16%).

The farm geographic distribution is illustrated in Figure 1, which represents the location of the farms studied in a gray background. The farm density was geographically non-homogeneous, and two areas had higher densities: (1) the northwest, which is the main dairy production area in France, with 61% of the total number of farms, and (2) the north tip, with 6% of the total number of farms. By contrast, the southeast (i.e. the Mediterranean area) had a very low farm density. The map of ASCS spatial intensity (Fig. 2) showed that the north-central area and the southwest had relatively high ASCS values of around 3.5.

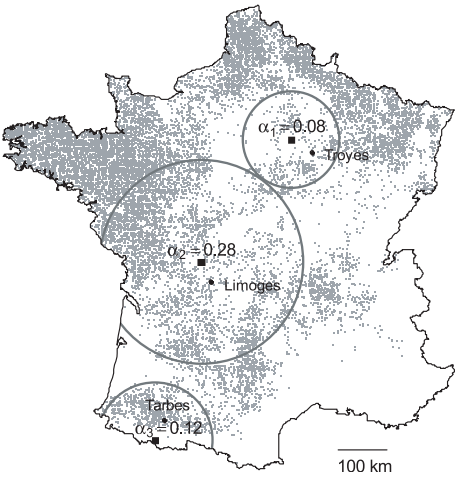

Figure 1. Farm location (gray background) and clusters of high annual somatic cell score detected by spatial hazard modeling in the study sample of dairy herds in France (n = 34 142, year 2000). ■: cluster centre; α: cluster strength; • main cities close to the clusters; ◯: cluster range (ρ).

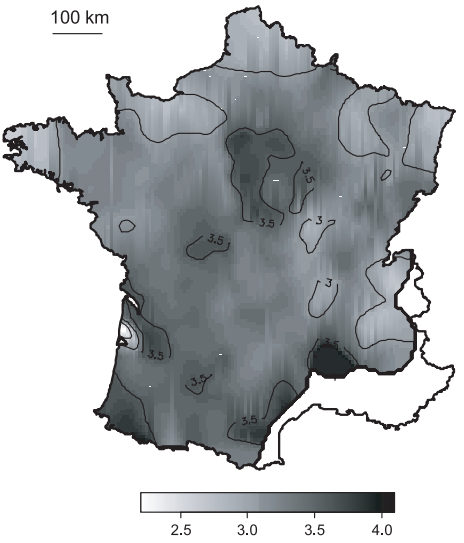

Figure 2. Intensity of the annual somatic cell score in the study sample of dairy herds in France (n = 34 142, year 2000).

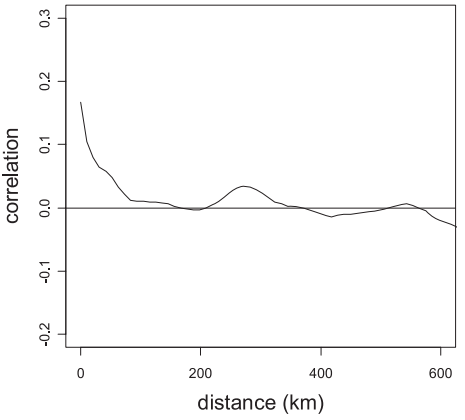

Figure 3. Correlogram of the annual somatic cell score in the study sample.

The correlogram of ASCS (Fig. 3) showed a positive and non negligible spatial correlation under a distance of 150 km with an approximated exponential form. Over this distance, it could be considered as constant around 0. The behavior of the correlogram near distance 0 pointed out a strong nugget effect (autocorrelation of 1 at a null distance if absence of a nugget effect), i.e. the presence of a relatively high white noise (non spatial correlation) of about 70% of the total variability.

3.2. Spatial modeling of mastitis risk

The model with K = 3 ($\mathcal{M}_{\beta,3}$) was selected since the presence of 3 clusters was highly significant, while the 4th cluster was not (Tab. I). The detailed results of estimations for this model are presented in Table II. Hazard of occurrence of the ASCS value of a farm was significantly decreased by increased mean parity (1 parity decreased the risk by $e^{0.5749}$ = 1.78), and thus this factor was positively associated with the occurrence of high ASCS. The percentage of winter and spring calvings was significant and, even if the association was low, it was positively associated with ASCS. The last risk factor, the herd size,

Table I. Tests of the different hazard models of the annual milk somatic cell score.

Model	Variables number	Deviance	LRS			*P* value
			Test	Value	(df)	
$\mathcal{M}_{0,0}$: no covariable	0	644496.3				
$\mathcal{M}_{\beta,0}$: risk factors alone	3	642493.2	$\mathcal{M}_{0,0}$ vs. $\mathcal{M}_{\beta,0}$	2003.1	(3)	$P < 0.001$
$\mathcal{M}_{\beta,1}$: risk factors+1 cluster	7	641136.1	$\mathcal{M}_{\beta,0}$ vs. $\mathcal{M}_{\beta,1}$	1357.1	(4)	$P < 0.001$
$\mathcal{M}_{\beta,2}$: risk factors+2 clusters	11	640719.8	$\mathcal{M}_{\beta,0}$ vs. $\mathcal{M}_{\beta,2}$	1773.4	(8)	$P < 0.001$
			$\mathcal{M}_{\beta,1}$ vs. $\mathcal{M}_{\beta,2}$	416.30	(4)	$P < 0.001$
$\mathcal{M}_{\beta,3}$: risk factors+3 clusters	15	640681.4	$\mathcal{M}_{\beta,0}$ vs. $\mathcal{M}_{\beta,3}$	1811.8	(12)	$P < 0.001$
			$\mathcal{M}_{\beta,1}$ vs. $\mathcal{M}_{\beta,3}$	454.7	(8)	$P < 0.001$
			$\mathcal{M}_{\beta,2}$ vs. $\mathcal{M}_{\beta,3}$	38.4	(4)	$P < 0.001$
$\mathcal{M}_{\beta,4}$: risk factors+4 clusters	19	640675.3	$\mathcal{M}_{\beta,0}$ vs. $\mathcal{M}_{\beta,4}$	1817.9	(16)	$P < 0.001$
			$\mathcal{M}_{\beta,1}$ vs. $\mathcal{M}_{\beta,4}$	460.8	(12)	$P < 0.001$
			$\mathcal{M}_{\beta,2}$ vs. $\mathcal{M}_{\beta,4}$	44.5	(8)	$P < 0.001$
			$\mathcal{M}_{\beta,3}$ vs. $\mathcal{M}_{\beta,4}$	6.1	(4)	NS

LRS: Likelihood Ratio Statistic; df: degrees of freedom; $\mathcal{M}_{\beta,k}$: model with risk factors (β part) and k clusters.

was not significant. The multiple linear regression demonstrated a similar relationship for the 3 covariables (Tab. III).

The three spatial clusters were attractive. The first one was detected in the northeast (Fig. 1), close to the city of Troyes. The second one spread in the center-west, around the city of Limoges. The third cluster was located in the southwest, close to the city of Tarbes.

3.3. Model properties under simulations

The detailed results of the simulation study are presented in Table IV. In 97% of the cases the right number of clusters was detected. Among those 97%, 5% detected an extra repulsive focus: when high values are concentrated on some areas, it can happen mechanically that low values are concentrated as well elsewhere, forming a repulsive focus. The mean distance between centers of detected and simulated clusters, i.e. the precision of localization, was 0.047 in the spatial domain of 1 by 1 unit.

3.4. Spatial scan statistic with a normal model

Specifying the upper limit for cluster size as a circular geographic region of radius 250 km, 6 significant clusters were detected with this method (Fig. 4).

4. DISCUSSION

4.1. Biological results

The results of the spatial hazard model concerning the introduced risk factors for ASCS were consistent with previously published results, indicating a significant association of parity and calving season with ASCS used as an indicator of subclinical mastitis. Increased mean parity increases the risk of high ASCS levels [18]; that can be due to the rise of persistence

Table II. Spatial hazard model of the annual milk somatic cell score (K = 3 clusters).

		Coefficient estimation	Standard deviation	exp(coef)	LRS	(df)	P value
Mean parity		–0.5749	0.0176	0.56	723.30	(1)	$P < 0.001$
Winter-spring calving		–0.0091	0.0004	0.99	651.70	(1)	$P < 0.001$
Number of cows		0.0321	0.0140	1.03	2.10	(1)	NS
Cluster 1	α	0.0783	0.0149				
	ρ	0.1105	0.0127				
	x_c	0.2462	0.0081				
	y_c	0.0942	0.0107				
Cluster 2	α	0.2775	0.0373				
	ρ	0.2328	0.0126				
	x_c	0.0444	0.0200				
	y_c	–0.1766	0.0248				
Cluster 3	α	0.1213	0.0217				
	ρ	0.1336	0.0104				
	x_c	–0.0597	0.0243				
	y_c	–0.5721	0.0257				

LRS: likelihood ratio statistic; df: degrees of freedom; α: cluster strength; ρ: cluster range; x_c: cluster x coordinate; y_c: cluster y coordinate.

Table III. Multiple linear regression model of the annual milk somatic cell score.

	Coefficient estimation	Standard deviation	P value
Intercept	2.136	0.023	$P < 0.001$
Mean parity	0.3408	0.0094	$P < 0.001$
Winter-summer calving	0.0042	0.0002	$P < 0.001$
Number of cows	0.0119	0.0074	NS

and intensity of mammary infections with parity. A high percentage of winter and spring calvings is a risk factor for high ASCS values, since weaker cow body condition and housing hygiene during this period increase the risk of subclinical mastitis [5]. Several studies highlighted that herd size was negatively associated with SCS [2, 23]. In this study, no significant effect was detected, but French herds are of little size (less than 0.7% of farms have more than 100 cows), and this can be a reason why the effect of heard size did not appear. Some other known risk factors for ASCS were not available in the present work. Particularly, information on hygienic and milking conditions was lacking [3, 8]. Nevertheless, the method allows complementary variation factors to be easily integrated in the model if available later.

Table IV. Results of the simulation process (spatial domain of 1 by 1 with a homogeneous point process, marks distributed according to the presence of 0 up to 3 "attractive" clusters).

Number of clusters simulated	Number of simulations	Number of clusters detected	Number of simulations (%)
0	50	0	50 (100)
1	50	1 attractive	43 (86)
		1 attractive + 1 repulsive	4 (8)
		2 attractive	3 (6)
2	50	2 attractive	47 (94)
		3 attractive	3 (6)
3	50	2 attractive	1 (2)
		3 attractive	44 (88)
		3 attractive + 1 repulsive	6 (12)

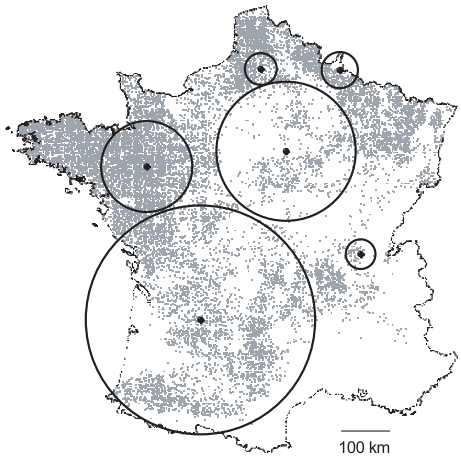

Figure 4. Farm location (gray background) and clusters of high annual somatic cell score detected by the spatial scan statistic in the study sample of dairy herds in France (n = 34 142, year 2000).

Concerning the second and original part of the model, the presence of 3 clusters was highly significant. The detected clusters, located in areas with low farm density (close to Troyes, Limoges and Tarbes), corresponded to regions mainly focused on bovine and ovine meat and on cereal production. It is consistent with the specialization in dairy production being linked in France with lower ASCS [5]. Introducing the farm density as a covariable in the model could be a way to approximate this specialization factor. Nevertheless, a local analysis would be necessary to precisely explain the factors associated with the clusters identified, since only local staff could have accurate and relevant information on local events or singularities having influenced ASCS.

4.2. Comparison with the spatial scan statistic with normal model

The spatial scan statistic method identified six clusters. The two larger ones, in the southwest and the northeast, included approximately the same regions as the three clusters detected with our spatial hazard model, but the four last ones were different.

The number of clusters detected by the spatial scan is not determined in an objective way, and depends on the chosen upper limit for cluster size. In our dataset the clusters were located in regions with

low farm density, consequently the classical use of 50% of the population as the upper limit was not adapted. We had to try several upper limits and chose the one that seemed to give the better results. Indeed, by its very construction of the alternative hypothesis, the scan tests can not theoretically handle the issue of multiple cluster detection. Moreover, with the spatial scan statistic, the integration of covariables had to be done in a first step before cluster detection.

4.3. Spatial hazard model for cluster detection

The properties of this new model tested via simulations demonstrated good detection ability and precision of localization. The spatial hazard model we developed has several advantages. First, the method applies to continuous variables; such a possibility is of recent growing concern in the issue of cluster detection. Secondly, the model is adjusted for risk factors as in Klassen et al. [16], and takes into account the potential heterogeneity of the background population. It focuses on unexplained spatial singularities, which can be detected even in low density areas. Moreover, the model being parametric, allows for the tests of comparative hypotheses on the two components, risk factors and cluster presence. As the classical hazard model, this one can easily accommodate censored data. Moreover, if the baseline hazard function is available (e.g. exponential, log Gaussian, Gamma type) a true parametric likelihood can be derived.

On the contrary, the present drawback of this method is the need to fix the number of clusters a priori and then to test submodels to retain only significant cluster components. As for true likelihood methods, one could bypass this issue by introducing an a priori parameterized probability distribution for the number K of cluster components. For example, if K is Poisson distributed with parameter λ, the corresponding log-likelihood in λ, β and γ is written: $L = \sum_{k=0}^{\infty} I_{\{k=k\}}(\log(L(\lambda, \beta, \gamma^{(k)}) - \lambda + k\log(\lambda) - \log(k!))$.

Clearly, the maximum of L was achieved for one value k of K, and this amounted simply to penalize the pseudo likelihood by the term $-\lambda + k\log(\lambda) - \log(k!)$ as the Bayesian Information Criterion (BIC) or Akaike Information Criterion (AIC) usually did.

Another consequence of using a parametric model is the need for a definition of the form of the cluster. In this study we chose a standard Gaussian form, characterized by intensity and range parameters and a circular form, a choice which was convenient for the further interpretation of the parameters. However, other parametric forms of the cluster function could better fit the problem [29] if they were supported by particular epidemiological arguments. But without prior assumptions about the possible spread of the disease, a circular form seems to be a good default choice.

This method of cluster detection based on a spatialized hazard model allows combining two important fields of epidemiological studies: the classical analysis of risk factor effects, and the spatial analysis of the disease. Moreover, this method applies to continuous as well as discrete variables, and gives quantitative results.

The approach of the spatialized risk with a cluster component is generic; it is also intended to apply to other diseases and to classical survival models considering death or occurrence times of infection as the variable of interest.

ACKNOWLEDGEMENTS

The authors would like to thank Joël Chadœuf for his helpful advice and his support on statistics.

REFERENCES

[1] Abrial D., Calavas D., Lauvergne N., Morignat E., Ducrot C., Descriptive spatial analysis of BSE in western France, Vet. Res. (2003) 34:749–760.

[2] Allore H.G., Oltenacu P.A., Erb H.N., Effects of season, herd size, and geographic region on the composition and quality of milk in the northeast, J. Dairy Sci. (1997) 80:3040–3049.

[3] Barkema H.W., Schukken Y.H., Lam T.J.G.M., Beiboer M.L., Benedictus G., Brand A., Management practices associated with low, medium, and high somatic cell counts in bulk milk, J. Dairy Sci. (1998) 81:1917–1927.

[4] Barnouin J., Chacornac J.P., Aissaoui C., El Idilbi N., Mazur A., Comment dépister les déséquilibres biologiques et les troubles de santé chez la vache laitière dans le cadre d'études écopathologiques? Vet. Res. (1994) 25:104–109.

[5] Barnouin J., Chassagne M., Bazin S., Boichard D., Management practices from questionnaire surveys in herds with very low somatic cell score through a national mastitis program in France, J. Dairy Sci. (2004) 87:3989–3999.

[6] Busato A., Trachsel P., Schallibaum M., Blum J.W., Udder health and risk factors for subclinical mastitis in organic dairy farms in Switzerland, Prev. Vet. Med. (2000) 44:205–220.

[7] Carpenter T.E., Methods to investigate spatial and temporal clustering in veterinary epidemiology, Prev. Vet. Med. (2001) 48:303–320.

[8] Chassagne M., Barnouin J., Le Guenic M., Expert assessment study of milking and hygiene practices characterizing very low somatic cell score herds in France, J. Dairy Sci. (2005) 88:1909–1916.

[9] Cressie N.A.C., Geostatistics, in: Barnett V., Bradley R., Fisher N., Hunter J., Kadane J., Kendall D. et al. (Eds.), Statistics for spatial data, John Wiley and Sons, New York, 1991, pp. 58–67.

[10] Doran R.J., Laffan S.W., Simulating the spatial dynamics of foot and mouth disease outbreaks in feral pigs and livestock in Queensland, Australia, using a susceptible-infected-recovered cellular automata model, Prev. Vet. Med. (2005) 70:133–152.

[11] Ely L.O., Smith J.W., Oleggini G.H., Regional production differences, J. Dairy Sci. (2003) 86:E28-E34.

[12] Gerbier G., Bacro J.N., Pouillot R., Durand B., Moutou F., Chadoeuf J., A point pattern model of the spread of foot-and-mouth disease, Prev. Vet. Med. (2002) 56:33–49.

[13] Harmon R.J., Physiology of mastitis and factors affecting somatic cell counts, J. Dairy Sci. (1994) 77:2103–2112.

[14] Hill C., Com-Nougué C., Kramar A., Moreau T., O'Quigley J., Senoussi R., Chastang C., Analyse statistique des données de survie, INSERM Médecine-Sciences Flammarion, Paris, 1990.

[15] Huang L., Kulldorff M., Gregorio D., A spatial scan statistic for survival data, Biometrics 63 (2007) 63:109–118.

[16] Klassen A.C., Kulldorff M., Curriero F., Geographical clustering of prostate cancer grade and stage at diagnosis, before and after adjustment for risk factors, Int. J. Health Geogr. (2005) 4:1.

[17] Kulldorff M., A spatial scan statistic, Commun. Stat.-Theory Methods (1997) 26:1481–1496.

[18] Laevens H., Deluyker H., Schukken Y.H., De Meulemeester L., Vandermeersch R., De Muelenaere E., De Kruif A., Influence of parity and stage of lactation on the somatic cell count in bacteriologically negative dairy cows, J. Dairy Sci. (1997) 80:3219–3226.

[19] Lawson A.B., Cluster modelling of disease incidence via RJMCMC methods: a comparative evaluation, Stat. Med. (2000) 19:2361–2375.

[20] Marshall R.J., A review of methods for the statistical analysis of spatial patterns of disease, J. R. Stat. Soc. A (1991) 154:421–441.

[21] Norstrom M., Pfeiffer D.U., Jarp J., A space-time cluster investigation of an outbreak of acute respiratory disease in Norwegian cattle herds, Prev. Vet. Med. (1999) 47:107–119.

[22] Odoi A., Martin S.W., Michel P., Middleton D., Holt J., Wilson J., Investigation of clusters of giardiasis using GIS and a spatial scan statistic, Int. J. Health Geogr. (2004) 3:11.

[23] Oleggini G.H., Ely L.O., Smith J.W., Effect of region and herd size on dairy herd performance parameters, J. Dairy Sci. (2001) 84:1044–1050.

[24] Patil G.P., Taillie C., Geographic and network surveillance via Scan Statistics for critical area detection, Stat. Sci. (2003) 18:457–465.

[25] Ratziu V., Massard J., Charlotte F., Messous D., Imbert-Bismut F., Bonyhay L., Tahiri M., Munteanu M., Thabut D., Cadranel J.F., Le Bail B., de Ledinghen V., Poynard T., Diagnostic value of biochemical markers (FibroTest-FibroSURE) for the prediction of liver fibrosis in patients with non-alcoholic fatty liver disease, BMC Gastroenterol. (2006) 6:6.

[26] Romain H.T., Adesiyun A.A., Webb L.A., Lauckner F.B., Study on risk factors and their association with subclinical mastitis in lactating dairy cows in Trinidad, J. Vet. Med. B Infect. Dis. Vet. Public Health (2000) 47:257–271.

[27] Silverman B.W., The kernel method for univariate data, in: Cox D., Hinkley D., Rubin D., Silverman B. (Eds.), Density estimation for statistics and data analysis, Chapman and Hall, London, 1986, pp. 34–94.

[28] Stevenson M.A., Benard H., Bolger P., Morris R.S., Spatial epidemiology of the Asian honey bee mite (Varroadestructor) in the North Island of New Zealand, Prev. Vet. Med. (2005):241–252.

[29] Tango T., Takahashi K., A flexibly shaped spatial scan statistic for detecting clusters, Int. J. Health Geogr. (2005) 4:11.

[30] Ward M.P., Carpenter T.E., Techniques for analysis of disease clustering in space and in time in veterinary epidemiology, Prev. Vet. Med. (2000) 45:257–284.

[31] Yang G.J., Vounatsou P., Zhou X.N., Tanner M., Utzinger J., A Bayesian-based approach for spatio-temporal modeling of county level prevalence of *Schistosoma japonicum* infection in Jiangsu province, China, Int. J. Parasitol. (2005) 35:155–162.