Détection d'agrégats spatiaux dans le cas d'une variable continue : application à un indicateur de l'infection mammaire chez les bovins

Emilie E. Gay

UNIVERSITE BLAISE PASCAL
CLERMONT II

UNIVERSITE D'AUVERGNE
CLERMONT I

ECOLE DOCTORALE
DES SCIENCES DE LA VIE ET DE LA SANTE

Année 2006
N° d'ordre :

THESE

Présentée à l'Université d'Auvergne pour l'obtention du grade de
DOCTEUR D'UNIVERSITE

Spécialité : Epidémiologie

Soutenue le 23 février 2006
par

Emilie GAY

TITRE :

Détection d'agrégats spatiaux dans le cas d'une variable continue : application à un indicateur de l'infection mammaire chez les bovins

JURY

M. Jacques BARNOUIN	Directeur de thèse	Directeur de recherches, INRA Clermont-Fd-Theix
M. Jean-Jacques BENET	Rapporteur	Professeur, Ecole Nationale Vétérinaire d'Alfort
M. Daniel COMMENGES	Rapporteur	Directeur de recherches, INSERM Bordeaux
M. Laurent GERBAUD	Président	PUPH, CHU Clermont-Ferrand
M. Mounir MESBAH	Rapporteur	Professeur, Université Parsi VI
M. Rachid SENOUSSI	Co-directeur de thèse	Directeur de recherches, INRA Avignon

Remerciements

Mes sincères remerciements à mes deux encadrants de thèse : Jacques qui a été à l'origine de mon parcours dans la recherche épidémiologique, et Rachid qui a su me faire surnager dans le flot des statistiques.

Mes remerciements vont aussi à Jean-Jacques Bénet, Daniel Commenges, Laurent Gerbaud, et Mounir Mesbah pour avoir accepté d'évaluer ce travail en participant à mon jury de thèse.

Un grand merci à l'ensemble de l'unité de biométrie d'Avignon :

Sabrina, compagnon de bureau, de travail, de galère et de fous rires, et Malika, avisée confidente du "bureau du fond", deux amies dont j'espère encore croiser la route ;

Joël dont la porte est toujours ouverte, et qui donne son aide, son temps et son attention sans compter ;

Hervé et André, je ne peux plus compter le nombre de services qu'ils m'ont rendus ;

le groupe "biomjeunes" avec Edith, Etienne, Gaël, Nathalie, Nico et Samuel, ainsi que les nouveaux membres qui n'auraient pas manqué d'en faire partie, Edwige, Gwendal, Lionel et Radu ;

Véronique et Marie-Claude, sans qui tout aurait été beaucoup plus compliqué et surtout beaucoup moins sympathique ;

Denis que je remercie encore de ses démarches m'ayant permis une fin de thèse dans les meilleures conditions ;

Pascal grâce à qui mes représentations cartographiques sont bien meilleures qu'elles ne l'étaient initialement ;

et enfin Claude, Franck et Vincent, pour les conseils et astuces qu'ils ont su me donner.

Merci à l'ensemble de l'unité d'épidémiologie animale de Clermont-Ferrand :

Gwenaël et Séverine, grâce à qui je n'ai jamais été sdf à Clermont, ainsi que Valérie, avec un souvenir ému de nos collaborations, confidences et rires toutes les 4 ;

Christian, Patrick, David et Michelle, pour l'attention qu'ils ont portée à mon travail et leurs remarques constructives ;

Nelly, Françoise et Ginette dont la disponibilité et l'accueil chaleureux m'ont rendu ce travail plus facile et plus agréable ;

et les "petits nouveaux", Chloé, Jocelyn, Anne-Sophie et Myriam, qui sauront j'en suis sûre perpétrer l'ambiance si conviviale de cette unité.

Je voudrais aussi remercier ma famille :

mes parents, qui ont su m'entourer et me conseiller dans cet ultime diplôme et ce premier "vrai boulot" à la fois ;

ma sœur et la si précieuse connivence qui nous unit ;

et mon frère que je vois avec amusement glisser sur la même pente que ses deux sœurs.

Merci à mes amis de toujours (ou presque !), Jessy, Delphine, Coralie, Emily, Sylvie, Caroline, Céline, Vanessa, Wilfried, Dicki, Monika et Maren ; et à mes amis plus récents "de thèse", Magali, Virginie et Valérie.

Merci à JP, qui le plus naturellement du monde n'a rien perturbé mais a tout changé.

Finally a special thanks to Greg, the first veterinary epidemiologist I met. Working with him 6.5 years ago in Australia certainly had been the starting point of all of this PhD adventure.

Table des matières

I. Introduction

I.1. Contexte épidémiologique

L'épidémiologie est l'étude des facteurs de santé des populations. S'il est largement reconnu que certains de ces facteurs possèdent une composante spatiale, cet aspect de la connaissance épidémiologique n'apparaît pas le plus documenté. Pourtant les différences régionales tenant à la population à risque, aux paramètres environnementaux, aux politiques de prévention, ou à la survenue d'évènements locaux peuvent influencer fortement les caractéristiques d'une maladie. La détection des spécificités locales peut permettre de mieux connaître les circonstances d'apparition et de développement des phénomènes pathologiques, de suggérer des hypothèses de causalité et d'aider à la mise en place de mesures de prévention et de contrôle prenant en compte les particularités locales.

Ces dernières années, l'importance de la surveillance et de l'étude spatiale des maladies a été remise sur le devant de la scène à l'occasion de la survenue de crises sanitaires ayant concerné des maladies émergentes ou réémergentes chez l'animal ou l'Homme (Encéphalopathie Spongiforme Bovine ou ESB, fièvre aphteuse, fièvre à virus West Nile, Syndrome Respiratoire Aigu Sévère ou SRAS, grippe aviaire). Ces crises sanitaires, dans lesquelles les déplacements, les migrations et les échanges commerciaux ont joué un rôle notable, ont été l'occasion d'une réflexion renouvelée sur les stratégies de détection, d'analyse et de contrôle des émergences (Barnouin et Vourc'h, 2004). Dans ce cadre, la mise au point de méthodes génériques de détection d'agrégats est apparue comme un objectif important, dans le but d'aider la réflexion de l'épidémiologiste et l'action du gestionnaire de la santé.

Une émergence est l'augmentation significative de l'incidence d'une maladie dans un cadre spatio-temporel donné (Lederberg et al., 1992), augmentation pouvant se traduire par différents schémas épidémiologiques. Le plus classique de ces schémas est le développement d'une maladie rare ou nouvelle (ESB, SRAS). Il s'agit alors de repérer le plus précocement possible quand une maladie s'écarte de sa ligne de base, en termes d'incidence et de répartition. Un autre schéma épidémiologique à prendre en compte concerne l'augmentation localisée d'incidence ou de gravité d'un évènement pathologique à caractère endémique. Dans cette situation, le but de l'analyse spatiale consiste à identifier l'émergence de foyers de maladie, c'est-à-dire à détecter si la distribution spatiale des cas est « normale » (répartition de la maladie au hasard), ou bien s'il existe des agrégats de cas. Un agrégat se traduit par des cas plus regroupés dans l'espace que ne le laisse présager le hasard, compte tenu de la distribution de la population.

Les travaux concernant la détection spatiale des agrégats se sont multipliés ces dernières années en épidémiologie humaine (Fèvre et al., 2001; Thomas et Carlin, 2003), mais aussi en épidémiologie animale (Perez et al., 2002; Olea-Popelka et al., 2003).

Ces travaux s'appuient sur les différentes façons d'observer, de représenter et d'étudier les caractéristiques spatiales d'une maladie. Les informations spatiales peuvent être classées en 2 grands types (Lawson, 2001) :

- données ponctuelles, correspondant au fait que chaque individu est spatialement localisé par un point portant une marque de maladie (processus ponctuels marqués spatiaux) ;
- données groupées pour lesquelles les descripteurs de la maladie sont comptabilisés dans chaque région subdivisant la zone d'étude.

Généralement, l'information épidémiologique disponible porte sur la présence de la maladie. Si les renseignements sur la structure de la population générale ne sont pas disponibles, l'étude ne porte que sur les cas ; si ces renseignements sont disponibles, les cas sont rapportés à la population sous jacente divisée en cas et en témoins. Les témoins peuvent rassembler l'ensemble des non-cas de la population totale, ou seulement un échantillon de celle-ci. Les données spatialisées sur la maladie sont ainsi décrites soit sous forme de processus ponctuels marqués, avec une marque égale à 0 pour les témoins et à 1 pour les cas (informations spatiales ponctuelles), soit sous forme d'effectif ou de proportion de malades dans chaque région subdivisant la zone d'étude (informations spatiales groupées).

D'autres situations ne se prêtent pas au classement des individus en cas et en témoins. En effet, certaines pathologies, comme le diabète chez l'Homme ou l'acidose ruminale chez les bovins, ont un caractère progressif et ne peuvent être approchées que par l'intermédiaire de variables continues, en l'occurrence des indicateurs biologiques (glycémie, pH). C'est cette situation qui prévaut en matière d'infection mammaire chez les bovins, pathologie endémique d'intérêt majeur pour la filière bovine laitière, dépistée par la mesure d'un indicateur de l'inflammation mammaire, le comptage des leucocytes du lait. Ainsi, les données disponibles concernant ce comptage seront utilisées pour concrétiser la réflexion sur les méthodes de détection d'agrégats menée dans cette thèse.

Pour les maladies mesurées par l'intermédiaire d'une variable continue, les données existent soit sous forme de processus ponctuels marqués par la valeur de la variable, soit sous forme de moyennes (éventuellement pondérées) de la variable dans chaque région subdivisant la zone d'étude. Dans notre cas, l'attribution, pour chaque élevage bovin repéré dans l'espace, d'une marque correspondant à la valeur du comptage leucocytaire place l'étude dans le cadre des processus ponctuels.

Pour les données spatiales ponctuelles avec marque dichotomique 0/1, de nombreux outils de détection d'agrégats ont été développés (cf. paragraphe I.3. et chapitre II). Nous évaluerons dans quelle mesure ces outils pourraient s'appliquer à notre indicateur continu de l'infection mammaire, et surtout quelles sont les limites de leur adaptation. Nous proposerons ensuite d'autres approches pour contourner ces limites.

Afin de développer une méthode répondant correctement à la problématique biologique, nous présenterons tout d'abord la maladie sur laquelle porte cette thèse.

I.2. Les infections mammaires des bovins

Les infections mammaires des bovins, ou mammites, sont des inflammations de la glande mammaire, dues essentiellement au développement de micro-organismes dans la mamelle, mais pouvant aussi avoir une origine traumatique ou chimique. La forme pathologique la plus courante est l'infection mammaire subclinique, forme inapparente non détectée par le seul examen clinique. Ce type de mammite résulte de l'évolution à bas bruit de foyers infectieux au sein du parenchyme mammaire, évolution qui peut se dérouler sur une période assez longue et aboutir à une fibrose des quartiers atteints, stoppant leur production laitière.

Les infections mammaires, en particulier subcliniques, sont les maladies infectieuses les plus fréquentes du troupeau bovin laitier. Busato et al (2000) ont estimé la prévalence des infections mammaires subcliniques, en définissant un quartier comme infecté si le California Mastitis Test (test visuel basé sur la floculation de l'ADN des leucocytes du lait) était positif. Ces auteurs ont obtenu, à l'échelle de l'animal, une estimation de la prévalence des infections subcliniques de 47,8% (±3,7) pour la période 7-100 jours post-partum, et de 61,5% (±2,6) pour la période 100-305 jours post-partum. Cette maladie a un fort impact économique, puisqu'elle entraîne des coûts de traitement, une éventuelle réforme anticipée des animaux, une diminution de la qualité du lait et des pertes de production laitière (DeGraves et Fetrow, 1993; Coulon et al., 2002; Seegers et al., 2003).

Les pathogènes en cause dans les infections mammaires sont principalement des bactéries : *Staphylococcus aureus*, des staphylocoques coagulase-négative, des streptocoques dont *Streptococcus agalactiae*, *S. dysgalactiae* et *S. uberis*, et des entérobactéries dont *Escherichia coli* (Busato et al., 2000). Ces infections ont un caractère contagieux intra élevage, mais pas inter élevages. Leur dépistage repose sur le comptage des cellules somatiques du lait (CCS), terme consacré pour désigner la numération leucocytaire effectuée sur le lait de mélange des 4 quartiers. Le CCS est en effet un témoin internationalement reconnu (Reneau, 1986) du niveau d'inflammation mammaire (via une mobilisation de leucocytes à partir du pool sanguin circulant), et donc indirectement du niveau d'infection (Harmon, 1994). Le CCS est soit utilisé de façon brute pour une évaluation individuelle du niveau d'infection mammaire, soit transformé en score de cellules somatiques (SCS) afin de normaliser sa distribution ($SCS = \log_2(CCS/100\,000)+3$) (Ali et Shook, 1980) pour une évaluation collective du niveau d'infection du troupeau. La mesure des CCS est effectuée en routine, une fois par mois environ, sur toutes les vaches productrices non traitées dans tous les élevages adhérant à un organisme de Contrôle Laitier. Ces données sont ensuite regroupées dans la base de données nationale de France Contrôle Laitier.

Les principaux facteurs de risque des infections mammaires ont été largement étudiés, et peuvent être divisés en 3 niveaux : niveau de l'animal, niveau de l'élevage, et niveau de l'environnement.

Les facteurs tenant à l'animal sont les suivants :

- la race (Coulon et al., 1996; Barnouin et al., 1999; Rupp et al., 2000) ;
- le rang de lactation (Coulon et al., 1996; Laevens et al., 1997; Busato et al., 2000; Rupp et al., 2000) ;
- le stade de lactation (Coulon et al., 1996; Laevens et al., 1997; Busato et al., 2000).

Les facteurs tenant à l'élevage sont plus nombreux :

- effectif du troupeau (Oleggini et al., 2001; Skrzypek et al., 2004) ;
- proportion de primipares (Coulon et al., 1996; Barnouin et al., 1999; Rupp et al., 2000) ;
- répartition des périodes de vêlage (Coulon et al., 1996; Rupp et al., 2000; Barnouin et al., 2004) ;
- conditions de logement (Busato et al., 2000) ;
- conditions de traite (Barkema et al., 1998; Barnouin et al., 2004; Skrzypek et al., 2004; Chassagne et al., 2005) ;
- conduite du tarissement (Barkema et al., 1998; Skrzypek et al., 2004) ;
- conditions générales d'hygiène (Barkema et al., 1998; Barnouin et al., 2004) ;
- gestion de l'alimentation (Barkema et al., 1998; Barnouin et al., 2004; Skrzypek et al., 2004) ;
- spécialisation et attention de l'éleveur (Barnouin et al., 1999; Busato et al., 2000; Barnouin et al., 2004).

Enfin, les facteurs tenant à l'environnement sont les suivants :

- saison (Coulon et al., 1996; Skrzypek et al., 2004) ;
- région (Rupp et al., 2000; Oleggini et al., 2001; Ely et al., 2003).

La composante spatiale des infections mammaires n'a été que très peu étudiée, et uniquement abordée par le biais de l'introduction de l'effet région en tant que facteur de risque dans une comparaison de moyennes (Rupp et al., 2000) ou dans un modèle linéaire généralisé (Oleggini et al., 2001; Ely et al., 2003). Néanmoins, ces analyses simples ont mis en évidence des variations significatives entre les SCS de différentes régions. Les différences spatiales dans les ressources naturelles, les structures d'élevage, les systèmes d'exploitation ou l'organisation du conseil technique aux éleveurs peuvent provoquer des variations dans les performances techniques des élevages, notamment au niveau de la gestion des infections mammaires. Connaître ces particularités spatiales peut permettre d'identifier des causes locales non évidentes d'une agrégation, et de mettre en place des mesures de prévention adaptées.

I.3. Les outils pour la détection d'agrégats

Les méthodes utilisées pour la détection des phénomènes d'agrégation sont classiquement divisées en 2 catégories (Wakefield et al., 2000; Lawson, 2001) :

- méthodes de détection d'une tendance globale d'agrégation. Ces méthodes dites non spécifiques ne visent pas à localiser des agrégats, mais à tester de manière générale la présence d'agrégation.
- méthodes de détection spécifique d'agrégats. Ces méthodes permettent de localiser et tester les agrégats.

Les méthodes spécifiques peuvent se décliner en méthodes de détection ciblée ou non ciblée. La détection ciblée ne s'applique que dans le cas où l'on connaît a priori une ou plusieurs sources potentielles d'agrégation (zones industrielles, centrales nucléaires…), et où la présence d'agrégats autour de ce(s) point(s) est testée (Morris et Wakefield, 2000). Quant à la détection non ciblée, elle ne donne pas d'a priori sur la localisation ni le nombre des agrégats (chapitre II.1.2).

Pour l'une ou l'autre de ces catégories de détection existent des approches paramétriques et non paramétriques. L'approche paramétrique spécifie un modèle statistique avec paramètres d'intérêt, alors que l'approche non paramétrique ne spécifie pas de modèle mais seulement des hypothèses d'intérêt.

Le tableau 1 présente les différentes méthodes de détection disponibles (qui seront détaillées dans le chapitre II), classées selon le type de formalisation (paramétrique ou non) et le type de détection (spécifique ou non). Les données et variables auxquelles ces outils peuvent être appliqués sont déclinées, ainsi que les paramètres à fixer a priori dans chacun des cas.

Type de formalisation	Type de détection	Méthode	Principe	Type de données	Type de variable	Paramètres a priori
non paramétrique (tests)	non spécifique	coefficient de Moran	autocorrélation	P ou G	D ou C	matrice de pondération
		indice des proches voisins	proches voisins	P	D	nombre de proches voisins
		test de Cuzick et Edwards	proches voisins	P	D	nombre de proches voisins
		K-fonctions	moment d'ordre 2	P	D	–
	spécifique	GAM	fenêtres mobiles	P ou G	D	risque α, taille grille et disques
		test de Besag et Newell	fenêtres mobiles	P ou G	D	nombre critique de cas
		test du scan spatial	fenêtres mobiles	P ou G	D	–
		test de Tango	fenêtres mobiles	P ou G	D	mesure de proximité
paramétrique (modèles)	non spécifique	lois log-normale/gamma	sur-dispersion	P ou G	D ou C	choix de la loi
		modèles auto-régressifs	autocorrélation	G	D ou C	choix du modèle
		champ aléatoire caché	lissages, géostatistique	P ou G	D ou C	choix du champ
	spécifique	foyers d'agrégation	MCMC	P ou G	D	foyers, forme de l'agrégation

P : données ponctuelles
G : données groupées
D : variable dichotomique
C : variable continue

Tableau 1 : Présentation synthétique des différentes méthodes de détection d'agrégats

I.4. Adéquation des outils au cas de variables continues

Exception faite du coefficient d'autocorrélation de Moran, les méthodes non paramétriques disponibles ne s'appliquent que pour des maladies approchées par des variables dichotomiques de type cas/témoins, situation la plus fréquemment rencontrée en matière d'étude d'évènements pathologiques. Il en va de même avec la méthode paramétrique de détection spécifique. Les méthodes paramétriques non spécifiques, quant à elles, peuvent être utilisées avec une variable continue, mais l'objectif de telles méthodes est d'avantage la prise en compte de l'agrégation pour de meilleurs modèles prédictifs que l'analyse de cette agrégation en elle-même.

Le cadre de notre problématique est le suivant : la maladie est mesurée par une variable continue et l'objectif est de détecter et de localiser les agrégats (approche spécifique). Le coefficient de Moran est utilisable avec notre indicateur, mais cette méthode non spécifique ne répond pas à notre problématique.

Par ailleurs, les tests de détection spécifiques pourraient être utilisés moyennant quelques adaptations :

- choix d'un seuil pour la variable continue afin de se replacer dans le cadre d'une variable dichotomique ;
- ou choix de seuils multiples pour donner plus de souplesse à la variable, et adaptation des tests pour plusieurs niveaux d'une variable discrète (possible pour la GAM et le scan spatial).

Mais la discrétisation entraîne une perte notable de l'information sur la variable. Elle ne fait que déplacer le problème du domaine quantitatif au domaine qualitatif par le choix très délicat du ou des seuils de la variable. De plus, modifier les tests pour les adapter à des seuils multiples rend les calculs complexes et l'interprétation difficile.

Enfin, la méthode paramétrique de détection spécifique modélise seulement l'intensité de cas (0/1), alors que dans notre cadre il s'agit de modéliser l'intensité d'un processus ponctuel marqué par une variable continue.

Dans le contexte de notre problématique, l'idéal serait en fait de pouvoir disposer de méthodes de détection d'agrégats comblant les lacunes des méthodes précédentes, et satisfaisant le cahier des charges suivant :

- méthodes applicables aux variables continues ;
- méthodes prenant en compte l'éventuelle hétérogénéité de la population sous-jacente ;
- méthodes prenant en compte les facteurs de risque connus de la maladie ;
- méthodes ne nécessitant pas d'a priori sur le nombre, la forme et la localisation des agrégats.

I.5. Objectif et stratégie d'analyse

Les objectifs de la thèse sont de proposer des méthodes génériques de détection d'agrégats satisfaisant au cahier des charges précédemment défini, et en particulier applicables aux variables continues (marqueur biologique), et d'analyser dans ce contexte l'indicateur des infections mammaires subcliniques.

L'agrégat est défini en tant que groupe de valeurs proches (une fois pris en compte les facteurs de risque connus de la maladie) spatialement liées et ayant peu de chances d'apparaître sous l'effet du hasard.

L'intensité des points marqués, c'est-à-dire dans notre cas l'intensité des élevages marqués par leur valeur de SCS, peut se décomposer, tout comme les densités de probabilité de 2 variables, de la façon suivante :

$$\overline{\lambda}(x,z) = \lambda_0(x)\lambda(z|x) \quad (1)$$

où $\lambda_0(x)$ est l'intensité de points, c'est-à-dire la densité d'élevages, et $\lambda(z|x)$ est l'intensité de la marque conditionnellement aux points, c'est-à-dire l'intensité de SCS conditionnellement à la position des élevages.

Notre étude ne s'intéresse pas dans le cas présent à la densité des élevages $\lambda_0(x)$ en elle-même, elle porte seulement sur l'intensité de SCS conditionnellement à la position des élevages. Pour détecter de façon spécifique les agrégats de valeurs de score, 2 approches sont là aussi possibles : une approche paramétrique et une approche non paramétrique.

L'approche non paramétrique ne donnera pas de forme à $\lambda(z|x)$. Nous utiliserons une méthode basée sur le calcul de la distance d'Hellinger entre distributions spatiales. Nous calculerons la distance entre la distribution spatiale à différents niveaux de score et la distribution initiale (densité d'élevages globale).

Sous l'approche paramétrique, on modélise de façon spécifique $\lambda(z|x)$. Cette intensité de score conditionnellement aux points, tout comme la densité d'une loi de survie, sera reliée à une fonction de risque $r(z|x)$ (Droesbeke et al., 1989) par la formule :

$$\lambda(z|x) = r(z|x)\exp\left(-\int_0^z r(u|x)du\right) \quad (2)$$

Nous choisirons donc de construire un modèle de survie spatialisée pour les SCS, en modélisant la fonction de risque instantané conditionnelle $r(z|x)$ de disparition de points en fonction du niveau de score.

Revue bibliographique des méthodes de détection d'agrégats

II. Revue des principales méthodes de détection d'agrégats

II.1. Méthodes non paramétriques : tests de détection d'agrégats

Il existe de nombreuses versions pour chacun des tests que nous allons présenter, certains auteurs ayant adapté ces tests à leur propre problématique ou ayant comblé certaines de leurs lacunes initiales. Nous détaillerons les 8 méthodes principales (Marshall, 1991; Ward et Carpenter, 2000; Carpenter, 2001), elles-mêmes à l'origine de nombreuses adaptations.

Les notations adoptées pour la présentation des différentes méthodes sont les suivantes :

- W l'espace d'étude ;
- Φ la réalisation d'un processus ponctuel (d'intensité λ si nécessaire) dans W ;
- N le nombre d'individus (population totale) ;
- N^+ le nombre total de cas parmi N ;
- x_i la position dans l'espace, soit du point i appartenant à Φ (données ponctuelles), soit du centroïde de la région (données groupées) ;
- z_i la marque du point i pour un processus ponctuel (par exemple 1/0 pour les cas/témoins), ou la valeur globale de la région pour les données groupées.

II.1.1. Détection non spécifique de l'agrégation

II.1.1.1. Coefficient d'autocorrélation spatiale : coefficient de Moran

Le coefficient de Moran (Moran, 1950) est le plus utilisé des outils se basant sur l'autocorrélation spatiale (corrélation entre les différentes valeurs d'une même variable). Il s'agit d'un coefficient d'autocorrélation calculé avec des pondérations adaptées aux positions spatiales des individus. Il s'en distingue toutefois par l'estimation de la variance (ici le dénominateur de la fraction) qui est calculée sans pondération :

$$I = \frac{N \sum_i \sum_j w_{ij} (z_i - \bar{z})(z_j - \bar{z})}{\left(\sum_i \sum_j w_{ij} \right) \sum_k (z_k - \bar{z})^2} \qquad (3)$$

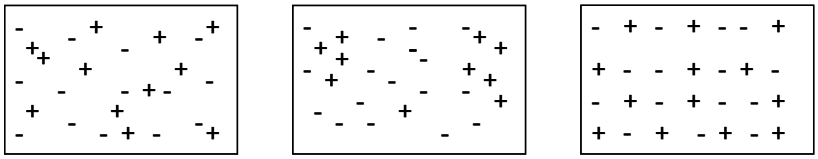
où $\bar{z}$ est la valeur moyenne sur le domaine W, et w_{ij} la matrice de pondération.

La matrice de pondération w_{ij} définit les relations spatiales entre sites i et j. Les relations spatiales peuvent, par exemple, être définies par de simples relations de voisinage : w_{ij} est alors une variable dichotomique prenant la valeur 1 si i et j sont voisins, et la valeur 0 sinon. De même, si les relations spatiales sont définies par des distances, w_{ij} peut être, par exemple, l'inverse de la distance entre i et j.

La variable z peut être le nombre ou la proportion de cas dans la région subdivisant la zone d'étude pour les données groupées, ou encore la valeur de la marque du point pour les processus ponctuels. Les variables continues, comme les indicateurs, peuvent donc être traitées par cette technique.

La justification d'un tel indice statistique est l'adéquation de son comportement pour différentes hypothèses. Ainsi la valeur moyenne de I est (Figure 1) :

- proche de 0 pour une répartition des marques z_i spatialement aléatoire (corrélation nulle) ;
- positive pour une agrégation spatiale des marques (corrélation positive) ;
- négative pour une dispersion (ou régularité) spatiale des marques (corrélation négative).



distribution aléatoire I=0 | agrégation I>0 | dispersion (régularité) I<0

Figure 1 : Valeur du coefficient I de Moran sous différentes hypothèses

L'utilité du coefficient de Moran, utilisable pour des marques de type variables discrètes ou continues, est cependant assez réduite, car la quantité d'information apportée sur la répartition spatiale est faible et peu discriminante. Elle est moindre, par exemple, que celle apportée par un corrélogramme ou un variogramme (Arnaud et Emery, 2000), lesquels intègrent la notion de portée des corrélations et la part de variabilité spatiale et non spatiale dans la variabilité totale. Le variogramme ou le corrélogramme ne sont pourtant pas classiquement utilisés pour la détection de l'agrégation, mais plutôt en tant qu'outils descriptifs préalables à la construction de modèles spatiaux.

Le calcul du coefficient de Moran ne prend pas en compte l'hétérogénéité potentielle de la répartition de la population sous-jacente en cas d'utilisation de données groupées : un rapport de 1 cas pour 100 individus, ou de 1000 cas pour 100 000 individus aura ainsi la même valeur statistique alors qu'il n'a pas la même précision ni la même valeur biologique. Enfin, le calcul du coefficient de Moran implique de choisir l'a priori des pondérations et du nombre de proches voisins, choix qui n'est pas toujours facile à justifier.

II.1.1.2. Indice des proches voisins

L'indice des proches voisins (Clark et Evans, 1954) est le rapport de la distance moyenne entre cas voisins sur la distance moyenne entre cas attendus sous l'hypothèse d'une distribution aléatoire uniforme :

$$R = \frac{\overline{D}_{observé}}{\overline{D}_{attendu}} = \frac{\sum_{i|z_i=1} \sum_{j \in U_k(i)|z_j=1} d_{ij} / N}{0{,}5\sqrt{A/N}} \quad (4)$$

où $U_k(i)$ est l'ensemble des k plus proches voisins de i, d_{ij} la distance entre i et j, et A l'aire de la zone.

L'indice R est égal à 0 pour une agrégation maximale, à 1 pour une distribution aléatoire uniforme, et il est supérieur à 1 en cas de dispersion.

Pour calculer la significativité, on utilise une statistique similaire à R, dont la loi de probabilité asymptotique est gaussienne sous l'hypothèse d'une répartition aléatoire uniforme dans le domaine, et pour un nombre conséquent de cas :

$$\frac{(\overline{D}_{observé} - \overline{D}_{attendu})}{\sigma_{\overline{D}_{attendu}}} \text{ avec } \sigma_{\overline{D}_{attendu}} = \frac{0{,}26136}{N/\sqrt{A}} \quad (5)$$

Dans ce calcul, seuls les cas sont comptabilisés. Si la population sous-jacente est spatialement homogène, cette caractéristique ne pose pas de problème, les agrégats détectés ne pouvant qu'être reliés à la maladie. Par contre, si la population est hétérogène, il n'est pas possible de savoir si les agrégats détectés sont attribuables à la maladie ou à la distribution de population. Un autre inconvénient de cette méthode est le choix a priori du nombre k de voisins considérés, qui peut être problématique. Enfin, cet indice dépendant de l'estimation de l'aire A du domaine considéré, il pose le problème de l'échelle spatiale des évènements.

II.1.1.3. Test de Cuzick et Edwards

Le test de Cuzick et Edwards (Cuzick et Edwards, 1990) est une évolution de la technique des proches voisins, qui a l'avantage de tenir compte des témoins dans le calcul. Il s'agit du comptage, pour chaque cas i, du nombre de cas présents parmi les k plus proches voisins (les voisins étant des cas ou des témoins) :

$$T_k = \sum_i \sum_{j \in U_k(i)} \delta_i \delta_j \quad (6)$$

où $U_k(i)$ est l'ensemble des k plus proches voisins de i, et δ_i une variable dichotomique prenant la valeur 1 si i est un cas et la valeur 0 si c'est un témoin.

Sous l'hypothèse d'indépendance, l'espérance de T_k conditionnellement au nombre de cas total N^+ est la suivante :

$$E(T_k) = N^+ k\left(\frac{N^+ - 1}{N - 1}\right)$$

Une valeur élevée de T_k indique une tendance à l'agrégation. La significativité est calculée par l'approximation gaussienne de la statistique τ suivante :

$$\tau = \frac{T_k - E(T_k)}{\sqrt{\mathrm{var}(T_k)}} \tag{7}$$

Le choix de l'ordre k de voisinage est délicat : en effet plus cet ordre est grand plus le type d'agrégation est complexe, mais plus il est alors possible de passer à côté d'une agrégation à courte distance.
Si l'échantillonnage est correctement réalisé, la prise en compte des témoins permet de tenir compte de la répartition de la population sous jacente. Cependant, si l'hétérogénéité de population est très forte, la proximité des cas ne sera pas prise en compte, car les voisins éloignés ou proches sont gérés de la même façon. Ross et Davis (1990) ont proposé une solution à ce problème, en ajoutant un terme de pondération basé sur la distance entre i et j. Cette solution permet aussi de contourner la question du choix du nombre de voisins k, mais pose en revanche le problème du choix de la fonction de pondération.

II.1.1.4. K-fonctions

Cette méthode examine la différence entre les K-fonctions (mesure du moment d'ordre 2) des cas et des témoins en fonction des distances (Diggle et Chetwynd, 1991). Pour un processus ponctuel stationnaire d'intensité λ (nombre d'évènements par unité d'aire), la K-fonction à la distance d est définie par :

$$K(d) = \lambda^{-1} E\left[\text{nombre d'autres évènements à une distance} \le d \text{ de l'évènement considéré}\right]$$

Par exemple, pour un processus de Poisson standard (complètement aléatoire), on a : $K(d)=\pi d^2$. Pour un processus agrégé $K(d)>\pi d^2$. Enfin, pour un processus régulier $K(d)<\pi d^2$.

Pour tester l'agrégation des données, Diggle considère $K_1(d)$ et $K_2(d)$ les K-fonctions respectives des cas et des témoins. Pour une distance d fixée, la différence positive $D(d)=K_1(d)- K_2(d)$ mesure la sur agrégation des cas par rapport aux témoins. En choisissant quelques distances d_i sur un intervalle d'intérêt, on construit un test statistique global d'agrégation de la façon suivante :

$$\hat{D} = \sum_k \frac{\hat{D}(d_k)}{\sqrt{\mathrm{var}\left(\hat{D}(d_k)\right)}} \tag{8}$$

où $\hat{D}(d) = \hat{K}_1(d) - \hat{K}_2(d)$, $\hat{K}$ étant un estimateur avec correction des effets de bords de K.

Sous l'hypothèse d'indépendance des cas, $\hat{D}$ est centrée. La variance de $\hat{D}(d_k)$ n'est pas calculable et est estimée sous l'hypothèse nulle par une procédure de Monte Carlo (MC). Le graphe de $\hat{D}(\cdot)$ selon d permet de voir à partir de quelle distance d on s'éloigne du hasard.

On peut reprocher à cette méthode l'utilisation de la statistique K, qui n'est bien définie et interprétable que pour les distributions de population homogènes, ce qui est loin d'être le cas général. Cependant la symétrisation des rôles des cas et des témoins relativise le non-sens, et permet de tenir compte de l'hétérogénéité de la distribution de la population sous-jacente.

Remarque générale sur les tests de détection non spécifique

Pour ces méthodes de détection non spécifique de l'agrégation, nous avons présenté des tests utilisant les lois asymptotiques pour la détermination des seuils de significativité des statistiques observées. A l'heure actuelle, les procédures de randomisation (Manly, 1991) sont de loin les plus utilisées avec les mêmes statistiques.

II.1.2. Détection spécifique des agrégats

Les tests de détection spécifique qui vont être présentés sont basés sur le principe d'un balayage du domaine d'étude par le biais de fenêtres mobiles.

II.1.2.1. Geographic analysis machine (GAM)

La GAM est la première méthode utilisant les fenêtres mobiles qui a été mise au point (Openshaw et al., 1987). C'est une méthode graphique. Le principe est de superposer une grille sur la zone d'étude, en plaçant des disques de rayon donné, centrés sur les points de la grille, de façon à ce que les disques se recouvrent partiellement (jusqu'à 80% de recouvrement).

Pour chaque disque, un test local permet de comparer le nombre observé de cas dans le disque au quantile α d'une loi de Poisson dont la moyenne est le nombre attendu de cas (dépendant du nombre total de cas, du nombre total d'individus, et du nombre d'individus dans le disque).

Les disques avec un test significatif sont dessinés sur la carte, une zone comptant de nombreux disques étant définie comme un agrégat.

L'avantage de cette méthode est qu'elle permet de prendre en compte l'éventuelle hétérogénéité de la population sous-jacente. Par contre, la mise en œuvre simultanée de nombreux tests non indépendants constitue une limite importante, puisque les propriétés statistiques sont alors difficiles à évaluer.

Cette méthode implique de fixer a priori de nombreux paramètres. Tout d'abord le choix du seuil de significativité α est délicat. Cette probabilité doit être petite (classiquement α = 0,005) car les tests sont multiples, mais il n'existe pas de développement théorique pour fixer un critère de choix ; pourtant le choix de α est déterminant pour la sensibilité et la spécificité de la détection. La méthode dépend aussi fortement de la taille k de la grille (et donc du rayon des disques), puisque la fixation d'un seuil α fera apparaître au final en moyenne $k.\alpha$ disques significatifs sous hypothèse de totale indépendance. Ensuite, le choix du rayon des disques n'est pas évident, la procédure est souvent répétée pour différentes valeurs de rayon sans critère de sélection au final. Le fait de garder le même rayon de disque quand la distribution de population de base n'est pas homogène fait que les tests locaux ont le même niveau α, mais des puissances distinctes.

Il faut aussi souligner que cette approche a été initialement mise au point pour détecter des agrégats d'évènements rares, puisqu'elle est testée par rapport à une hypothèse poissonnienne. Enfin il faut noter que cette méthode est gourmande en temps de calcul.

II.1.2.2. Test de Besag et Newell

Le test de Besag et Newell (Besag et Newell, 1991) dérive de l'approche GAM, mais considère la surface nécessaire pour rassembler un nombre k de cas. Il s'applique aux données groupées par zones géographiques. Pour 1 cas dans une zone A_0, les zones A_i autour de A_0 sont classées en fonction de la distance croissante par rapport à A_0 (distance de centroïde à centroïde). On définit N_i^+ le nombre cumulé d'autres cas dans A_i, et M l'indice nécessaire pour accumuler k cas :

$$M = \min\left\{ i : N_i^+ \geq k \right\}$$

L'agrégation locale autour de A_0 est indiquée par un M petit. La significativité du test local est donnée par $P(M \leq m)$ calculée sous H_0 (hypothèse poissonnienne) pour m une réalisation de M :

$$p = P(M \leq m) = 1 - P(M > m) = 1 - \sum_{s=1}^{k-1} \frac{\exp(-E)(E)^s}{s!} \qquad (9)$$

où E est le nombre des cas attendus dans la zone délimitée par m.

Bien que ce ne soit pas le but de la méthode, l'agrégation globale peut être testée via une procédure de Monte Carlo. Si R est le nombre de tests significatifs pour un α fixé, la valeur attendue de R sous H_0 est alors approximativement αN^+, avec N^+ le nombre de cas total donc le nombre de tests locaux. Si la valeur observée de R dépasse αN^+, un test de Monte Carlo est utilisé pour déterminer la significativité.

De même que dans la méthode GAM, le test de Besag et Newell permet de prendre en compte l'hétérogénéité de la population sous-jacente. L'avantage par rapport à la GAM est de baser la définition non pas sur une taille de cercles, mais sur un nombre de cas, ce qui permet de détecter des agrégats dans des zones à faible densité de population par exemple, mais surtout d'avoir la même puissance pour tous les tests locaux. Le choix a priori de k reste toujours délicat et les calculs sont souvent effectués pour plusieurs valeurs, ce qui pose les problèmes liés aux tests multiples. Enfin, comme pour la GAM, cette méthode a été initialement mise au point dans le cadre de la détection d'agrégats d'évènements rares.

II.1.2.3. Test du scan spatial

Le principe du scan, une fenêtre mobile balayant la zone d'étude, a été d'abord utilisé par Turnbull et al (1990). Le test proposé par ces auteurs détecte, parmi des fenêtres avec une population constante N_R, la fenêtre contenant le nombre maximal de cas M_R. L'hypothèse nulle de distribution aléatoire est rejetée si M_R est plus élevé qu'une valeur V, déterminée par la distribution sous H_0 de M_R et par un niveau de significativité α. La distribution sous H_0 de M_R est non calculable, elle est obtenue par simulations de Monte Carlo.

Le test du scan spatial, développé ensuite par Kulldorff (Kulldorff et Nagarwalla, 1995; Kulldorff, 1997), permet la comparaison entre la proportion de cas et la proportion attendue de cas à l'intérieur d'une fenêtre mobile de rayon variable (0 à 50% de la zone d'étude).

La proportion attendue peut être calculée sous 2 modèles probabilistes distincts, basés sur les distributions de Poisson ou de Bernouilli. La distribution de Bernouilli s'applique pour des données ponctuelles de type cas/témoins, celle de Poisson s'appliquant pour des données groupées. Quand l'incidence de cas est faible, la loi de Poisson est une bonne approximation de la loi de Bernouilli, et demande moins de temps de calcul.

Si p est la probabilité d'avoir un cas à l'intérieur de la fenêtre, et q la probabilité des cas à l'extérieur de la fenêtre, les hypothèses H_0 et H_1 s'écrivent de la façon suivante :

- H_0 : "le nombre de cas dans une zone est proportionnel à la taille de la population dans cette zone : $p=q$" ;
- H_1 : "la proportion de cas est plus élevée dans la fenêtre qu'en dehors : $p>q$".

Les fenêtres ne contenant pas toutes le même nombre d'individus, le critère du nombre maximal de cas n'est pas adéquat, pas plus que la proportion maximale de cas, puisque les variances ne sont pas égales. Le critère choisi est celui du test du rapport de vraisemblances.

Modèle de Bernouilli

Pour le modèle de Bernouilli la fonction de vraisemblance, dans chaque fenêtre B (cercle de centre c et de rayon r), s'écrit à une constante près :

$$L(B,p,q) = p^{N_B^+}(1-p)^{N_B-N_B^+} q^{N^+-N_B^+}(1-q)^{(N-N_B)-(N^+-N_B^+)} \tag{10}$$

où N est la taille de la population totale, N_B la taille de la population dans la fenêtre B, N^+ le nombre total de cas, N_B^+ le nombre de cas dans B.

Le test du rapport du maximum de vraisemblance est :

$$S = \frac{\sup_{B,p>q} L(B,p,q)}{\sup_{p=q} L(B,p,q)} = \frac{\max_B L(B)}{L_0} \tag{11}$$

avec L_0 une constante dépendant uniquement du nombre de cas.

Modèle de Poisson

Pour le modèle de Poisson, l'écriture de la fonction de vraisemblance et du rapport de vraisemblance est un peu plus complexe (Kulldorff, 1997).

Pour les 2 modèles, la distribution de S est inconnue et ne présente pas de forme analytique simple. Par contre, des procédures de Monte Carlo peuvent être utilisées pour échantillonner la distribution exacte de S. La fenêtre qui constitue l'agrégat le plus probable est alors celle qui, parmi les fenêtres B, maximise S. Les agrégats secondaires (définis comme les seconds plus probables et qui sont situés en dehors de la zone du premier agrégat) sont ensuite identifiés.

L'éventuelle hétérogénéité de la population sous-jacente est prise en compte, les témoins étant comptabilisés. L'avantage de cette méthode, par rapport aux 2 précédentes, est qu'il n'est pas nécessaire de choisir la taille de la fenêtre puisque le rayon est variable. En conséquence le problème des tests multiples ne se pose pas. Par contre, le rapport de vraisemblance teste l'hypothèse nulle contre l'hypothèse alternative qu'il n'y a qu'un seul agrégat dans la zone d'étude. S'il y a plus d'un agrégat, l'hypothèse alternative n'est pas correctement spécifiée et la puissance du test diminue, tendant à la sur-détection d'agrégats (Tango, 2000).

II.1.2.4. Test de Tango

Le test de Tango (Tango, 1995) permet de détecter une tendance globale à l'agrégation et de localiser les centres des agrégats. La zone d'étude étant divisée en m régions (ou fenêtres), le principe est de comparer les nombres observés et attendus de cas, pondérés par une mesure de proximité a_{ij} entre les régions i et j :

$$C_\lambda = \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij}(\lambda)(N_i^+ - E_i)(N_j^+ - E_j) = \sum_{i=1}^{m} U_i(\lambda) \qquad (12)$$

où N_i^+ est le nombre observé de cas dans la région i, et E_i le nombre attendu de cas dans cette même région sous l'hypothèse nulle de répartition aléatoire des cas.

Plus les cas se concentrent dans un nombre restreint de régions, plus C_λ est élevé. La mesure de proximité a_{ij} peut avoir différentes expressions, dépendant toutes du paramètre λ appelé échelle d'agrégation. Voici quelques exemples pour a_{ij} :

- fonction exponentielle décroissante simple : $a_{ij} = \exp(-d_{ij}/\lambda)$ où d_{ij} est la mesure de la distance entre les régions i et j ;
- fonction exponentielle plus complexe : $a_{ij} = \exp(-4(d_{ij}/\lambda)^2)$;
- composante dichotomique valant 1 pour $d_{ij} \leq \lambda$ et 0 sinon.

La significativité globale est testée par la statistique suivante :

$$T = \frac{C_\lambda - E(C_\lambda)}{\sqrt{\mathrm{var}(C_\lambda)}} \qquad (13)$$

Asymptotiquement sous H_0, T dûment re-normé et recentré suit une loi du Chi-2. En cas d'agrégation significative, les centres des agrégats sont indiqués par les centroïdes des régions avec une valeur U_i élevée.

Un inconvénient de cette méthode est le choix a priori de λ. Pour répondre à ce problème Tango a fait évoluer sa méthode (Tango, 2000). Le test statistique devient alors le minimum du profil des P-valeurs de C pour λ, où λ varie de façon continue entre une valeur proche de 0 et environ la moitié de la zone d'étude :

$$P_{\min} = \min_{\lambda} \Pr\left\{ C_\lambda > c_\lambda \mid \mathrm{H}_0, \lambda \right\} = \Pr\left\{ C_\lambda > c_\lambda \mid \mathrm{H}_0, \lambda = \lambda^* \right\} \quad (14)$$

où c_λ est la statistique de test observée en tant que fonction de λ, et λ^* la valeur de λ atteignant la P-valeur minimum de C_λ.

La distribution de P_{min} sous H_0 est obtenue par simulations de Monte Carlo.

Cette méthode a plusieurs avantages. Tout d'abord, elle prend en compte l'hétérogénéité de la distribution de la population sous-jacente. Ensuite, tout comme pour le test du scan spatial, aucun a priori n'est nécessaire sur le nombre, la forme ou la distribution des agrégats. De plus, des simulations de Monte Carlo ont montré que la puissance du test de Tango est supérieure à celle du scan spatial. Par contre, dans cette méthode, seule la significativité globale de l'agrégation est testée, il n'existe pas de tests locaux pour apprécier la significativité de chaque agrégat.

II.1.3. Conclusion sur les méthodes non paramétriques de détection

Le problème majeur des tests est la prise en compte de l'hétérogénéité de la population sous-jacente. L'intégration des témoins dans les statistiques de test permet d'ajuster au niveau pratique les différences de répartition de population.

Avec une standardisation adaptée, il est possible d'inclure l'effet de certaines covariables (race, sexe…). La standardisation est une méthode de neutralisation de facteurs de confusion, elle permet de comparer des taux dans 2 populations en ajustant sur les facteurs de confusion, lesquels peuvent ne pas être répartis de façon égale dans les 2 populations. Cependant, cette standardisation peut difficilement intégrer tous les facteurs de risque connus de la maladie sans devenir extrêmement complexe.

Pour aboutir à une prise en compte satisfaisante des facteurs de risque et de l'hétérogénéité de la population sous-jacente, une approche paramétrique semble être mieux adaptée.

II.2. Méthodes paramétriques : modélisation spatiale

L'approche paramétrique consiste à intégrer dans un modèle une composante d'agrégation spatiale en plus des facteurs de risque. Les méthodes sont nombreuses et s'adaptent à différents contextes (Marshall, 1991; Diggle, 2000). Nous en verrons ici les grands principes.

II.2.1. Introduction d'un paramètre statistique général

Lors du processus de modélisation de la maladie, il est possible de prendre en compte une sur-dispersion des valeurs du modèle en incluant un paramètre d'agrégation des observations sans structure spatiale particulière. La sur-dispersion peut avoir 2 causes distinctes : une hétérogénéité spatiale ou une autocorrélation spatiale.

II.2.1.1. Paramètre d'hétérogénéité spatiale

L'hétérogénéité spatiale, non corrélée, peut être prise en compte par l'introduction d'un paramètre de sur-dispersion, en utilisant par exemple les lois log-normale ou gamma dans la modélisation de l'incidence de la maladie (Clayton et Kaldor, 1987). Cette méthode est plus efficace pour capturer les changements régionaux progressifs que pour inclure des changements abrupts locaux de type agrégats. De plus, elle sous entend l'indépendance des observations entre des régions voisines, ce qui, la plupart du temps, n'est pas une hypothèse raisonnable.

II.2.1.2. Paramètre d'autocorrélation

L'autocorrélation spatiale peut, quant à elle, être prise en compte dans des modèles de type autorégressifs (Ferrandiz et al., 1999). Classiquement, il s'agit de processus sur grille régulière, mais les modèles autorégressifs ont été adaptés aux processus ponctuels. Il existe 4 types de modèles autorégressifs : les simultanés, les conditionnels, les intrinsèques et les modèles "somme" (Richardson, 1992).

Modèles autorégressifs simultanés

Les modèles autorégressifs simultanés (SAR) relient la valeur z_i à une fonction linéaire des autres valeurs z dans les unités voisines et à une erreur aléatoire sans structure spatiale.

$$z_i - \bar{z} = a \sum_j w_{ij} (z_j - \bar{z}) + \varepsilon_i \qquad (15)$$

où $\bar{z}$ est l'espérance supposée constante des z_i, a le paramètre de dépendance spatiale, w_{ij} une matrice normée de pondération spatiale donnée a priori, et ε_i un bruit aléatoire suivant une loi normale $N(0,\sigma^2)$.

Modèles autorégressifs conditionnels

Les modèles autorégressifs conditionnels (CAR) définissent l'espérance de la valeur z_i conditionnellement aux autres valeurs z_j comme une fonction linéaire des valeurs de z sur les unités voisines, avec une variance conditionnelle de $z_i|z_j$ constante.

$$\begin{aligned} E(z_i \mid z_j, i \neq j) &= \bar{z} + a\sum_j w_{ij}(z_j - \bar{z}) \\ \mathrm{var}\,(z_i \mid z_j, i \neq j) &= \sigma^2 \end{aligned} \tag{16}$$

Modèles autorégressifs conditionnels intrinsèques

Les modèles autorégressifs conditionnels intrinsèques définissent l'espérance de z_i comme une moyenne des valeurs voisines dans le cadre d'une matrice w de contiguïté en 0-1.

$$\begin{aligned} E(z_i \mid z_j, i \neq j) &= w_{i.}^{-1}\sum_j w_{ij} z_j \\ \mathrm{var}\,(z_i \mid z_j, i \neq j) &= \sigma^2 / w_{i.} \end{aligned} \tag{17}$$

Modèles "somme"

Enfin, les modèles "somme" rajoutent aux modèles intrinsèques une source de variation supplémentaire sans structure spatiale.

L'estimation des paramètres de ces modèles autorégressifs se fait de manière numérique par maximum de vraisemblance. Le choix du modèle est orienté par le calcul de l'autocorrélation (coefficient de Moran) et par celui de la variabilité dans et entre les différentes échelles géographiques (canton, département, région) (David et al., 2002).

Cette approche est non spécifique, elle permet d'introduire des relations spatiales dans les modèles (donc une éventuelle agrégation), mais n'a pas pour but l'étude de l'agrégation en elle-même.

II.2.2. Introduction d'un champ aléatoire caché

Une autre approche, très prisée actuellement pour les variables dichotomiques, consiste à utiliser un modèle linéaire généralisé. Le principe est de prendre en compte les facteurs de risque de la maladie en effets fixes, et d'intégrer une composante d'agrégation spatiale sous forme d'un champ aléatoire caché additionnel, pour expliquer la variabilité des observations et leur manque d'adéquation au modèle initial. Ces modèles sont du type :

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \alpha + \sum_j \beta_j W_j(x) + S(x) \tag{18}$$

où $p(x)$ est la probabilité d'être un cas, α le risque de base, W_j le vecteur des valeurs du facteur de risque j avec β_j le coefficient associé, et $S(x)$ le champ aléatoire caché.

Il existe 2 niveaux d'étude statistique pour ce champ aléatoire. Le premier, non paramétrique, vise à reconstituer la cartographie du champ $S(x)$ par des méthodes de lissage et d'interpolation spatiale (Kelsall et Diggle, 1998). Le second utilise une paramétrisation des lois de probabilité de ce même champ aléatoire (Banfield et Raftery, 1993), et vise à estimer ces paramètres pour évaluer la structure de corrélation sous-jacente. Dans le cadre des modèles géostatistiques (Diggle et al., 1998), la structure de corrélation spatiale est donnée par le variogramme. Comme la vraisemblance est alors un mélange complexe de distributions, les paramètres du modèle sont estimés dans un cadre bayésien, et l'implémentation est effectuée par méthodes MCMC (Monte Carlo Markov Chain).

Une grande part d'a priori réside dans la construction de ces modèles et dans la spécification du champ caché. De plus, bien que la nature stochastique de l'agrégation soit mieux spécifiée qu'avec un paramètre statistique général, ces méthodes ne permettent qu'une détection non spécifique de l'agrégation.

II.2.3. Introduction de foyers d'agrégation

Cette approche permet une détection spécifique des agrégats de cas, mais est beaucoup moins utilisée que les précédentes (Lawson, 1995). L'intensité de cas $\lambda(x)$ (x coordonnées dans l'espace) intègre la position dans l'espace de foyers d'agrégation, et a une forme du type :

$$\lambda(x) = g(x)\left\{1 + \sum_k h(x - y_k)\right\} \tag{19}$$

où $g(x)$ est l'intensité de présence de la population à risque (obtenue avec une fonction de lissage par exemple), k le nombre de foyers d'agrégation, y_k la position du foyer k, et $h(.)$ la fonction de distribution des éléments de l'agrégat, décrivant la relation entre les cas et le foyer.

Conditionnellement à la donnée des foyers d'agrégation, les cas sont gouvernés par un processus poissonnien hétérogène. Un cadre bayésien, avec des a priori sur les foyers et la forme des agrégats, peut être développé par des techniques MCMC pour le calcul des vraisemblances (Gangnon et Clayton, 2000; Lawson, 2000).

Ce type de modèle est explicatif à l'ordre 1, c'est-à-dire qu'il ne fait que modéliser l'intensité de présence de cas, et se restreint à ce niveau à un comportement purement poissonnien. Il ne permet pas d'interpréter les interactions (ou corrélations) entre les cas (ordre 2).

II.3. Bilan sur les méthodes de détection d'agrégats

Les méthodes de détection d'agrégats sont assez diversifiées, et basées sur des concepts à la fois différents et complémentaires. Leur comparaison sensu stricto est difficile car la définition de l'agrégat utilisée n'est pas identique, comme le montre le tableau 2.

Méthode	Critère définissant l'agrégation
coefficient de Moran	présence d'autocorrélation
indice des proches voisins	distance entre cas
test de Cuzick et Edwards	nombre de cas parmi les k plus proches voisins
K-fonctions	sur agrégation des cas par rapport aux témoins
GAM	nombre observé de cas par rapport au nombre attendu
test de Besag et Newell	surface nécessaire pour rassembler k cas
test du scan spatial	proportion observée de cas par rapport à la proportion attendue
test de Tango	proportion observée de cas par rapport à la proportion attendue
sur-dispersion	hétérogénéité spatiale
autocorrélation	présence d'autocorrélation
champ aléatoire caché	présence d'une structure spatiale dans les données
foyers d'agrégation	présence de foyers de cas

Tableau 2 : Définition de l'agrégation selon la méthode utilisée

Pour un contexte donné de détection d'agrégat, il convient donc de bien examiner le cadre de travail (données, contexte…) et la question posée. Il faut en outre vérifier que les hypothèses sous-jacentes aux outils utilisés sont correctes, et que la méthode répond au mieux à la problématique. Ainsi, dans le domaine de l'épidémiologie, quelques tests sont jugés inappropriés par Tango (1999) dans la mesure où ils ne prennent pas en compte l'hétérogénéité de la population sous-jacente (coefficient de Moran, indice des proches voisins, et K-fonctions dans une certaine mesure). La possibilité de prendre en compte des facteurs de risque connus est aussi un point important si le but est de détecter les agrégats inexpliqués. Enfin, de nombreuses méthodes impliquent de préciser a priori certains paramètres, ce qui peut poser problème si les connaissances sur la maladie ne permettent pas de guider ce choix.

Approche non paramétrique pour variables continues

III. Méthode de détection d'agrégats basée sur la distance d'Hellinger

III.1. Introduction

III.1.1. Principe

La distance d'Hellinger est une métrique initialement utilisée pour quantifier l'écart entre mesures de probabilité sur des espaces quelconques, afin d'étudier la convergence en loi (Gibbs et Su, 2002). Dans le cadre de notre étude nous l'utilisons pour mesurer la distance entre distributions spatiales de points sur un domaine D. Plus précisément, nous calculons la distance d'Hellinger entre la distribution globale (ou initiale) des élevages $\overline{p}_{z_0}(x)$ et les distributions des élevages $\overline{p}_{z_1}(x)$ à différents niveaux de score z_j (Figure 2). Les densités de points sont normalisées.

$$Hel(z_j) = \sqrt{\frac{1}{2}\int_D \left(\sqrt{\overline{p}_{z_0}(x)} - \sqrt{\overline{p}_{z_j}(x)}\right)^2 dx} \qquad (20)$$

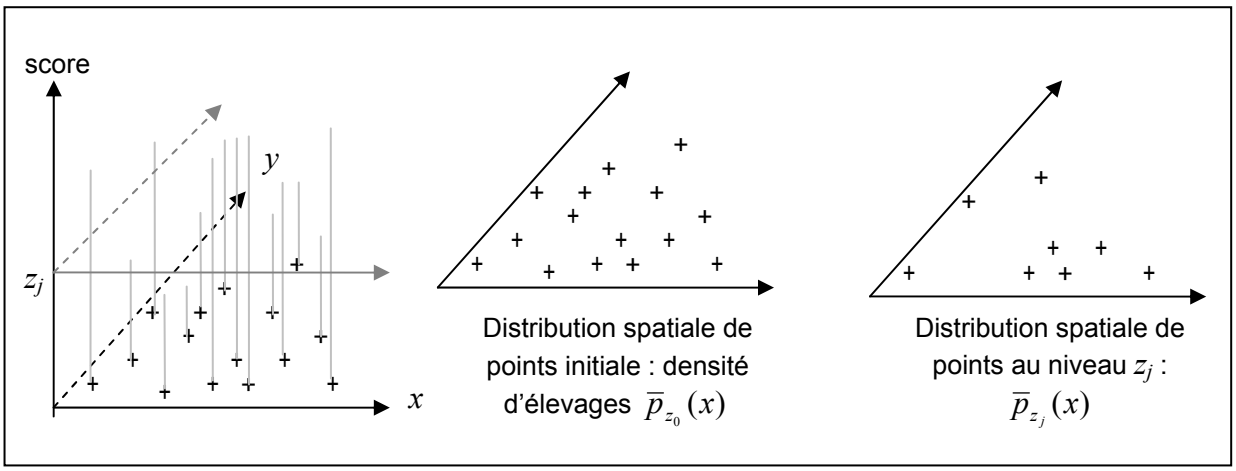


Figure 2 : Représentation des points des élevages situés dans l'espace selon les différents niveaux de la marque (valeur de score cellulaire)

Les hypothèses de travail que nous considérons sont définies comme suit :

- H_0 : "les valeurs de score sont distribuées de façon indépendante dans l'espace" ;
- H_1 : "les valeurs de score présentent une structuration spatiale avec l'existence de zones agrégatives".

Sous l'hypothèse nulle, la distribution normalisée de points, quel que soit le niveau de score z_j considéré, doit être proche de la distribution initiale qu'est la densité d'élevages. La distance d'Hellinger sera donc faible pour tous les z_j. Sous l'hypothèse alternative, la distribution de points à certains niveaux de score z_j présentera des agrégats, elle sera donc plus éloignée de la distribution initiale, et la distance d'Hellinger sera d'autant plus grande.

La distance d'Hellinger globale est définie comme la somme des valeurs des distances d'Hellinger aux différents niveaux de score z_j, pondérée par la proportion de points n_j à chaque niveau :

$$Hel = \sum_j Hel(z_j) \frac{n_{j-1} - n_j}{n} \quad (21)$$

La significativité de la statistique des distances d'Hellinger globale et par niveau z_j est calculée par tests de permutations spatiales. Le processus de permutations assigne de façon aléatoire chaque valeur de score à un point de l'espace pour constituer un nouveau jeu de données permuté, et ce un grand nombre de fois (Manly, 1991). Pour un nombre d'observations n, le jeu de données possible vaut $n!$, qui vaut par approximation de Stirling : $n! \approx n^n e^{-n} \sqrt{2\pi n}$, ce qui, déjà pour n=100, dépasse largement le nombre d'atomes dans l'univers (10^{80}).
Pour un seuil α donné, ces nouveaux jeux de données permettent d'établir un intervalle de confiance d'acceptation de H_0.

Après l'appréciation globale de l'agrégation, chaque niveau de score est analysé. La cartographie des densités d'élevages, pour chacun de ces niveaux, permet de visualiser la concentration progressive, si elle existe, des points dans des zones spécifiques. En effet certaines zones présentent peu de disparition de points car les scores sont élevés, la densité normalisée par rapport au reste de la zone d'étude tend donc à augmenter au fur et à mesure des scores croissants, ce sont les agrégats. Le niveau de score apparaissant comme le plus discriminant est choisi pour déterminer les limites spatiales des agrégats.

III.1.2. Etapes de mise en œuvre

1) Intégration des facteurs de risque connus et disponibles de la maladie

Afin d'intégrer les facteurs de risque de la maladie, nous avons modélisé dans un premier temps le score cellulaire à l'aide d'une régression linéaire. Ainsi, la détection d'agrégats est effectuée sur les résidus du modèle linéaire.

2) Estimation des densités d'élevages aux différents niveaux de score

L'hétérogénéité de distribution des points pose un certain nombre de problèmes. Elle peut entraîner une mauvaise estimation des densités (estimation utilisée dans le calcul de la distance d'Hellinger), et rendre difficile la visualisation des agrégats sur les cartes de densité de points aux différents niveaux de score. Pour palier ces difficultés, nous avons opté pour une technique d'homogénéisation de la distribution par transformation radiale de l'espace (Senoussi et al., 2000). La transformation permet de contracter ou de dilater localement l'espace en fonction du nombre de points, tout en conservant la topologie de voisinage. De telles techniques d'homogénéisation par transformation de l'espace ont été développées dès les années 60 (Selvin et al., 1988; Gastner et Newman, 2004).

3) Détection d'agrégats

Nous avons ensuite effectué la détection d'agrégats proprement dite, qui implique le calcul de la distance d'Hellinger globale observée, la mise en œuvre des tests de permutations, l'identification des concentrations de points et la cartographie des agrégats.

III.2. Article n°1 : Présentation de la méthode et analyse spatiale des SCS

Cet article, intitulé "A spatial clustering analysis for continuous variables with application to milk somatic cell score in France", a pour vocation de décrire cette nouvelle méthode de détection d'agrégats basée sur la distance d'Hellinger, et d'analyser la distribution des SCS. L'étude est réalisée sur un échantillon de 5000 élevages français inscrits au Contrôle Laitier en 1996.

Cet article a été soumis à la revue Preventive Veterinary Medicine en juin 2005.

A spatial clustering analysis for continuous variables with application to milk somatic cell score in France

E. Gay [1 2 *], R. Senoussi [1], J. Barnouin [2]

[1] Unité de Biométrie, INRA Domaine St-Paul, Site Agroparc, 84914 Avignon Cedex 9, France
[2] Unité d'Epidémiologie Animale, INRA, 63122 St-Genès-Champanelle, France

[*] Corresponding author:

Unité de Biométrie
INRA Domaine St-Paul, Site Agroparc
84914 Avignon Cedex 9
FRANCE
00 33 (0)4 32 72 21 56
fax: 00 33 (0)4 32 72 21 82
emilie.gay@avignon.inra.fr

Abstract

This paper proposes a spatial analysis method based on the Hellinger distance between distributions to explore spatial patterns and detect geographical clusters of continuous epidemiological variables. The method was applied to herd milk somatic cell score (SCS) on a national scale in France.

Available data were dairy farm coordinates, mean annual SCS of the herd (ASCS) and basic information on herd characteristics. To explore this data, maps of farm density and ASCS intensity were drawn using kernel smoothing. This analysis was completed by variogram estimation to investigate the structure of spatial correlations. A cluster was defined as a geographically bound group of close values being unlikely to have occurred at random, after taking account of the known and available variation factors for ASCS. Therefore a linear model was performed integrating those factors, and further analysis was focused on ASCS residuals. We used the Hellinger distance to compare geographical distribution of the farms at different levels of ASCS residuals. To enhance the estimation of farm density in the method, a space deformation was performed to make farm distribution homogeneous. The observed Hellinger distances were compared to the ones obtained by spatial random permutations of the data. Finally densities were represented on deformed map at the different levels of residuals to visualise the clusters, and these clusters were overlain on the original map.

The mean ASCS was 3.16. The global Hellinger distance was higher than that which could have been expected by random. Mapping for the different levels of ASCS residuals displayed a progressive concentration on 3 clusters: 1) one focused cluster in the Morbihan department, 2) a diffuse cluster in the area corresponding to the departments of Aube, Yonne, Loiret and Sarthe, and 3) another diffuse cluster in the area corresponding to the departments of Indre, Creuse, Haute-Vienne, Vienne and Indre-et-Loire.

Assumptions were postulated to explain these clusters, which were a very high animal density for the Morbihan, and the non-specialisation in dairy production for the other areas. The model only included some known variation factors for ASCS, but new factors could easily be included in the model if available. The Hellinger distance could be a useful notion in spatial cluster detection in epidemiology as it can manage continuous variables as well as qualitative ones.

Keywords: mastitis, cattle-diseases, epidemiology, spatial analysis, clustering, France

Introduction

Mastitis is a key pathological problem in dairy cattle. The economic impact of the disease is heavy, including reduced milk yield and quality, early culling and drug costs (Coulon et al. 2002; Seegers et al. 2003). The health status of the udder is reliably approached by the determination of milk somatic cell count (SCC), which is an internationally recognised indicator for subclinical mastitis control (Reneau 1986; Harmon 1994). SCC variation factors have been widely studied for a long time. A lot of factors have been associated with SCC through observational studies (Barkema et al. 1998; Barnouin et al. 1999; Busato et al. 2000; Barnouin et al. 2004). They have mainly concerned cow characteristics (breed, parity), dairy management practices (milking, udder hygiene, housing system, calving conditions), seasonal factors (climate) and their interactions.

In addition to knowledge of general risk factors, it could be relevant for better SCC control in herds to highlight through spatial studies local singularities playing a role in SCC variation. Such studies can be performed in France, as SCC values are routinely determined on a national basis by the Dairy Herd Improvement Association (DHIA) using a standard method. Regional differences in somatic cell score (SCS) have been shown previously in the United States (Ely et al. 2003). Nevertheless, research on spatial SCC or SCS features is still uncommon in epidemiology, although it can assist in identifying the reasons for specific foci of subclinical mastitis by suggesting hypotheses. Concerning other diseases, there are a large number of studies on the development and use of methods for identifying and analysing spatial clusters (Elliott et al. 2000; Carpenter 2001). The first step in this type of approach is to adopt a definition for a cluster (Elliott and Wakefield 2001; Wartenberg 2001). The main methods used to detect spatial clusters rely on the notion of nearest neighbour and of the scan statistic. The statistical tests based on the nearest neighbour notion are:

- the k-nearest neighbour test, which computes the mean distance between neighbours (Norstrom et al. 1999; Perez et al. 2002; Olea-Popelka et al. 2003)
- the Cuzick and Edwards'test, which measures the number of cases nearer to a specific case than the k^{th} nearest control (Perez et al. 2002)

The scan statistic test compares the observed number of cases inside a moving window to the expected number of cases under a Poisson or a Bernouilli hypothesis, and is based on the maximum likelihood ratio statistic across all windows (Kulldorff 1997). The scan statistic, which is widely used in epidemiology (Norstrom et al. 1999; Fevre et al. 2001; Hanson and Wieczorek 2002; Perez et al. 2002; Thomas and Carlin 2003; Brooker et al. 2004; Gangnon and Clayton 2004), aims to identify the most probable cluster(s).

All the reviewed methods of spatial cluster detection deal with binary variables (cases and controls), and are not suitable for the detection of clusters concerning continuous variables such as SCC. Consequently, the contribution of this paper is to propose a spatial analysis method based on the Hellinger distance between distributions to explore spatial patterns and detect geographical clusters of continuous epidemiological variables. The method was applied to herd milk somatic cell score (SCS) at a national scale in France.

Material and methods

Data

The statistical unit for the entire study was the herd-year. Annual herd SCS (ASCS) was the outcome variable. For each selected herd, ASCS was computed as the arithmetic mean of all monthly SCS values determined during 1996, according to Ali and Shook (1980):

$$SCS = \log_2\left(SCC/100{,}000\right) + 3$$

ASCS were calculated for 1996, the first year for which all monthly SCC values were available through the national DHIA database. To check the value of the proposed method of cluster detection, a random sample of 5,000 Holstein herds with more than 20 cows was selected from the DHIA database, representing around 15% of the total number of eligible herds (n=33,890). In fact two-thirds of French dairy cows are Holstein, and breed is a key variation factor for SCS (Barnouin et al. 1999).

The explanatory herd variables were mean parity, numbers of calvings per season, herd size, farm altitude and geographic coordinates. Calving seasons were classified according to Huffman et al. (1984) as spring (March to May), summer (June to August), autumn (September to November) and winter (December to February).

Cluster definition

A cluster was defined as a geographically bound group of close values being unlikely to have occurred at random, after taking account of known and available risk factors.

Statistical analysis

Statistical procedures were conducted using R 2.0.1 (R Development Core Team, 2004, R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0).

Descriptive analysis

As a first step, a descriptive analysis of the study variables was carried out. Then, the spatial patterns were explored via mapping representations. The interpolation technique of kernel smoothing (Silverman 1986) was used to represent local farm density (Formula 1), as well as variable intensity (Formula 2).

$$\hat{p}(x) = \sum_{i \in I} \frac{1}{h^2} k\left(\frac{x_i - x}{h}\right) \quad (1)$$

$$\hat{z}(x) = \frac{\sum_{i \in I} z_i k\left(\frac{x_i - x}{h}\right)}{\sum_{i \in I} k\left(\frac{x_i - x}{h}\right)} \quad (2)$$

$\hat{p}(x)$: estimated farm density at location *x* in a spatial domain D (France)

$\hat{z}(x)$: estimated variable intensity at location *x* in a spatial domain D (France)

I : farm set under observation z_i : value of the variable *z* (e.g. ASCS) of the farm *i*

h : bandwidth, which stands for width of the local window x_i : spatial coordinates of the farm i

k : kernel function (the Gaussian density was chosen here)

x : any spatial coordinates of a point in the space where density or intensity is to be calculated

As a second step, assuming ASCS as a continuous random field since the size of the sampling sites was large, data exploration was completed by spatial correlation aspects. The variogram (Cressie 1991) of the ASCS and its confidence bounds calculated by permutation procedure were carried out till a distance of about 600 km consistent with sufficient data. The permutation procedure is a resampling method that allows quantifying uncertainty by calculating confidence intervals. The permutations randomly assign the existing values (here the ASCS) over the geographical points (Manly 1991). If the computed value of the observations is beyond the bounds of the confidence interval, the null hypothesis is rejected. The null hypothesis in the present study stated that the ASCS were equally distributed and spatially independent.

Model fitting to known risk factors

A linear regression model of ASCS was performed, to take into account several available risk factors W_k (mean parity, numbers of calvings in spring, summer, autumn and winter, herd size, farm altitude) as follows:

$$ASCS = \beta_0 + \sum_k \beta_k W_k + \varepsilon \quad (3)$$

where $\beta = (\beta_k)$ denotes the mean value and the parameters for each risk factor, and ε the residuals of the model. The residuals could be considered as the unexplained part of the ASCS (without the effects of the integrated risk factors and the mean value). At this stage, the residuals were assumed to be Gaussian, equally distributed and independent. So the adequation of the model was assessed through a Kolmogorov Smirnov test of the Gaussianity of the residuals, and through a plot of residuals vs. predicted values to check homoscedasticity (Draper and Smith 1996).

As regards the independent assumption of residuals, the same descriptive analysis as for the rough ASCS was performed, to point out spatial features such as aggregation of high or low values.

Cluster detection via Hellinger distance based statistics

The method described below was applied to the residuals ε of the ASCS modelling (Formula 3) to detect clusters not related to available risk factors. The aim was to test whether the distribution of residuals was spatially neutral, or if there were localised foci of high or low values. To achieve this aim, the density $\overline{p}_z$ of farm location x_i whose residuals values were $\varepsilon_i \geq z$ (any given residual level *z*) was compared to the one at the initial or reference level z_0 (farms density $\overline{p}_0$). The Hellinger distance between 2 probability distributions $\overline{p}_z$ and $\overline{p}_0$ was used:

$$Hel(z) = \frac{1}{\sqrt{2}}\left(\int_D \left(\sqrt{\overline{p}_0(x)} - \sqrt{\overline{p}_z(x)}\right)^2 dx \right)^{\frac{1}{2}} \cong \frac{1}{\sqrt{2}}\left(\sum_{x_g \in G_D} \left(\sqrt{\overline{p}_0(x_g)} - \sqrt{\overline{p}_z(x_g)}\right)^2 \right)^{\frac{1}{2}} \cdot \Delta x \quad (4)$$

In this case, $\overline{p}_0(x)$ and $\overline{p}_z(x)$ were the normalised version of point densities calculated by kernel smoothing (Formula 1). The last term of (4) refers to an approximation of the integral, calculated on a regular grid G_D with cells of area Δx .

The Hellinger distance is a metric (symmetric, fulfils the triangle inequality property), which assumes bounded values in [0,1], and allows the addition of information (Gibbs and Su 2002).

To have a better estimation *Hel*(*z*) of the distance of the theoretical densities, good estimations of $\overline{p}_0(x)$ and $\overline{p}_z(x)$ were required. It is known that the use of a global bandwidth *h* in formulas 1 and 2 may give poor results when the density is too irregular or too far from homogeneity. To overcome this problem, one approach called variable bandwidth kernel estimations consists in taking locally bandwidth according to the point density value. Another way was chosen to overcome the difficulty, by using a homogenisation technique of point distribution before calculation. Therefore, a radial deformation of the space was performed. A reference point *g* in the space (gravity centre), assumed to be 0, was chosen and then, whatever the point *x* in the space, a homothetic transformation of x to $\Phi(x)$ along the line $[0, x]$ was computed as follows:

$$\Phi(x) = x\varphi(x)$$

$$\varphi(x) = x\left(2\int_0^1 \hat{p}(sx)\, s\, ds \right)^{1/2} \quad (5)$$

Function $\varphi(x)$ integrated the point density along segment $[0, x]$, and thus performed a directional dilatation or contraction of the space with a running parameter *s* from 0 to 1 (Senoussi et al. 2000):

The farm pattern became homogenous with a constant density, and so did the efficiency of the kernel estimator.

To test the significance of the null hypothesis H_0 "the ε are independently distributed" against H_1 "the ε are spatially structured", a permutation procedure was applied to the following global Hellinger distance statistic $Hel = \int Hel(z)dz \cong \sum_l Hel(z_l)$ (where *l* was the number of levels of *z* chosen). Significantly higher value of *Hel* than the one of the confidence bound carried out by the permutations meant that the density of points was not spatially retained for the different levels of *z*. Consequently H_0 could be rejected, and it could be concluded that there was a non homogenous ε density, i.e. that some areas presented clusters (of high or low values), while some others were emptier than expected by random. This allowed testing the global presence of clusters for the whole domain and all the ε values.

Following this global testing, the Hellinger distance was mapped at different levels *z*, to identify significant levels and then significant areas. The maps of farm density performed for levels of *z* where *Hel*(*z*) revealed a significant discrepancy with the farm density at the initial level were used. The areas and respective farms on the map were determined. Moreover, since the discrepancy was easier to notice if the initial level was homogenous, the deformed maps were used to visually identify at which location there had been a cluster formation on the deformed map. Finally using the original map of France, it was determined which farms were involved.

Results

Herd statistical description

In the studied herds, mean ASCS was 3.16 (± 0.57), mean herd size was 40 cows (± 16) and mean parity was 2.6 (± 0.32). Moreover, it was observed that of all calvings, 13% occurred in spring, 22% in summer, 41% in autumn, and 24% in winter.

The geographic distribution of the farms is illustrated in figures 1 and 2: Figure 1 represents farm locations, while Figure 2, which is an interpolation by a kernel smoothing technique, displays the farm density. Clearly, the farm density was spatially non-homogenous, as the majority of herds were located in the main agricultural area in France, i.e. the north-west region (61% of the farms). By contrast, the south-east region, which corresponded to the Mediterranean area, had a very low farm density (6% of the farms).

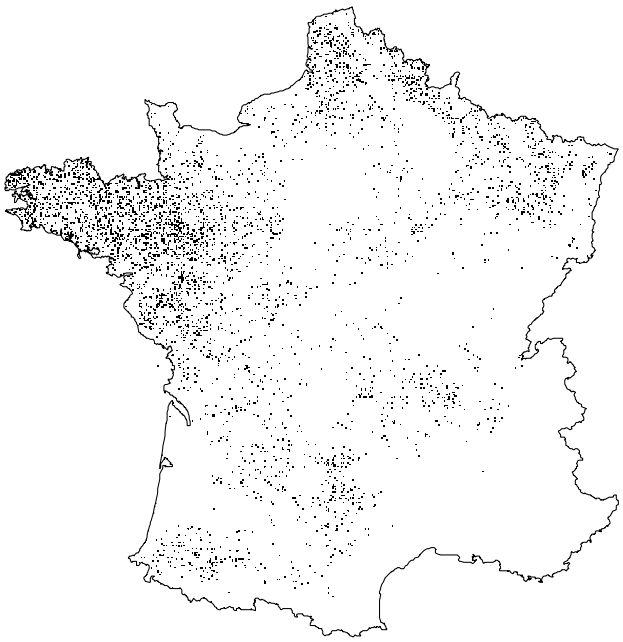


Figure 1: Geographic distribution of the study population (Holstein dairy herds in France, n=5000)

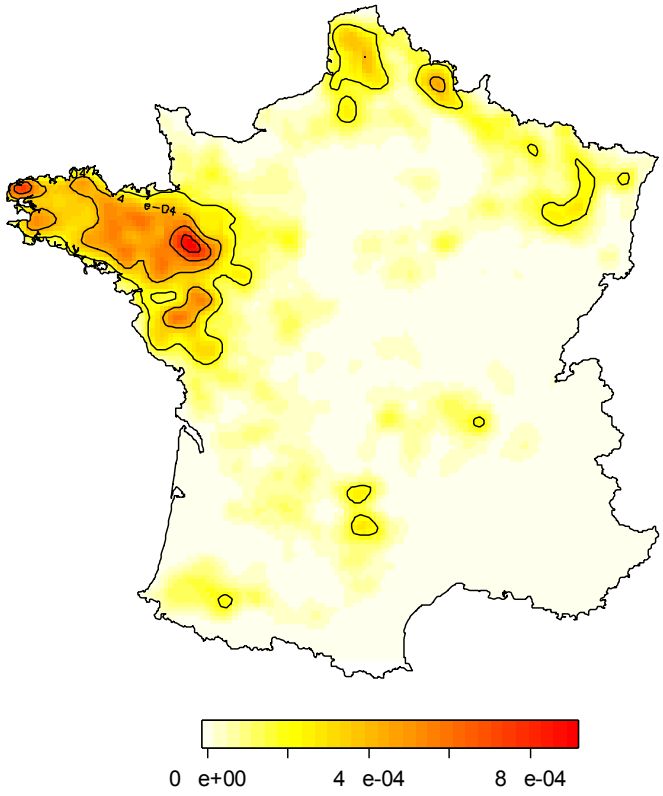


Figure 2: Kernel estimation of density of Holstein farms involved in the study (Holstein dairy herds in France, n=5000)

Figure 3, which represents maps of ASCS intensity, shows that the centre-north region had high ASCS, around 3.5 to 4. The high value observed in the Alps, in the south-east region, corresponded to the highest smoothed value, but was due to the lack of points in the region (only 1 point on this bandwidth region).

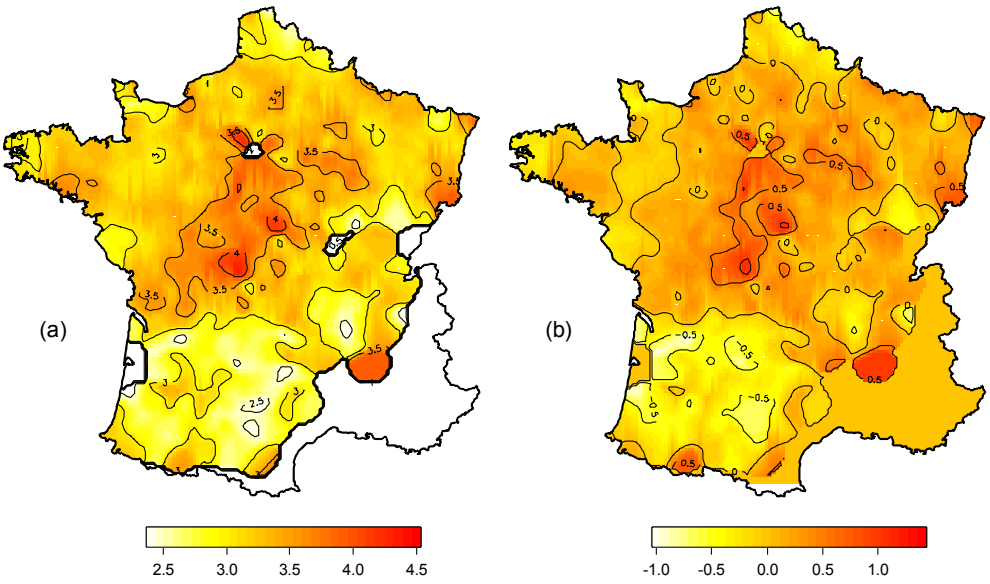


Figure 3: Variable intensity in the study population of Holstein dairy farms in France: (a) ASCS and (b) residuals of ASCS modelling

The variogram of the ASCS and the confidence bounds (50 steps, 1000 permutations, α=1%) on Figure 4 shows that for small distances (≤ 200 km), there were spatial correlations between the ASCS. The behaviour of the variogram near 0 pointed out a nugget effect, suggesting a high noise, of about 0.26/0.035 ≈ 3/4 of the total variability. The spatial structure part represented 0.09/0.35 ≈ 1/4 of the variability.

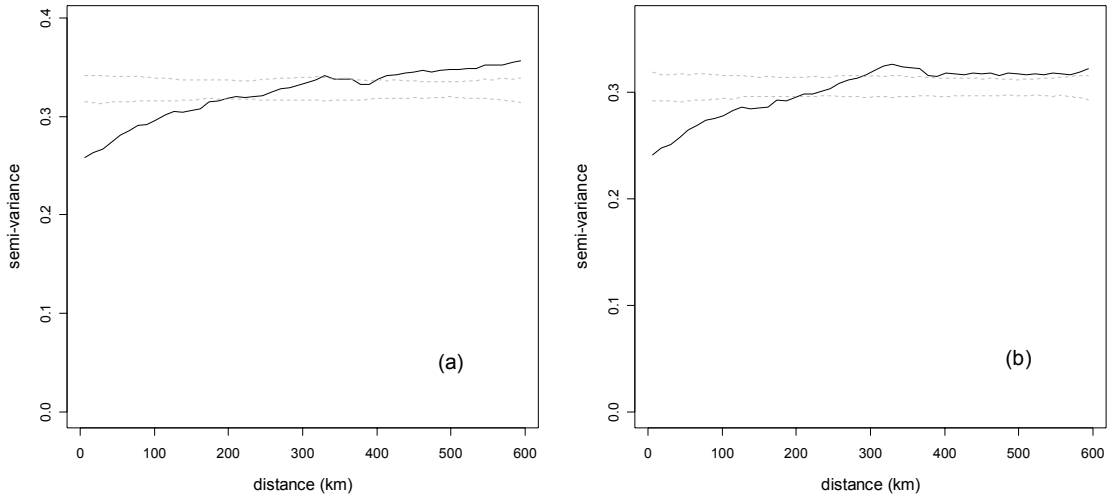


Figure 4: Variograms with confidence bounds of (a) the ASCS and (b) the residuals of ASCS modelling

Covariable effects and residuals

The linear regression model showed (Table 1) that ASCS was significantly raised by increasing mean parity, increasing calving numbers in spring and winter, and increasing farm altitude, while the herd size had no significant influence. The R^2 was low (6.6%), but the model was globally very significant. The hypothesis of normality of residual distribution was not rejected by the Kolmogorov Smirnov test (p=0.6599). Moreover, the symmetry of the plot of the residuals along the y-coordinate of predicted values assessed the homoscedasticity hypothesis.

Table 1: Linear regression model of the ASCS (annual somatic cell score) in the study population of Holstein dairy herds in France (n = 5000).

Variable	Coefficient estimation	Standard deviation	p value	
mean parity	2.83E-01	2.51E-02	<2E-16	***
summer calving	-9.91E-04	1.50E-03	0.51	
autumn calving	-2.37E-03	1.27E-03	0.06	
winter calving	7.75E-03	1.76E-03	1.10E-05	***
spring calving	1.13E-02	2.14E-03	1.31E-07	***
herd size	8.40E-03	1.43E-02	0.55	
farm altitude	-4.83E-04	5.04E-05	<2E-16	***
adjusted R^2	0.066			
global p value	<2E-16	***		

* : p<0.05

** : p<0.01

*** : p<0.001

The range of residuals was -1.84 to 2.28. The map of intensity of residuals is illustrated in Figure 3. The centre-north region remained an area with relatively high values, but the effect appeared somewhat muted. Figure 4, the variogram of residuals, shows the same behaviour as the ASCS one.

Cluster detection

The result of the spatial deformation is displayed in Figure 5.

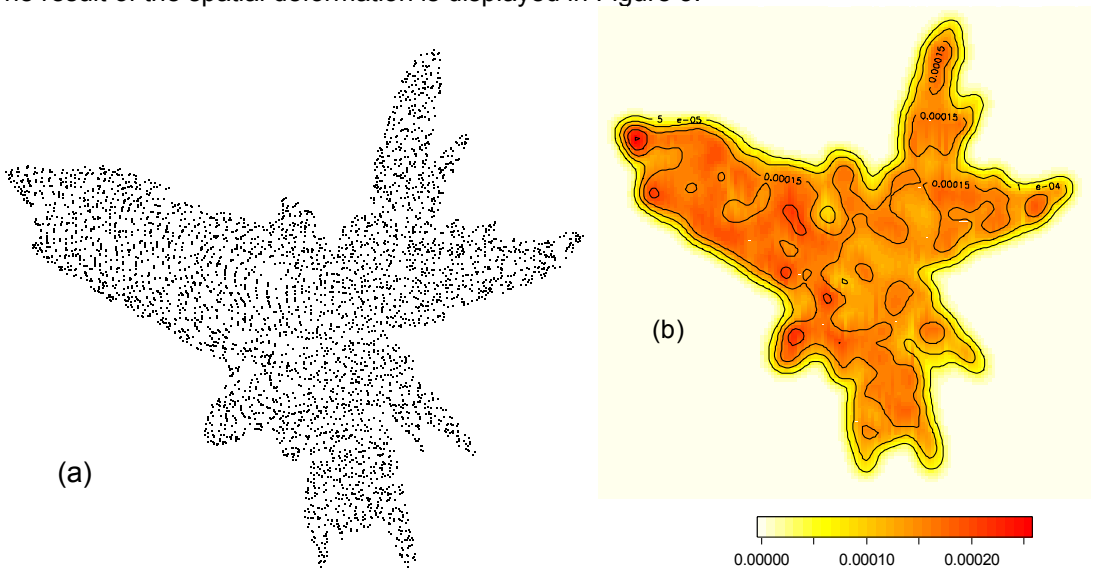


Figure 5: Geographic distribution after deformation of (a) farms and (b) farm density (Holstein dairy herds in France, n=5000)

The global Hellinger distance statistic *Hel* was 3.10. The 5% confidence bounds calculated by permutations were [1.95, 2.70]. As the global Hellinger distance value was beyond the bounds, one could not reject the hypothesis of a significant cluster formation of the ASCS residuals.

Figure 6 shows the computation of the Hellinger distance for 15 residuals levels and 1000 permutations to build the confidence bounds at α=5%. From a level of residuals of -0.5 to a level of 1.2, there was a significant difference from configuration of values distributed at random. This represented 96% of the total number of points.

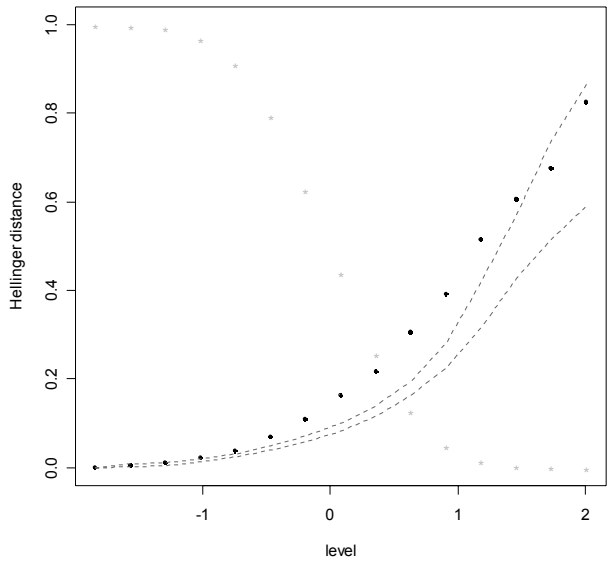


Figure 6: Hellinger distance for each level of ASCS residuals, with permutations in space
- - - : confidence intervals obtained by permutations of the data (1000 permutations, α = 5%)
** : survival curve of ASCS residuals*

Mapping of farm density for the different significant levels of cluster formation displayed the same trend, i.e. a progressive concentration on 3 clusters. Figure 7 is the map for the level of 0.36, at which the 3 areas were clearly isolated and the number of points was considered sufficient for calculation.

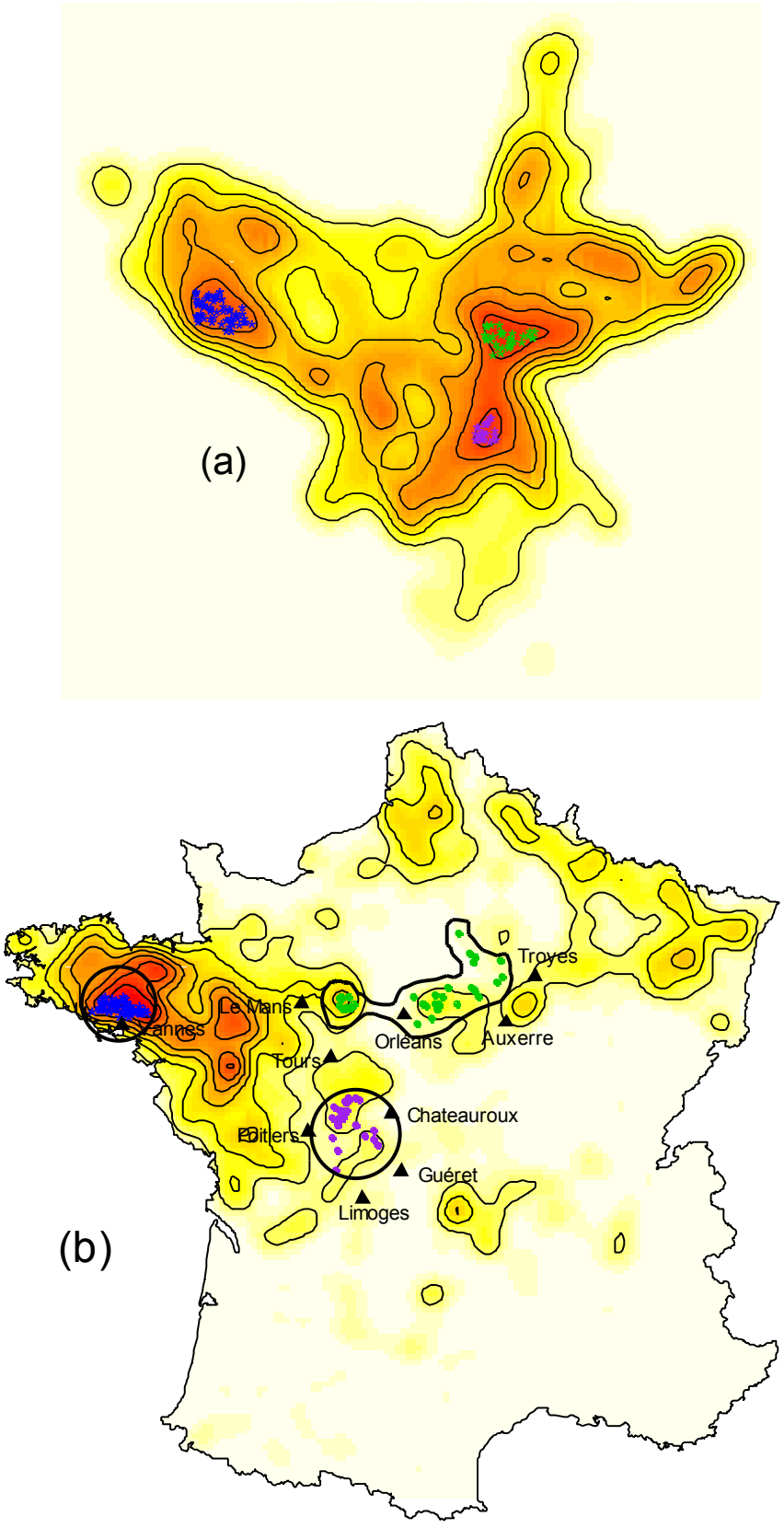


Figure 7: Localisation of clusters of ASCS residuals (level 0.36): (a) deformed map and corresponding points on (b) original map including main cities near the significant clusters

Finally, locating the farms for the 3 clusters on the original map identified 3 areas:

- the Morbihan department, with a focused cluster above Vannes
- an area corresponding to the departments of Aube, Yonne, Loiret, and Sarthe with a diffuse cluster around Orléans
- an area corresponding to the departments of Indre, Creuse, Haute-Vienne, Vienne and Indre-et-Loire, with another diffuse cluster surrounded by Chateauroux, Guéret, Limoges and Poitiers

Discussion

The present work identified 3 geographical clusters that were not perceptible directly from the map of ASCS intensity. To satisfy this study's definition of a cluster, a model for ASCS was first built to integrate covariable effects. The results of the linear regression were consistent with previously published results (Laevens et al. 1997; Barnouin et al. 2004). But a great number of known risk factors for ASCS variation were not available in the present study. Particularly, hygienic conditions and climatic parameters, which are important to consider, were not included in the model. Consequently, the remaining clusters could be linked to some of the missing factors. For the 3 clusters detected, the following explanations are suggested:

- the cluster in Morbihan could be linked to the very high farm density of this area and is consistent with the behaviour of the variogram at small distances. High density could favour microbial spread between the herds and involve less available DHIA advisors.
- the 2 other clusters of the centre-north region could be due to the non specialisation in dairy production of these areas (Barnouin et al. 2004), which are focused on bovine and ovine meat, and on cereal production.

If data corresponding to hypothesised reasons for clusters become available later, they could be included in the model. A new cluster detection could then be performed to confirm or deny the assumptions. The method can as well be used to detect clusters of low values, to highlight regions with very good control of ASCS.

It is important to distinguish a time-permanent cluster from an ephemeral cluster, because they point to different assumptions. In the present study, the year 1996 was analysed. The tendencies detected have to be compared with those of subsequent years, as the proposed method can be extended to time, and even to space and time detection. This would help better control of herd mastitis risk, and the identification of areas such as:

- areas in which new pathogens or deleterious strains of customary pathogens could emerge
- areas experiencing specific environmental conditions unfavourable to udder health

The method evolved by the present study to detect clusters has many advantages. First of all, the method applies to continuous variables, a possibility which was lacking in spatial statistics. Such an option is useful because even for case/control variables, model residuals are continuous. Thus, the proposed methodology based on the Hellinger distance can be used both on rough variables and after modelling, as in this study, to take into account covariables as often needed in studies (Timander and McLafferty 1998). The cluster detection method used can be equally applied both to binary variables (only 2 levels instead of several), and more generally to qualitative data (1 level per state).

Secondly, the method proposed does not need an *a priori* precision of either the number or the position of clusters. At a given level of any variable of interest, the statistic relies on the excess of points observed compared to values randomly distributed, and so it is not defined by a maximum number of cases. Therefore, a cluster is not always detected, as the highest SCC must occur somewhere (Waller 2000). The Hellinger distance is a true measure to compare distributions, which takes into account the underlying non-homogeneous population density. Coupled with a space deformation, the cluster detection based on the Hellinger distance allows detection even for low farm density areas.

On the other hand, the Hellinger distance cannot deal with isolated extreme values, because the number of points becomes insufficient. Therefore, the statistic based on the Hellinger distance seems more appropriate to investigate diseases not focused on maximum values, as was the present case. In the same way, its computation needs density estimation, and consequently requires a sufficiently global number of points.

At this stage of development of the method, using Hellinger distance based statistics via permutation tests is conditional on observations for statistical significance. Further modelling may be essential to elaborate more appropriate tests for specified hypotheses.

To summarise, there is a wide range of methods to detect clusters, and there are approximately as many cluster definitions and criteria as there are methods. For a given problem of cluster detection, it is necessary to explore the underlying hypothesis with precision, and to be sure that the chosen method accurately responds the biological question. This study's method, based on the Hellinger distance, can be applied to other diseases and to other variables such as biological markers, as it manages continuous variables, as soon as there is a sufficient sample of points. The comparison of spatial patterns of 2 diseases in the same population can be another interesting usage.

Conclusion

A new method of cluster detection based on the Hellinger distance allowed the highlighting of 3 clusters of high ASCS values in France. One of these clusters, corresponding to the Morbihan department, was very focused, while the 2 others, which were located in the centre-north part of France, were more geographically diffuse. In conclusion, the Hellinger distance coupled with a space deformation technique appears to be a useful tool in cluster detection that can manage continuous variables as well as qualitative ones.

References

Ali, A. K. A. and Shook, G. E., 1980. An optimum transformation for somatic cell concentration in milk. J Dairy Sci. 63(3), 487-490.

Barkema, H. W., Schukken, Y. H., Lam, T. J. G. M., Beiboer, M. L., Benedictus, G. and Brand, A., 1998. Management practices associated with low, medium, and high somatic cell counts in bulk milk. J Dairy Sci. 81(7), 1917-1927.

Barnouin, J., Chassagne, M., Bazin, S. and Boichard, D., 2004. Management practices from questionnaire surveys in herds with very low somatic cell score through a national mastitis program in France. J Dairy Sci. 87(11), 3989-3999.

Barnouin, J., Geromegnace, N., Chassagne, M., Dorr, N. and Sabatier, P., 1999. Facteurs structurels de variation des niveaux de comptage cellulaire du lait et de fréquence des mammites cliniques dans 560 élevages bovins répartis dans 21 départements français. Anim Prod. 12, 39-48.

Brooker, S., Clarke, S., Njagi, J. K., Polack, S., Mugo, B., Estambale, B., Muchiri, E., Magnussen, P. and Cox, J., 2004. Spatial clustering of malaria and associated risk factors during an epidemic in a highland area of western Kenya. Trop Med Int Health. 9(7), 757-766.

Busato, A., Trachsel, P., Schallibaum, M. and Blum, J. W., 2000. Udder health and risk factors for subclinical mastitis in organic dairy farms in Switzerland. Prev Vet Med. 44(3-4), 205-220.

Carpenter, T. E., 2001. Methods to investigate spatial and temporal clustering in veterinary epidemiology. Prev Vet Med. 48(4), 303-320.

Coulon, J. B., Gasqui, P., Barnouin, J., Ollier, A., Pradel, P. and Pomies, D., 2002. Effect of mastitis and related-germ on milk yield and composition during naturally-occurring udder infections in dairy cows. Anim Res. 51(5), 383-393.

Cressie, N. A. C., 1991. Geostatistics. In: Barnett V, Bradley RA, Fisher NIet al (eds), Statistics for spatial data. John Wiley and Sons, New York, pp. 58-67.

Draper, N. R. and Smith, H., 1996. Applied regression analysis. John Willey & Sons, New York.

Elliott, P. and Wakefield, J., 2001. Disease clusters: should they be investigated, and, if so, when and how? J. R. Statist. Soc. A. 164(part 1), 3-12.

Elliott, P., Wakefield, J., Best, N. and Briggs, D., 2000. Clustering, cluster detection and spatial variation in risk. In: Elliott P, Wakefield JC, Best NG and Briggs D (eds), Spatial epidemiology: methods and applications. Oxford University Press, New York, pp. 128-152.

Ely, L. O., Smith, J. W. and Oleggini, G. H., 2003. Regional production differences. J Dairy Sci. 86, E28-E34.

Fevre, E. M., Coleman, P. G., Odiit, M., Magona, J. W., Welburn, S. C. and Woolhouse, M. E., 2001. The origins of a new Trypanosoma brucei rhodesiense sleeping sickness outbreak in eastern Uganda. Lancet. 358(9282), 625-628.

Gangnon, R. E. and Clayton, M. K., 2004. Likelihood-based tests for localized spatial clustering of disease. Environmetrics. 15, 797-810.

Gibbs, A. L. and Su, F. E., 2002. On choosing and bounding probability metrics. Int Stat Rev. 70(3), 419-435.

Hanson, C. E. and Wieczorek, W. F., 2002. Alcohol mortality: a comparison of spatial clustering methods. Soc Sci Med. 55(5), 791-802.

Harmon, R. J., 1994. Physiology of mastitis and factors affecting somatic cell counts. J Dairy Sci. 77(7), 2103-2112.

Huffman, E. M., Mortimer, R., Olson, J. D., Ball, L. and Farin, P. W., 1984. Risk factors for prebreeding pyometra on four Colorado dairy farms. Prev Vet Med. 2(6), 785-790.

Kulldorff, M., 1997. A spatial scan statistic. Commun Stat-Theor M. 26(6), 1481-1496.

Laevens, H., Deluyker, H., Schukken, Y. H., De Meulemeester, L., Vandermeersch, R., De Muelenaere, E. and De Kruif, A., 1997. Influence of parity and stage of lactation on the somatic cell count in bacteriologically negative dairy cows. J Dairy Sci. 80(12), 3219-3226.

Manly, B. F. J., 1991. Randomization and Monte Carlo methods in biology. Chapman and Hall, London.

Norstrom, M., Pfeiffer, D. U. and Jarp, J., 1999. A space-time cluster investigation of an outbreak of acute respiratory disease in Norwegian cattle herds. Prev Vet Med. 47(1-2), 107-119.

Olea-Popelka, F. J., Griffin, J. M., Collins, J. D., McGrath, G. and Martin, S. W., 2003. Bovine tuberculosis in badgers in four areas in Ireland: does tuberculosis cluster? Prev Vet Med. 59(1-2), 103-111.

Perez, A. M., Ward, M. P., Torres, P. and Ritacco, V., 2002. Use of spatial statistics and monitoring data to identify clustering of bovine tuberculosis in Argentina. Prev Vet Med. 56(1), 63-74.

Reneau, J. K., 1986. Effective use of dairy herd improvement somatic cell counts in mastitis control. J Dairy Sci. 69(6), 1708-1720.

Seegers, H., Fourichon, C. and Beaudeau, F., 2003. Production effects related to mastitis and mastitis economics in dairy cattle herds. Vet Res. 34(5), 475-491.

Senoussi, R., Chadoeuf, J. and Allard, D., 2000. Weak homogenization of points processes by space deformations. Adv Appl Prob (SGSA). 32, 948-959.

Silverman, B. W., 1986. The kernel method for univariate data. In: Cox DR, Hinkley DV, Rubin D and Silverman BW (eds), Density estimation for statistics and data analysis. Chapman and Hall, London, pp. 34-94.

Thomas, A. and Carlin, B. P., 2003. Late detection of breast and colorectal cancer in Minnesota counties: an application of spatial smoothing and clustering. Stat Med. 22(1), 113-127.

Timander, L. M. and McLafferty, S., 1998. Breast cancer in West Islip, NY: a spatial clustering analysis with covariates. Soc Sci Med. 46(12), 1623-1635.

Waller, L. A., 2000. A civil action and statistical assessments of the spatial pattern of disease: do we have a cluster? Regul Toxicol Pharmacol. 32(2), 174-183.

Wartenberg, D., 2001. Investigating disease clusters: why, when and how? J. R. Statist. Soc. A. 164(part 1), 13-22.

III.3. Article n°2 : Application de la méthode au suivi annuel de SCS

Cet article, intitulé "Spatial and Temporal Patterns of Herd Somatic Cell Score in France", montre l'utilisation de la méthode de détection d'agrégats pour le suivi de l'évolution de la répartition spatiale des SCS. Les données sur les scores cellulaires d'un échantillon de 5000 élevages suivis sur 5 ans (années 1996 à 2000) permettent l'analyse des dynamiques spatiales et temporelles des infections mammaires subcliniques.

Cet article a été soumis à la revue Journal of Dairy Science et est en cours de révision.

Somatic cell score spatio-temporal patterns

Gay

Spatio-temporal patterns of somatic cell score were investigated in French dairy herds for the years 1996 to 2000. An original method allowed detection of regional singularities for mastitis risk. A significant spatial cluster formation was highlighted and 12 areas of interest were identified. The global temporal trend of a decreasing score may be the result of DHIA efforts to improve dairy herd management. Two progressive movements were detected: the disappearance of clusters in the north-west, and the increase of clusters in the south-west region.

Spatial and Temporal Patterns of Herd Somatic Cell Score in France

E. Gay,[1, 2] J. Barnouin[2] and R. Senoussi[1]

[1] Unité de Biométrie, INRA Domaine St-Paul,
Site Agroparc, 84914 Avignon Cedex 9, France
[2] Unité d'Epidémiologie Animale, INRA,
63122 St-Genès-Champanelle, France

Corresponding author:
Jacques Barnouin
Unité d'Epidémiologie Animale
INRA
63122 St-Genès-Champanelle
00 33 (0)4 73 62 42 61
Fax: 00 33 (0)4 73 62 45 48
barnouin@clermont.inra.fr

Abstract

Spatial and temporal patterns of annual milk somatic cell score (ASCS) were explored in French dairy herds between 1996 and 2000 to detect regional singularities for mastitis risk. An original cluster detection method was used, which was adapted to continuous variables, and allowed to take into account ASCS variation factors.

The statistical unit was the herd-year. A linear regression model for each year allowed adjustment for breed, mean parity, number of calvings for each season, herd size and farm altitude. Cluster detection was performed on residuals of the model through a method based on the Hellinger distance between spatial distributions. We computed the Hellinger distance between farm distributions at different levels of residuals. Temporal ASCS patterns were explored using a computation of correlations and comparisons between spatial structures of the different years.

Mean ASCS varied from 3.05 to 3.17. The general trend over the study period was a decrease. The global Hellinger distance, which was higher than what could have been expected by random for each of the 5 years, indicated that there was a significant spatial cluster formation. Cluster mapping over the 5 years identified 12 areas: St-Brieuc, Vannes, 2 areas around Angers, the centre-west along a line Niort-Tulle, the south-west around Toulouse, Amiens, Charleville-Mézières, the centre-north around Auxerre, Strasbourg, Vesoul and Bourg-en-Bresse. Temporal correlations between ASCS residuals for each year were positive and decreasing, and 1996 and 2000 appeared spatially different than the others.

The more affected areas were regions that were non-specialised in dairy production. During the study period, we detected 2 progressive movements: the disappearance of clusters in the north-west, and the increase of clusters in the south-west. The global trend of decrease may be the result of DHIA and dairy industry efforts to improve dairy herd management for better control of mastitis risk.

(**Key words:** spatial and temporal pattern, somatic cell score)

Abbreviation key: ASCS = annual somatic cell score, **DHIA** = Dairy Herd Improvement Association, **SCC** = somatic cell count, **SCS** = somatic cell score

Introduction

The health status of the bovine udder is reliably approached by the determination of milk somatic cell count (**SCC**), since the main factor causing an increase above SCC physiological level is inflammation due to infection (Reneau, 1986; Harmon, 1994). SCC or its logarithmic transformation the somatic cell score (**SCS**) ($SCS = \log_2(SCC/100{,}000) + 3$) (Ali and Shook, 1980) are internationally recognised indicators for subclinical mastitis control. Moreover, in many countries cow SCC (and therefore SCS) is available on a monthly basis through the Dairy Herd Improvement Associations (**DHIA**).

Risk and protection factors associated with SCS have been widely investigated through observational studies (Barkema et al., 1998; Busato et al., 2000; Barnouin et al., 2004) which highlighted the influence of cow characteristics (breed, parity), dairy management practices (milking, udder hygiene, housing system, calving conditions), seasonal factors (climate) and their interactions on SCS level.

Moreover, differences in natural resources, dairy farm structure and market caused different regions of a same country to implement different dairy management systems which may be reflected in different technical performance (Oleggini et al., 2001; Ely et al., 2003). Consequently, the assessment of regional differences in SCS through the study of spatial and temporal patterns may assist in formulating mastitis control and prevention programs and in improving milk quality. In the United States Ely et al. (2003) explored spatial variations of some key variables in dairy management, including SCS, and the differences between regions were noticeable. There are numerous studies on the use of methods for identifying and analysing spatial or temporal clusters of animal disease events (Ward and Carpenter, 2000; Carpenter, 2001). However these events are essentially concerned with dichotomic variables (e.g. cases/control), and not with continuous variables such as SCS.

To explore spatio-temporal clustering, it is first necessary to specify a cluster definition (Wartenberg, 2001). A cluster could be defined as a geographically and temporally bounded group of close values being unlikely to have occurred at random (Elliott and Wakefield, 2001). The interest very often focuses on unexplained clustering (Marshall, 1991), but the methods used rarely take into account known disease risk factors. Nevertheless a way to deal with this problem is to model the disease, and then to use the results to perform the cluster detection (Timander and McLafferty, 1998; Ahrens et al., 2001).

Once the cluster definition is specified, a method has to be chosen to detect clusters. Scan statistics are widely used with regards to disease events (Norstrom et al., 1999; Fevre et al., 2001; Perez et al., 2002). These methods compare the observed number of events inside a moving window to the expected number of events under a Poisson or a Bernoulli hypothesis. But such approaches, which deal only with binary variables, are not suitable for the detection of clusters among SCS.

The aim of this paper is to analyse the spatial and temporal patterns of SCS in France. We used an original cluster detection approach adapted to continuous variables, based on the Hellinger distance between spatial distributions and which take into account the known SCS variation factors.

Materials and methods

Data

The study population consisted of French dairy herds enrolled in DHIA between 1996 and 2000. The selected herds were pure-breed and had at least 20 cows. Following these criteria, the eligible herds amounted to 34,735. As a good compromise between the necessary representativeness of the sample and the computational facilities needed for a new approach, we selected a random sample representing 15% of the eligible herds, i.e. 5210 farms.

The statistical unit was the herd-year. The outcome variable was the annual herd SCS (**ASCS**), which was computed as the arithmetic mean of all monthly cow SCS values determined in a herd during a calendar year. The other variables were breed, mean parity, number of calvings per season, and herd size. The farm geographic coordinates and altitude were obtained via the French National Institute of Statistics and Economic Studies. Calving seasons were classified according to Huffman et al. (1984) as spring (March to May), summer (June to August), autumn (September to November) and winter (December to February).

Statistical Analysis

Statistical procedures were conducted using R 2.0.1 (R Development Core Team, 2004, R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0).

After a descriptive analysis of the study variables, we used mapping representations to explore the spatial and temporal patterns of the data. When a spatial interpolation was needed to represent local farm density and ASCS intensity, we used kernel smoothing (Silverman, 1986).

As a second analysis step, we performed a linear regression model of ASCS for each year, to take into account the available risk factors (breed, mean parity, number of calvings in spring, summer, autumn and winter, herd size and farm altitude). Correlations between covariables were computed before modelling to ensure that the factors included were independent. The model assumptions were assessed through a Kolmogorov Smirnov test of the Gaussianity of the residuals, and a plot of residuals vs. predicted values to check homoscedasticity (Draper and Smith, 1996). The residuals could be considered as the unexplained part of the ASCS (following removal of the effects of the integrated risk factors and the mean value), and became the variable of interest in the follow-up of the study in order to focus on unexplained clustering.

The aim of the third analysis step, the cluster detection, was to test whether the residuals were spatially randomly distributed, or if there were localised foci of high values.

The heterogeneity of the spatial distribution of farms could lead to a poor estimation of the farm densities and so make it difficult to visualise the foci (many high values would be likely if there were many farms in the area) (Waller, 2000). To overcome this difficulty, we used a homogenisation technique of point distribution before calculation via a radial deformation of the space. It allowed locally dilating or contracting the space in accordance with the number of points, without changing the neighbourhood (Senoussi et al., 2000). Consequently the farm pattern became homogenous with a nearly constant density.

Then for any given residual level z_j we compared the normalised density $\overline{p}_{z_j}$ of the n_j farm locations x_i whose values of the residuals were $\varepsilon_i \geq z_j$ to the farm density $\overline{p}$ via the Hellinger distance in the spatial domain D (Gibbs and Su, 2002):

$$Hel(z_j) = \sqrt{\frac{1}{2}\int_D \left(\sqrt{\overline{p}(x)} - \sqrt{\overline{p}_{z_j}(x)}\right)^2 dx}$$

The integral was approximated by calculation on a regular grid. This formula gave a Hellinger distance value for each residual level (the number of levels dividing the range of residuals had to be chosen). The global Hellinger distance (*Hel*) was defined as the sum of those different values weighted by the farm count at each level (j levels):

$$Hel = \sum_j Hel(z_j)\frac{n_{j-1} - n_j}{n}$$

Under the null hypothesis “the residuals are independently distributed in space”, the point distribution at any level of residuals should have been close to the initial farm distribution, and therefore the Hellinger distance should be low. To test this hypothesis of complete spatial randomness, we used spatial permutations. The permutation procedure randomly assigns the existing values (here the ASCS for a specified year) over the geographical points (Manly, 1991). Significantly higher value of the global Hellinger distance than the ones of the confidence bounds carried out by the permutations allowed the rejection of the null hypothesis. Consequently the residuals were considered as spatially structured, and some areas presented clusters while some others were emptier than expected by random.

Following this global testing, the Hellinger distance was mapped at the different levels of the residuals to identify significant levels and areas. To select the farms involved in areas of concentration and report them on the original map, we chose a level at which the clusters were clearly defined and where the point number was sufficient for calculation (at least 10% of the total number of points).

In a last analysis step, we explored the temporal stability of spatial patterns by 2 complementary approaches. The first one was descriptive; it computed the temporal correlations between residuals of each year. The second one was inferential and used temporal permutations techniques: a value was randomly assigned to each point among the 5 possible ASCS residuals of the 5 years and the global Hellinger distance was calculated. This allowed testing temporal singularities of a particular year compared to the whole 5 year period.

Results

Descriptive Analysis

Eighty one percent of herds had Holstein cows, while 13% had Montbéliarde cows, 5% Normande cows, and 1% other breeds. The mean farm altitude was 226m (±244). The farm characteristics were stable over the 5 years, except for the ASCS which decreased over the final 2 years (1999 and 2000) (Table 1). According to the year, ASCS varied from 3.05 to 3.17, which corresponded to 253,000 to 262,000 cells/mL.

Table 1: Characteristics of the sample of French dairy herds from 1996 to 2000 (n=5210)

	1996	1997	1998	1999	2000
Annual milk cellular score	3.12 (0.59)[1]	3.17 (0.58)	3.14 (0.57)	3.07 (0.55)	3.05 (0.57)
Herd size	40.2 (16.7)	40.5 (16.7)	39.9 (16.6)	40.1 (16.7)	39.9 (16.6)
Mean parity	2.62 (0.33)	2.61 (0.33)	2.56 (0.32)	2.49 (0.32)	2.46 (0.31)
Spring calving (%)	13.1	12.9	13.3	12.9	13.6
Summer calving (%)	21.9	21.7	21.3	21.6	22.6
Autumn calving (%)	40.9	41.2	41.2	40.8	39.6
Winter calving (%)	24.1	24.2	24.2	24.7	24.2

[1]mean (standard deviation)

The geographic distribution of the farms is illustrated in Figure 1 where (a) represents the location of the selected farms, and (b) farm density obtained by a kernel smoothing technique. The farm density was non-homogeneous in space, and there were 3 areas with higher densities: 1) the north-west region (Brittany, Normandy and Pays de Loire) which is the main dairy area in France, 2) the north region, and 3) the centre-east region. By contrast, the south-east region, i.e. the Mediterranean area, had very few dairy farms.

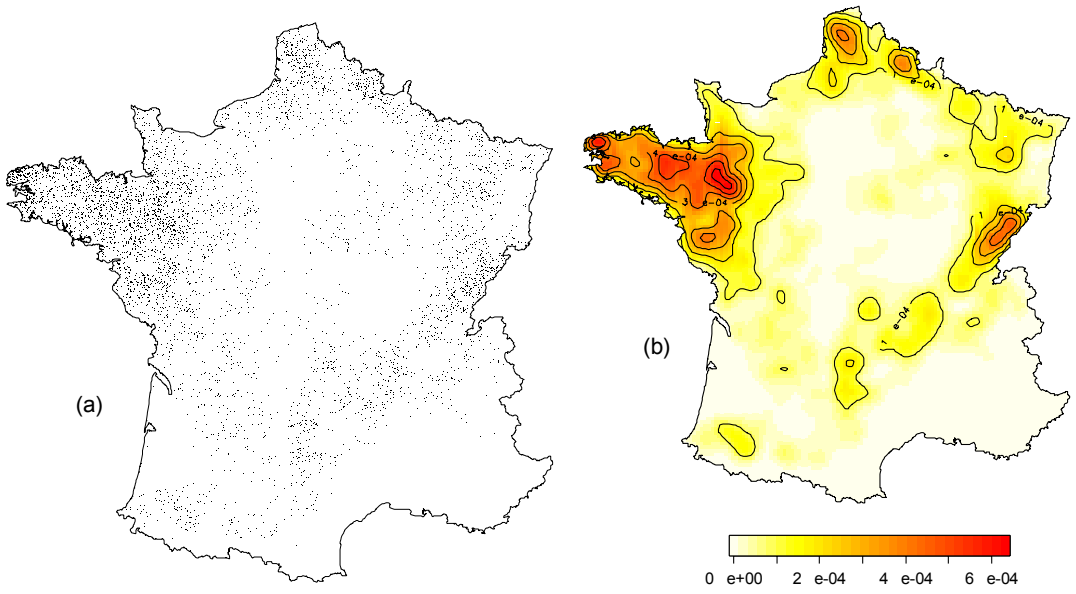


Figure 1: Spatial distribution of the studied population (dairy herds in France, n=5210): (a) farm locations and (b) farm density

The maps of ASCS spatial distribution (Figure 2) showed that during the whole study period, the centre-north region had relatively high ASCS intensities of around 3.5 to 4. Nevertheless during the last 2 years this area showed a notable reduction. Some spots with higher values than the surroundings appeared from 1998 in the south-west region. The high value observed in the department of Pyrénées Orientales, at the border of the Mediterranean area, was due to a single point as seen in Figure 1(a). This appeared to be an artefact of general smoothing technique and could be corrected via a homogenisation method.

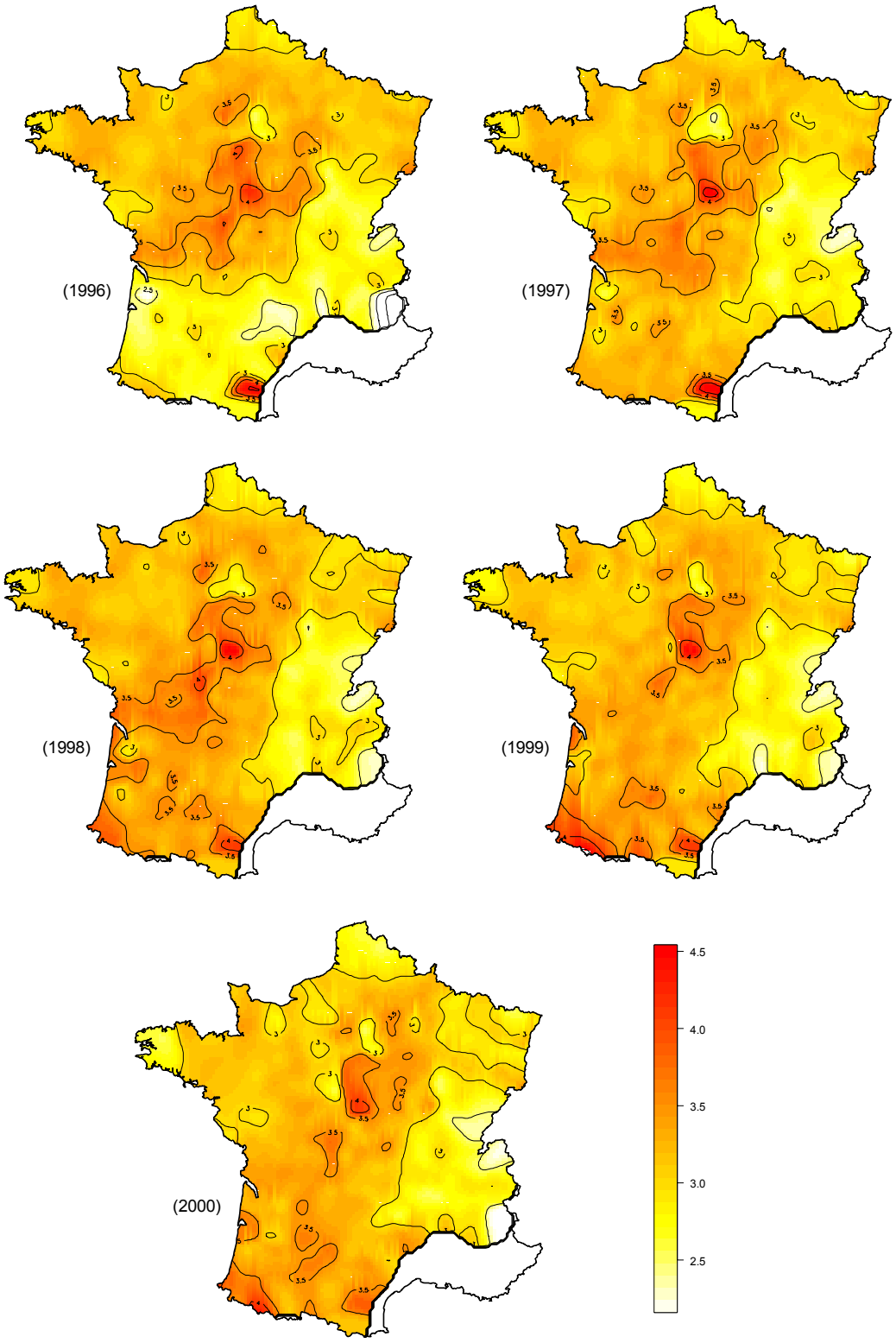


Figure 2: Annual somatic cell score intensity in the studied population of dairy farms in France (n=5210, 1996 to 2000)

The map of farms experiencing 4 successive increases or decreases during the study period is presented in Figure 3. There were 106 farms with 4 consecutive increases (2.0% of the farms), and 200 with 4 consecutive decreases (3.8%of the farms). The general trend was a decrease, as already seen with the ASCS in Table 1, which was lower in the last 2 years. But the concentration of farms with 4 decreasing ASCS was higher in western Brittany (Finistère department) and along the border of the centre-east region (Doubs department). By contrast, in the south-west region, numerous farms had 4 successive increases, while no farms showed 4 successive decreases.

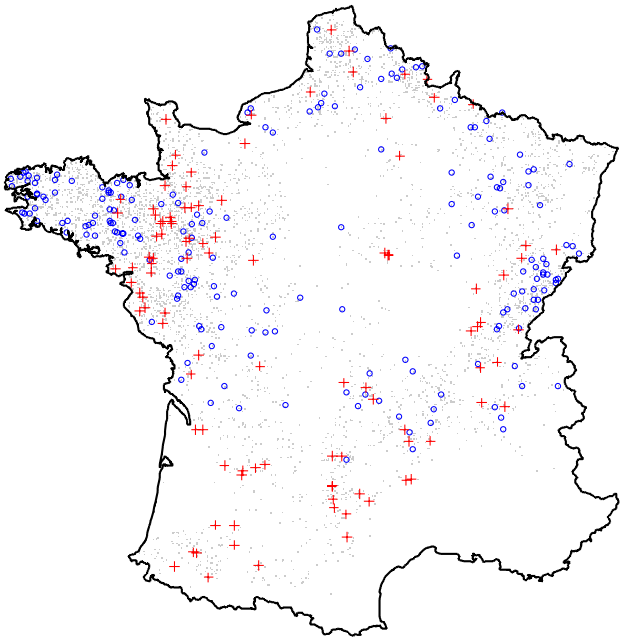


Figure 3: Dairy farms included in the 5 year period with 4 successive annual somatic cell score increases or decreases
+: 4 successive increases
○: 4 successive decreases

Risk factors effects

The 8 factors included in the regression models were: breed (Holstein was the reference to which Montbéliarde, Normande and other breeds were compared), mean parity, number of calvings in spring, summer, autumn and winter, herd size (less than 30 cows, 30 to 40 cows, 40 to 50 cows and more than 50 cows) and altitude.

The R^2 varied between 10 and 15% and all ASCS regression models were highly significant (Table2). Moreover, for each year studied the hypothesis of normality of residuals was not rejected by the Kolmogorov Smirnov test, and the plot of residuals vs. predicted values did not reject the hypothesis of homoscedasticity.

Finally, according to linear regression models, ASCS was significantly raised with increased mean parity and increased numbers of spring calvings. By contrast, ASCS was significantly decreased by an increased number of calvings in summer (except for 1996) and autumn (except for 2000), and by an increased altitude. For each year studied, the breed factor was highly significant, and the influence of each breed compared to Holstein taken as a reference was as follows: Montbéliarde decreased ASCS, and Normande increased the score. The herd size had no significant influence, except in 1997 where an increased herd size increased ASCS.

Table 2: Linear regression model of the annual milk cellular score in the studied population of dairy herds in France, 1996-2000

		Breed	Mean Parity	Summer Calving	Autumn Calving	Winter Calving	Spring Calving	Herd Size	Altitude (m)	Adjusted R^2	Global P value
	coef		0.2717	-0.0025	-0.0037	0.0049	0.0129	0.0186	-0.0030		
1996	SD		0.0250	0.0014	0.0012	0.0017	0.0021	0.0136	0.0004	0.11	***
	P	***	***	NS	**	**	***	NS	***		
	coef		0.3617	-0.0051	-0.0025	0.0026	0.0099	0.0259	-0.0010		
1997	SD		0.0237	0.0013	0.0011	0.0017	0.0020	0.0127	0.0004	0.13	***
	P	***	***	***	*	NS	***	*	*		
	coef		0.3446	-0.0043	-0.0038	0.0026	0.0133	0.0196	-0.0017		
1998	SD		0.0238	0.0014	0.0011	0.0016	0.0019	0.0128	0.0004	0.15	***
	P	***	***	**	***	NS	***	NS	***		
	coef		0.3171	-0.0054	-0.0029	0.0026	0.0095	0.0239	-0.0019		
1999	SD		0.0242	0.0013	0.0011	0.0016	0.0019	0.0126	0.0004	0.11	***
	P	***	***	***	*	NS	***	NS	***		
	coef		0.3099	-0.0052	-0.0005	-0.0001	0.0098	0.0053	-0.0012		
2000	SD		0.0258	0.0014	0.0012	0.0016	0.0019	0.0128	0.0004	0.10	***
	P	***	***	***	NS	NS	***	NS	**		

coef : coefficient estimation
SD: standard deviation
NS: P≥0.05

* : P<0.05
** : P<0.01
*** : P<0.001

Cluster Detection

The result of the spatial deformation of farm distribution is displayed in Figure 4. The map of homogenised farm density was the reference map to visualise further concentration along increasing values of the residuals.

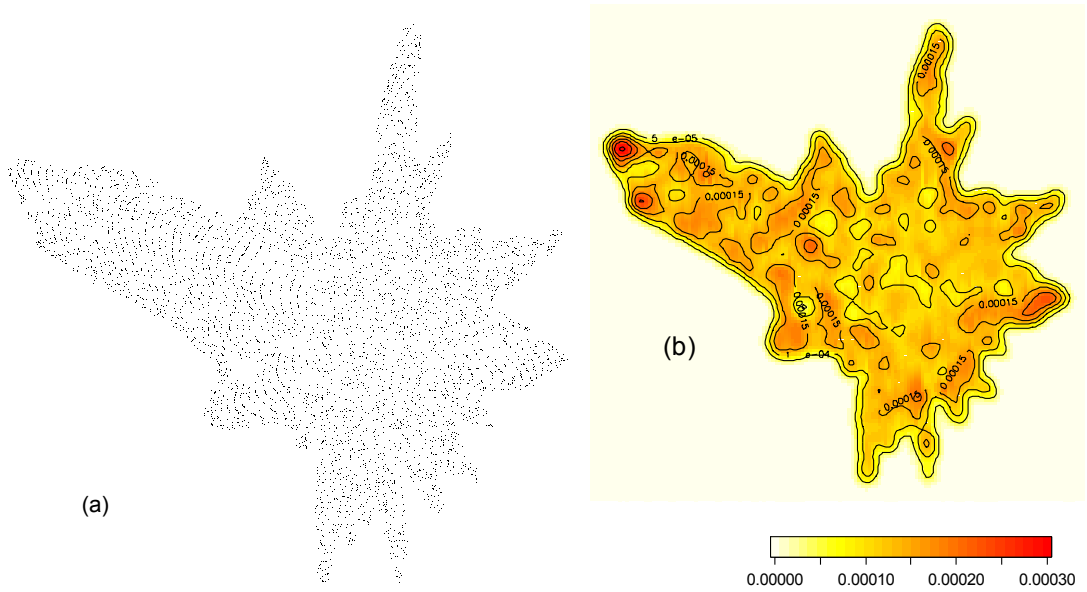


Figure 4: Spatial distribution after deformation (dairy herds in France, n=5210): (a) farms and (b) farm density

The global Hellinger distance statistic and the 5% confidence bounds calculated by permutations in space (1000 permutations) are presented in Table 3. As the global Hellinger distance value was beyond the bounds for the 5 years studied, the hypothesis of absence of spatial cluster formation of the ASCS residuals for the years 1996 to 2000 was rejected.

Table 3: Global Hellinger distance and 5% confidence bounds for the 5 years

	1996	1997	1998	1999	2000
Global Hellinger distance	0.1563	0.1505	0.1503	0.1516	0.1648
Confidence bounds	0.0929 - 0.1026	0.0939 - 0.1036	0.0944 - 0.1039	0.0951 - 0.1049	0.0953 - 0.1046

The process of selection of herds involved in clusters at level 0.64 is presented in Figure 5.

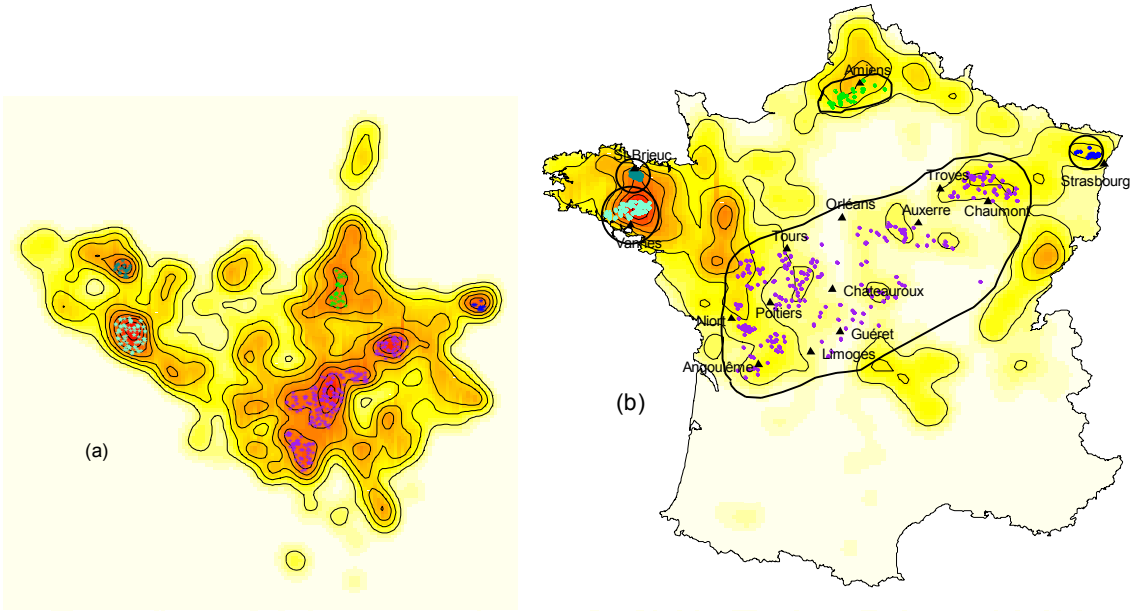


Figure 5: Localisation of clusters of annual somatic cell score residuals in 1996: (a) farm density on deformed map with farms included in clusters and corresponding farms on (b) original map including the main cities located near the significant clusters

Figure 6 shows the summary of the cluster detection results for the 5 years. In 1996 five areas were identified:

- the Morbihan department, with a focused cluster above the city of Vannes
- the Côtes d'Armor department, with another focused cluster below the city of St-Brieuc
- a large area in the centre-north of France, with a diffuse cluster around a line Angoulême-Troyes
- a small area close to the northern tip of France, with a focused cluster below Amiens
- the Bas-Rhin department, with a focused cluster close to the city of Strasbourg

In 1997, there were 7 clusters. Three were the same as in 1996: Vannes, St-Brieuc and Amiens. The 4 others were:

- the Maine-et-Loire department, with a focused cluster above the city of Angers
- the centre-west with a diffuse cluster along a line Niort-Tulle
- the centre-north with a diffuse cluster surrounded by the cities of Orléans, Auxerre, Troyes and Châlons-en-Champagne
- the Ardennes department with a cluster around the city of Charleville-Mézières

In 1998, we detected again the identified clusters of Vannes and Amiens, and the large cluster along the line Angoulême-Troyes as in 1996. Moreover a small cluster was identified around Angers, below the 1997 one.

In 1999, 2 areas were highlighted:

- the first was in the south-west, around the cities of Tulle, Cahors and Toulouse
- the second was in the centre-north, around the cities of Auxerre and Troyes

Six areas were detected for the year 2000. Two of them had already been identified in 1997, i.e. a cluster above Angers and another around Charleville-Mézières. The 4 others areas detected were:

- a large one along a line Niort-Toulouse
- the area close to Orléans-Auxerre-Troyes
- a small area around the city of Bourg-en-Bresse
- the Haute-Saône department, in the north-east, around Vesoul

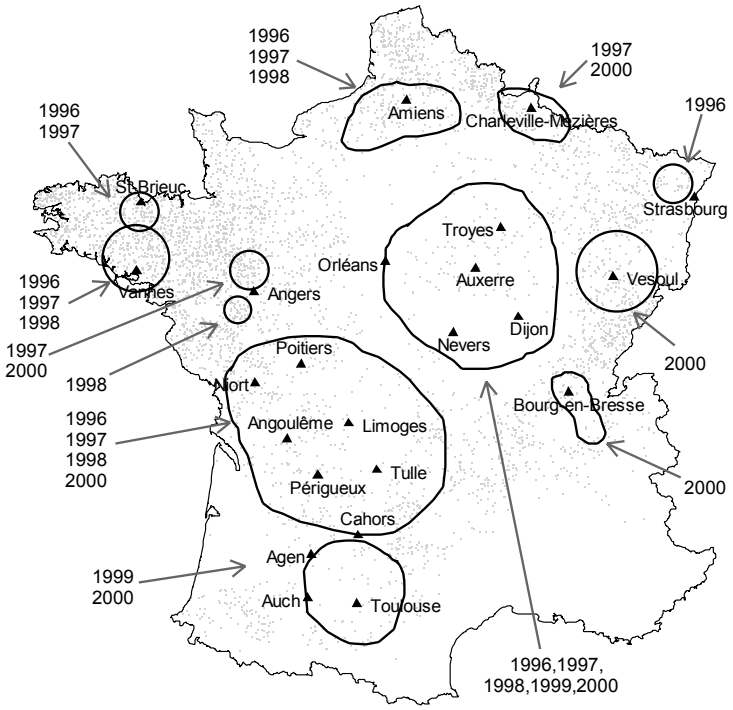


Figure 6: Clusters of somatic cell score residuals detected in France during the study period (1996 to 2000) with indication of detection years

Temporal correlations between ASCS residuals of each year were positive and decreasing: at 1 year lag (96-97, 97-98, 98-99, 99-00) the correlation was 0.76, while it was 0.64 at 2 years lag (96-98, 97-99, 98-00), 0.56 at 3 years lag, and 0.47 at 4 years lag.

The 5% confidence bounds of the Hellinger distance calculated after temporal permutations of ASCS residuals (1000 permutations independently for each farm) were 0.1392 to 0.1515. The global Hellinger distance *Hel* value was beyond these bounds for 1996 and 2000, pointing out that those years were spatially different from the others.

Discussion

The results of the linear regression models were consistent with previously published results indicating significant effects of breed (Barnouin et al., 1999), parity (Laevens et al., 1997) and calving season (Barnouin et al., 2004) on ASCS. Decreasing ASCS with increasing altitude could be explained by the high dairy specialisation of farms located in mountain areas, as the specialisation is associated in France with low ASCS (Barnouin et al., 2004). Nevertheless a great number of known risk factors for ASCS were not available in the present work. Particularly, housing, milking and climatic conditions, which are important factors to consider for a better ASCS control, were not included in the model. Consequently, the clusters detected could be linked to some missing factors. As the aim of the study was to propose a generic method to detect ASCS clusters, the factors included in the models could be considered mainly as examples. Moreover, complementary risk and protection factors could easily be integrated in the models if available later, and a new cluster detection could be performed.

During the 5 years of the study period, 12 clusters were detected. They were not all perceptible from the maps of ASCS intensity, which gave only some tendencies, and which were sensitive to isolated high values. Four clusters were focused on small areas with high farm density in the north-west region. Five other clusters were focused on areas with medium farm densities in the north and the east regions. The 3 last clusters, which were more spatially diffuse, corresponded to areas with low farm density located in the south-west, the centre-west and the centre-north regions. Some clusters were detected over several years, which was consistent with the strong correlation of ASCS residuals between years. The only cluster persistent over the 5 years was located around Auxerre in the centre-north region. The cluster around the line Niort-Tulle was detected for 4 years. Moreover this area progressively spread to the south: it spatially ended at Angoulême in 1996, went to Tulle in 1997 and 1998, and reached Toulouse in 1999 and 2000 (but in 1999 it spread only from Tulle to Toulouse). This result was consistent with the numerous ASCS successive increases highlighted in the south-west region. By contrast, we observed a high number of successive decreases in the north-west (Finistère), and the clusters of St-Brieuc and Vannes were not detected after 1998. Over the entire period studied, we noticed 2 progressive movements: a disappearance of clusters in the north-west, and an extension from the centre-west to the south-west region. Such a statement could explain the spatial singularities of 1996 and 2000 found by temporal permutations: 1996 could be the end of a previous movement of improvement in Brittany, and 2000 the beginning of a deterioration in the south-west. Moreover the singularity of 2000 was emphasised by the appearance of 2 new clusters in the centre-east.

The general trend of ASCS decrease in French dairy herds is consistent with the set up by DHIA of control programs and preventive measures to improve dairy management and mastitis control. The north-west region, the main dairy area in France (50% of the farms) and the more technically advanced, was less touched by clusters of high ASCS values (except Vannes). Controversially, the areas with several clusters detected were non-specialised in dairy production, and more focused on bovine and ovine meat, and on cereal production. But to explain precisely the factors associated with the clusters detected, a local analysis would be necessary, as only local DHIA staff could have accurate and relevant information on local events, decisions or singularities having influenced ASCS in a particular area and a stated period.

Taking into account the risk factors is one of the advantages of our cluster detection method. Moreover this approach based on the Hellinger distance can be used on rough variables and after modelling, and can be applied to quantitative or qualitative data. Another advantage is that it is not necessary to have an a priori precision of either the number or the position of clusters. Moreover, the association with space deformation allows not only taking into account the heterogeneity in the background population but cluster detection even in low farm density areas. On the other hand, this method is not accurate for detection of isolated extreme values, since the number of points becomes insufficient for efficient density estimation.

Nevertheless the method can also be used to detect clusters of low values, which could be useful to identify regions with a very good control of ASCS. Moreover, the comparison of spatial patterns of other milk production and hygiene parameters could be an efficient way to improve milk management at a national level.

References

Ahrens, C., N. Altman, G. Casella, M. Eaton, J. T. G. Hwang, J. Staudenmayer and C. Stefanescu, 2001. Leukemia clusters in update New York: how adding covariates changes the story. Environmetrics. 12, 659-672.

Ali, A. K. A. and G. E. Shook, 1980. An optimum transformation for somatic cell concentration in milk. J Dairy Sci. 63(3), 487-490.

Barkema, H. W., Y. H. Schukken, T. J. G. M. Lam, M. L. Beiboer, G. Benedictus and A. Brand, 1998. Management practices associated with low, medium, and high somatic cell counts in bulk milk. J Dairy Sci. 81(7), 1917-1927.

Barnouin, J., M. Chassagne, S. Bazin and D. Boichard, 2004. Management practices from questionnaire surveys in herds with very low somatic cell score through a national mastitis program in France. J Dairy Sci. 87(11), 3989-3999.

Barnouin, J., N. Geromegnace, M. Chassagne, N. Dorr and P. Sabatier, 1999. Facteurs structurels de variation des niveaux de comptage cellulaire du lait et de fréquence des mammites cliniques dans 560 élevages bovins répartis dans 21 départements français. Anim Prod. 12, 39-48.

Busato, A., P. Trachsel, M. Schallibaum and J. W. Blum, 2000. Udder health and risk factors for subclinical mastitis in organic dairy farms in Switzerland. Prev Vet Med. 44(3-4), 205-220.

Carpenter, T. E., 2001. Methods to investigate spatial and temporal clustering in veterinary epidemiology. Prev Vet Med. 48(4), 303-320.

Draper, N. R. and H. Smith, 1996. Applied regression analysis. John Willey & Sons, New York.

Elliott, P. and J. Wakefield, 2001. Disease clusters: should they be investigated, and, if so, when and how? J. R. Statist. Soc. A. 164(part 1), 3-12.

Ely, L. O., J. W. Smith and G. H. Oleggini, 2003. Regional production differences. J Dairy Sci. 86, E28-E34.

Fevre, E. M., P. G. Coleman, M. Odiit, J. W. Magona, S. C. Welburn and M. E. Woolhouse, 2001. The origins of a new Trypanosoma brucei rhodesiense sleeping sickness outbreak in eastern Uganda. Lancet. 358(9282), 625-628.

Gibbs, A. L. and F. E. Su, 2002. On choosing and bounding probability metrics. Int Stat Rev. 70(3), 419-435.

Harmon, R. J., 1994. Physiology of mastitis and factors affecting somatic cell counts. J Dairy Sci. 77(7), 2103-2112.

Huffman, E. M., R. Mortimer, J. D. Olson, L. Ball and P. W. Farin, 1984. Risk factors for prebreeding pyometra on four Colorado dairy farms. Prev Vet Med. 2(6), 785-790.

Laevens, H., H. Deluyker, Y. H. Schukken, L. De Meulemeester, R. Vandermeersch, E. De Muelenaere and A. De Kruif, 1997. Influence of parity and stage of lactation on the somatic cell count in bacteriologically negative dairy cows. J Dairy Sci. 80(12), 3219-3226.

Manly, B. F. J., 1991. Randomization and Monte Carlo methods in biology. Chapman and Hall, London.

Marshall, R. J., 1991. A review of methods for the statistical analysis of spatial patterns of disease. J. R. Statist. Soc. A. 154(part3), 421-441.

Norstrom, M., D. U. Pfeiffer and J. Jarp, 1999. A space-time cluster investigation of an outbreak of acute respiratory disease in Norwegian cattle herds. Prev Vet Med. 47(1-2), 107-119.

Oleggini, G. H., L. O. Ely and J. W. Smith, 2001. Effect of region and herd size on dairy herd performance parameters. J Dairy Sci. 84(5), 1044-1050.

Perez, A. M., M. P. Ward, P. Torres and V. Ritacco, 2002. Use of spatial statistics and monitoring data to identify clustering of bovine tuberculosis in Argentina. Prev Vet Med. 56(1), 63-74.

Reneau, J. K., 1986. Effective use of dairy herd improvement somatic cell counts in mastitis control. J Dairy Sci. 69(6), 1708-1720.

Senoussi, R., J. Chadoeuf and D. Allard, 2000. Weak homogenization of points processes by space deformations. Adv Appl Prob (SGSA). 32, 948-959.

Silverman, B. W., 1986. The kernel method for univariate data. In: Cox DR, Hinkley DV, Rubin D and Silverman BW (eds), Density estimation for statistics and data analysis. Chapman and Hall, London, pp. 34-94.

Timander, L. M. and S. McLafferty, 1998. Breast cancer in West Islip, NY: a spatial clustering analysis with covariates. Soc Sci Med. 46(12), 1623-1635.

Waller, L. A., 2000. A civil action and statistical assessments of the spatial pattern of disease: do we have a cluster? Regul Toxicol Pharmacol. 32(2), 174-183.

Ward, M. P. and T. E. Carpenter, 2000. Techniques for analysis of disease clustering in space and in time in veterinary epidemiology. Prev Vet Med. 45(3-4), 257-284.

Wartenberg, D., 2001. Investigating disease clusters: why, when and how? J. R. Statist. Soc. A. 164(part 1), 13-22.

Approche paramétrique pour variables continues

IV. Méthode de détection d'agrégat basée sur un modèle de survie spatialisée

IV.1. Introduction

Les modèles de survie sont classiquement utilisés pour quantifier la survie (ou le risque de décès/maladie) dans le temps d'une cohorte d'individus présentant diverses caractéristiques (Droesbeke et al., 1989; Hill et al., 1990). Dans cette thèse, nous utiliserons un modèle de survie pour suivre "l'évolution" de la distribution spatiale des élevages selon le niveau de score. Nous quantifierons la survie (ou le risque de disparition) spatialisée, pour des niveaux de score croissants, en tout point du domaine.

La fonction de risque instantanée r de disparition de points, au niveau de score z, est définie comme dépendante de facteurs de risque locaux W observés et de la présence de facteurs de risque cachés sous forme de "foyers" de maladie modélisés par la fonction ϕ :

$$r\left(z,x,W^x\right)=\underbrace{r_0(z)}_{\text{fonction de risque de base}}\exp\left(\underbrace{\sum_{j=1}^{J}\beta_j W_j^x}_{\text{effet des facteurs de risque connus de la maladie}}-\underbrace{\phi(\gamma,x)}_{\substack{\text{effet foyer :}\\ \text{somme de } K \text{ foyers spatiaux gaussiens}}}\right) \tag{22}$$

$$\phi(\gamma,x)=\sum_{k=1}^{K}\frac{\alpha_k}{2\pi\rho_k^2}\exp\left(-\frac{||x\text{-}c_k||^2}{2\rho_k^2}\right) \tag{23}$$

Le coefficient vectoriel β quantifie l'effet des facteurs de risque. Un β_j positif signifie que la fonction de risque de disparition de points est augmentée, et que donc la probabilité d'occurrence de valeurs de score plus élevées est diminuée.

La matrice $\gamma=(\alpha,\rho,c)$ contient les valeurs des paramètres des K foyers. Le paramètre α quantifie l'intensité du foyer, ρ sa portée, et c ses coordonnées géographiques. Un α positif signifie que la fonction de risque de disparition de points est diminuée, et que le foyer est "attractif", augmentant par là la probabilité d'occurrence de valeurs de score élevées. Par opposition, un α négatif signifie un foyer "répulsif" diminuant la probabilité d'occurrence de valeurs de score élevées.

IV.2. Article n°3

Cet article, intitulé “Spatial modelling of mastitis risk in France”, a pour vocation de décrire cette méthode paramétrique de détection d’agrégats utilisant un modèle de survie spatialisé. Le jeu de données adopté est le même que celui de l’article n°1 décrivant la méthode non paramétrique basée sur la distance d’Hellinger, ceci afin de permettre la comparaison entre les 2 méthodes.

Cet article a été soumis à la revue International Journal of Health Geographics en novembre 2005.

A spatial hazard model for cluster detection on continuous indicators of disease: application to bovine mastitis in France

Emilie Gay [1,2], Jacques Barnouin [2], Rachid Senoussi [1§]

[1]Unité de Biométrie, INRA Domaine St-Paul, Site Agroparc, 84914 Avignon Cedex9, France
[2]Unité d'Epidémiologie Animale, INRA, 63122 St-Genès-Champanelle, France

[§]Corresponding author

Email addresses:
EG: emilie.gay@avignon.inra.fr
JB: barnouin@clermont.inra.fr
RS: senoussi@avignon.inra.fr

Abstract

Background

Methods for spatial cluster detection are usually designed to deal with diseases quantified by dichotomous variables. In the dairy cow, mastitis is evaluated using a continuous indicator of udder inflammation, the somatic cell score (SCS). Consequently, spatialised risk and cluster components of mastitis were analysed through a new method based on a spatial hazard model of the SCS. The model allowed to estimate simultaneously the effects on SCS of the known risk factors and of potential clusters. The dataset contained annual SCS for a random sample of 5000 French dairy herds for the year 1996, and some important SCS risk factors such as mean parity, number of calvings per season and herd size. The hazard function depended on both observable local explanatory variables and on the presence of hidden foci of disease. The estimated foci and their range were then mapped.

Results

With a mean of 3.16, the annual SCS presented a spatial correlation highlighted by the form of the correlogram. Increasing mean parity and herd size significantly increased the risk of high values of annual SCS. The model with the presence of 3 foci was highly significant, and the 3 foci were of attractive type, i.e. increased the occurrence of higher SCS. The 3 localisations were: close to the city of Troyes in the centre-east region (highest strength and range), in the centre-west region (medium strength and range), and in the Morbihan department in the north-west (lower strength and range).

Conclusions

The parametric method based on spatial hazard modelling 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 with a spatially rigid form of clusters.

Background

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 occurrence of peculiar events such as cattle fair can have a great influence on disease spread and control. The available methods to explore spatial patterns range from geostatistics to point process approaches. Among these, the issue of cluster detection [1, 2] is of major interest, as it can assist control and prevention in pointing out possible causes of hot spots of the disease.

The main techniques used in cluster detection rely on scan statistic [3-6]. This method compare the observed number of cases inside a moving window to the expected number of cases under a Poisson or a Bernouilli distribution hypothesis [7]. Scan statistic however deals only with dichotomic variables (e.g. cases/controls), and can not take account of specific information at the individual level. Spatial modelling is another way to explore spatial patterns. Modelling allows the quantification of the effects of known disease risk factors, and attempts to focus on unexplained clustering [8]. Among the several approaches, some handle the concept of infectious potential, through Susceptible-Infected-Recovered (SIR) models [9], which can be linked to point-pattern methodology [10]. Others approaches use the classical framework of linear mixed models with risk factors as fixed effects, while spatial variations are included as a random effect with a geostatistical [11] or Bayesian point of view [12]. Some last methods specify the intensity of case events depending on location of clusters centres [13]. But as for the scan tests those methods deal only with dichotomic variables. Cluster detection for diseases measured by continuous variables, such as biological indicators, remains an unexplored field.

Bovine mastitis is an inflammation of the mammary gland, usually inducted by bacterial infection of udder tissues. This key problem of the dairy cattle, is reliably evaluated by the determination of milk somatic cell count (SCC), a biological indicator which is generally expressed through its logarithm transformation, named somatic cell score or SCS ($\mathrm{SCS} = \log_2(\mathrm{SCC}/100{,}000) + 3$). This continuous variable is internationally recognised as a valuable indicator for mastitis control at the herd level [14]. Risk factors associated with SCS had been widely investigated [15-17]. But in these studies the SCS spatial aspects were not taken into account, while SCS typically presents strong spatial variations that had been highlighted by several previous descriptive studies [18, 19]. Research on spatial SCS features, as for many other diseases indicators, is still uncommon in animal epidemiology. Differences in natural resources, farm structure and market rules can cause regions of a same country to implement specific dairy management systems, and call for the introduction of a spatial component in SCS data analysis.

A method of cluster detection has already been applied to SCS, but the results were only qualitative [20].

The purpose of this research was to quantitatively analyse the spatialised risk of SCS, using a spatial hazard model to simultaneously estimate precisely the effects on mastitis risk of known risk factors and of the location and form of potential spatial clusters. We used a dataset for which cluster detection had already been performed via a non parametric test procedure [20]. This dataset contained annual herd SCS (ASCS) for a random sample of 5000 French Holstein dairy herds for the year 1996, and some risk factors (mean parity, number of calvings per season and herd size).

Spatial patterns of ASCS were explored through a survival analysis, i.e. following the evolution of the spatial distribution of farms with increasing ASCS. The ASCS were considered as “lifetimes” whose hazard function depended on observable local explanatory variables and on the presence of *K* hidden foci, described by their strength α, range ρ and geographic coordinates c. If an explanatory variable in the model was positively associated with risk of disappearance of farms for increasing ASCS, this variable decreased the occurrence of higher ASCS levels. All the same, a focus with a positive strength α was “attractive” and increased the occurrence of high ASCS levels. To illustrate the results we, mapped the estimated foci and their range, and compared them to the results obtained by the non parametric cluster detection method based on the Hellinger distance developed by Gay et al (2005) [20]. The methods used are detailed in the method section.

Results

Descriptive analysis

The mean ASCS was 3.16 (±0.57), mean herd size was 40 cows (±16) and mean parity was 2.6 (±0.3). Calvings occurred mainly in autumn (41%), 24% occurring in winter, 22% in summer, and 13% in spring.

The geographic distribution of the farms is illustrated in Figures 1 and 2, where Figure 1 represents the location of the study farms, and Figure 2 the farm density obtained by a kernel smoothing technique. The farm density was non-homogeneous in space, and showed two areas with higher densities: 1) the north-west region, which is the main dairy production area in France (61% of the total number of farms), and 2) the north tip region, with 6% of the total number of farms. By contrast, the south-east region, i.e. the Mediterranean area, corresponded to very low farm density.

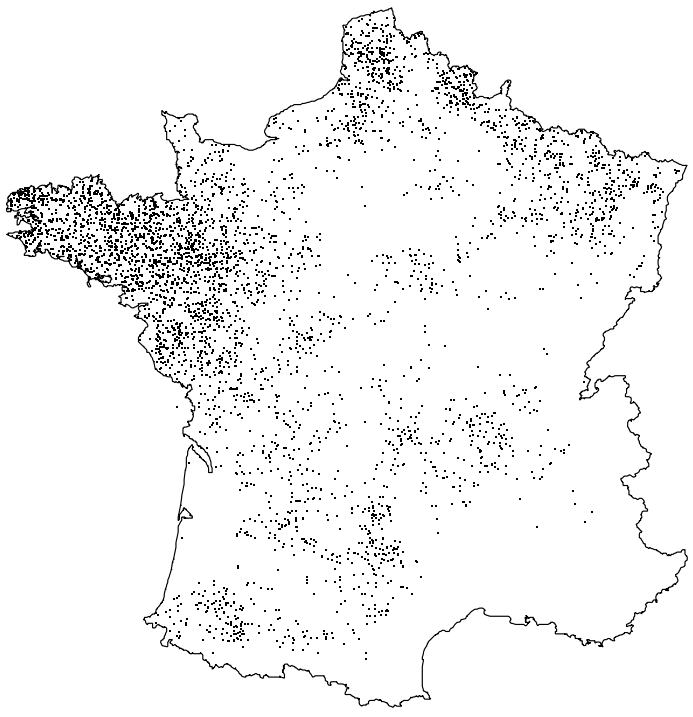


Figure 1 - Farm location of the study sample of dairy herds in France n=5000, year 1996

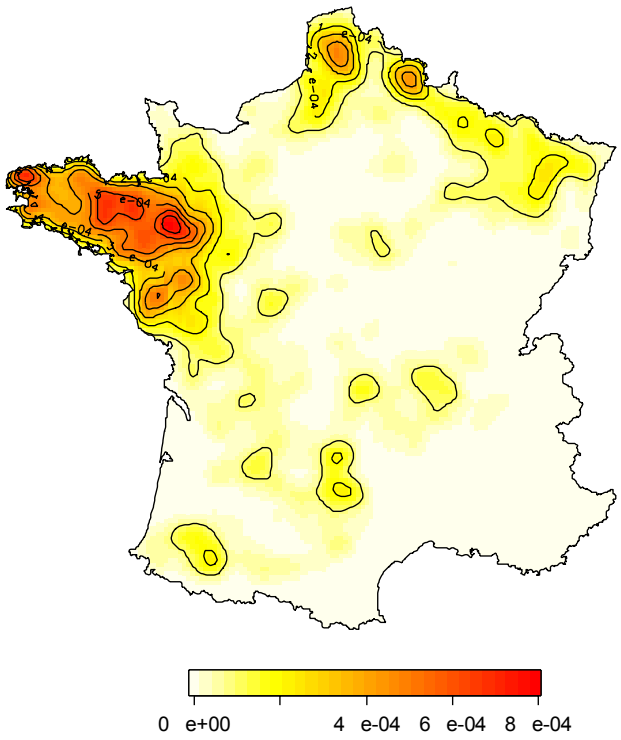


Figure 2 - Farm density of the study sample of dairy herds in France
unit: farms/km², n=5000, year 1996

The correlogram of ASCS (Figure 3) showed a positive and non negligible spatial correlation under a distance of 100 km with an approximated exponential form. Over this distance, it could be considered as constant around 0. The behaviour of the correlogram at distance 0 pointed out a strong nugget effect (autocorrelation of 1 at null distance if absence of nugget effect), and then the presence of a relative high white noise (non spatial correlation) of about 70% of the total variability.

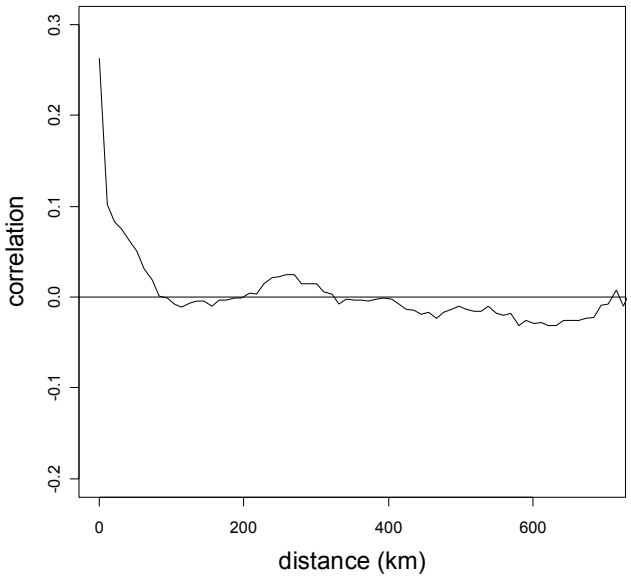


Figure 3 - Correlogram of the annual somatic cell score of the study sample

Spatial modelling

Eight risk factors were included as explanatory variables in the model, with a regression parameter β, as follows:

- 1 variable for mean parity;
- 4 variables for number of calvings occurring respectively in spring, summer, autumn and winter;
- 3 dummy variables for the 4 groups of the nominal herd size variable: group 1, corresponding to herds with less than 30 cows, which was taken as the reference group, group 2 concerning herds with 30 to 40 cows, group 3 corresponding to herds with 40 to 50 cows, and group 4 corresponding to herds with more than 50 cows.

We performed several models with different number of foci K, from K= 0, i.e. a model with only the risk factors ($\mathcal{M}_{\beta,0}$), to K= 4 ($\mathcal{M}_{\beta,4}$), and we tested their significance (Table 1).

Table 1 - Tests of the different hazard models of the annual milk somatic cell score

Model	Number of variables	Deviance	LRS test	LRS value	LRS (df)	P value
$\mathcal{M}_{0,0}$: no covariable	0	75190.37				
$\mathcal{M}_{\beta,0}$: risk factors alone	8	74964.25	$\mathcal{M}_{0,0}$ vs. $\mathcal{M}_{\beta,0}$	226.12	(8)	***
$\mathcal{M}_{\beta,1}$: risk factors + 1 focus	12	74715.71	$\mathcal{M}_{\beta,0}$ vs. $\mathcal{M}_{\beta,1}$	248.54	(4)	***
$\mathcal{M}_{\beta,2}$: risk factors + 2 foci	16	74564.81	$\mathcal{M}_{\beta,0}$ vs. $\mathcal{M}_{\beta,2}$	399.44	(8)	***
			$\mathcal{M}_{\beta,1}$ vs. $\mathcal{M}_{\beta,2}$	150.90	(4)	***
$\mathcal{M}_{\beta,3}$: risk factors + 3 foci	20	74529.72	$\mathcal{M}_{\beta,0}$ vs. $\mathcal{M}_{\beta,3}$	434.53	(12)	***
			$\mathcal{M}_{\beta,1}$ vs. $\mathcal{M}_{\beta,3}$	185.99	(8)	***
			$\mathcal{M}_{\beta,2}$ vs. $\mathcal{M}_{\beta,3}$	35.09	(4)	***
$\mathcal{M}_{\beta,4}$: risk factors + 4 foci	24	74526.93	$\mathcal{M}_{\beta,0}$ vs. $\mathcal{M}_{\beta,4}$	437.32	(16)	***
			$\mathcal{M}_{\beta,1}$ vs. $\mathcal{M}_{\beta,4}$	188.78	(12)	***
			$\mathcal{M}_{\beta,2}$ vs. $\mathcal{M}_{\beta,4}$	37.88	(8)	***
			$\mathcal{M}_{\beta,3}$ vs. $\mathcal{M}_{\beta,4}$	2.79	(4)	NS

LRS : Likelihood Ratio Statistic
df : degrees of freedom
NS : P≥0.05

* : P<0.05
** : P<0.01
*** : P<0.001

The model with *K*=3 ($\mathcal{M}_{\beta,3}$) was selected as the presence of 3 foci was highly significant, while the 4th focus was not. The detailed results of estimations for this model are presented in Table 2. Hazard of disappearance of farms was significantly decreased by increased mean parity (1 parity decreased the risk by $e^{0.4344}$ =1.54), so this factor increased the occurrence of high ASCS. The number of calvings for each season had little influence: only autumn and spring were significant, and the coefficients were low. Nevertheless, spring calvings increased ASCS, while autumn calvings decreased it. Concerning herd size, the 3 groups increased the risk of high ASCS compared to the reference group. The higher the herd size was, the stronger the influence was, except for the last group (more than 50 cows) for which the occurrence of high values of ASCS was still increased but less than for the previous ones.

Table 2 - Spatial hazard model of the annual milk somatic cell score (3 foci)

		Coef	SD	exp(coef)	LRS	(df)	P value
Mean parity		-0.4344	0.0459	0.65	139.12	(1)	***
Summer calving		0.0057	0.0028	1.01	1.40	(1)	NS
Autumn calving		0.0077	0.0022	1.01	40.78	(1)	***
Winter calving		-0.0005	0.0032	1.00	0.98	(1)	NS
Spring calving		-0.0269	0.0043	0.97	16.37	(1)	***
Number of cows	30-40	-0.0344	0.0396	0.97	6.32	(3)	*
	40-50	-0.2142	0.0537	0.81			
	>50	-0.1352	0.0817	0.87			
Focus 1	α	0.0286	0.0079				
	ρ	0.0676	0.0089				
	x_c	-0.2623	0.0063				
	y_c	0.0314	0.0101				
Focus 2	α	0.2968	0.0617				
	ρ	0.2041	0.0118				
	x_c	0.2892	0.0467				
	y_c	0.0681	0.0343				
Focus 3	α	0.1024	0.0520				
	ρ	0.1235	0.0231				
	x_c	0.0326	0.0421				
	y_c	0.1228	0.0340				

coef: coefficient estimation — NS: P≥0.05 — α : focus strength
SD: standard deviation — * : P<0.05 — ρ : focus range
LRS: likelihood ratio statistic — ** : P<0.01 — x_c: focus x coordinate
df : degrees of freedom — *** : P<0.001 — y_c: focus y coordinate

The 3 foci were attractive. The first one was detected in the Morbihan department (Figure 4), close to the city of Vannes, in the north-west region. This focus exercised the lowest attraction over the shortest distance. The second focus was located at the city of Troyes, in the centre-east region, and was the main one with the highest strength and range. The third focus, located in the centre-west region, was surrounded by the cities of Chateauroux, Limoges, Poitiers and Tours. It spread all over the departments of Indre, Creuse, Haute-Vienne, Vienne and Indre-et-Loire. Its strength and range were intermediate between those of the two other foci.

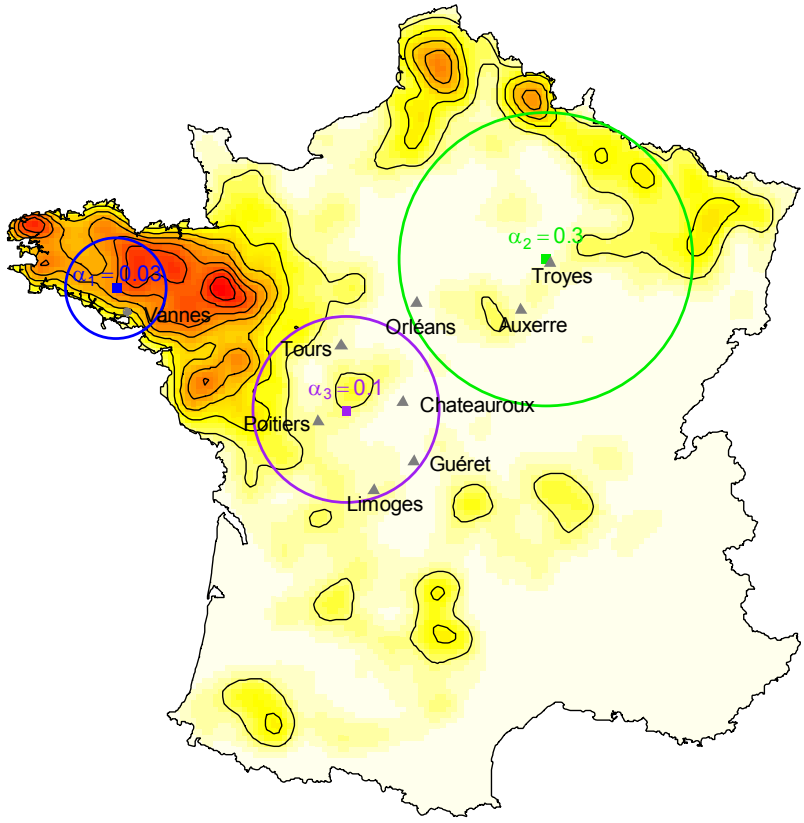


Figure 4 - Farm density and foci of high annual somatic cell score detected by spatial hazard modelling

■ *: focus centre*

α : focus strength

▲*: mains cities around the foci*

◯ *: focus range (ρ)*

The underlying hazard curve $R_0(z)$ of the hazard model ($\mathcal{M}_{\beta,3}$ containing risk factors and spatial effect of 3 foci) was estimated, and then compared to the underlying hazard curves of a model adjusted only for risk factors ($\mathcal{M}_{\beta,0}$), and a raw hazard function ($\mathcal{M}_{0,0}$, estimated with all the parameters equal to zero). Figure 5 showed that the risk of disappearance of points at any level was higher when adjusted on all covariables (under model $\mathcal{M}_{\beta,3}$), i.e. withdrawing the effect of the risk factors and of the foci. The risk adjusted only for risk factors (under model $\mathcal{M}_{\beta,0}$) was higher than the non-adjusted risk (under model $\mathcal{M}_{0,0}$), but lower than the one adjusted on both risk factors and foci. The curves confirmed that the effect of both risk factors and foci decreased the hazard function, and thus increased the occurrence of high ASCS values.

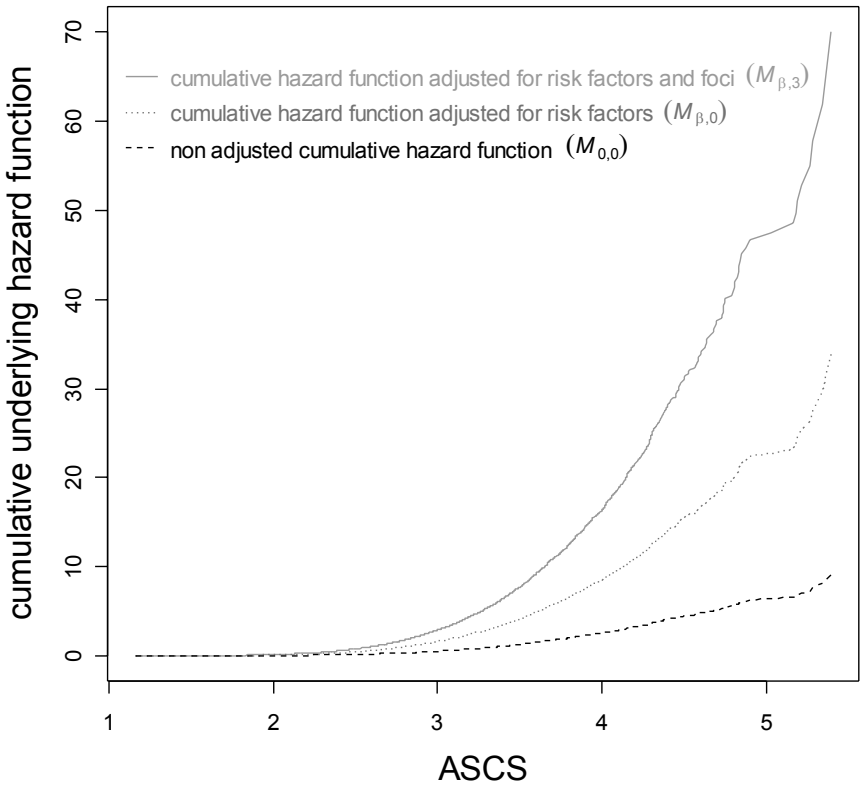


Figure 5 - Cumulative underlying hazard curves adjusted and non-adjusted for covariables of annual milk somatic cell score

The comparison of estimated foci with clusters detected by the Hellinger distance method (Figure 6) showed that the significant presence of the 3 foci corresponded spatially to the same regions. Due to our choice of Gaussian foci, the areas identified by the spatial hazard model were compelled to be disks and were larger. However, the aggregation was precisely quantified by the model, whereas it was only qualitative with the Hellinger distance detection.

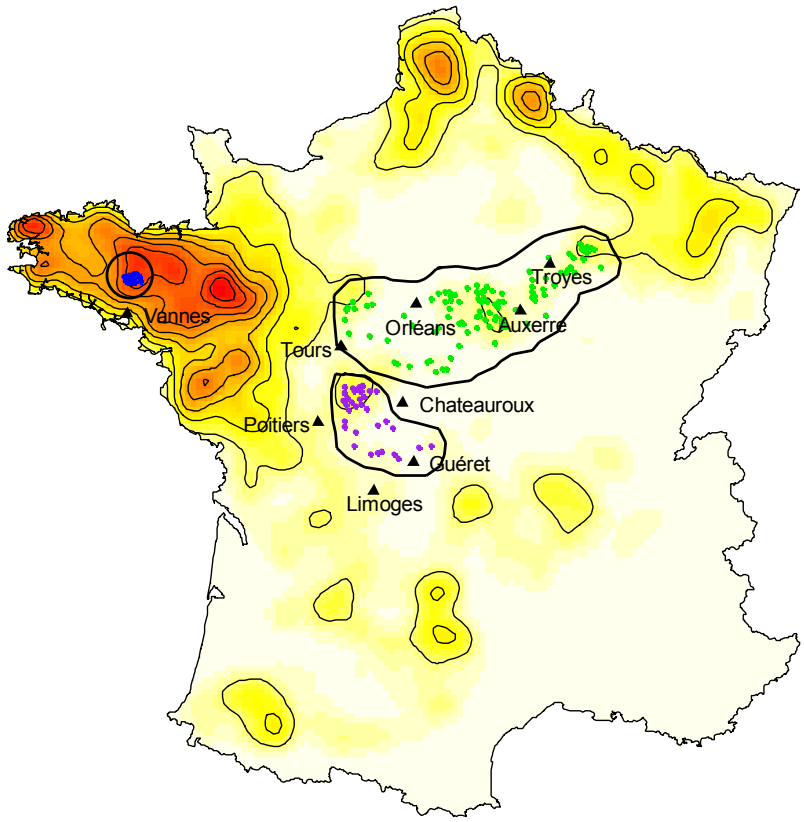


Figure 6 - Farm density and clusters of high annual somatic cell score detected by Hellinger distance method

• : farms involved in the cluster

▲: mains cities around the clusters

: demarcation of the clusters

Discussion

The results of the spatial hazard model concerning the introduced risk factors for ASCS were consistent with previously published results, whatever the method used, indicating a significant effect of parity, calving seasons, and herd size on ASCS. Increased mean parity increases the risk of high ASCS levels [21]; it can be due to the rise of persistence and intensity of mammary infections with parity. A high number of calvings in autumn is a protection factor against elevated ASCS values, while spring calving is a risk factor [17]. Spring calvings can be deleterious because cow body condition and housing hygiene are generally poorer at the end of winter. Occurrence of high ASCS is raised by growing herd size [22], which can be explained by an increased risk of transmission of udder pathogens from cow to cow, or by less time spent for udder surveillance in large herds. Parameter estimation suggested however a sill effect for high herd size, a result which could be linked to the higher level of specialisation of such farms [17].

A great number of other known risk factors for ASCS were not available in the present work. Particularly, for hygienic and milking conditions, which are very important to consider [23, 24], information were lacking. Nevertheless, the method allows complementary variation factors to be easily integrated in the model if available later.

Concerning the second and original part of the model, the focus effect, the 3 foci were highly significant, and approximately the same than the clusters detected by the method based on the Hellinger distance [20]. The similarities between the results of these 2 different approaches allowed having confidence in the results and the robustness of both methods. Actually the results showed the complementarity nature of the two methods: qualitative but spatially flexible cluster detection trough the non parametric method, and quantitative but spatially rigid form and location of clusters trough the parametric model developed here.

The two main foci, located in areas with low farm density (centre-east and centre-west regions), corresponded to French regions which were non-specialised in dairy production. As it is known that specialisation in dairy production is associated in France with lower ASCS [17], those regions, mainly focused on bovine and ovine meat and on cereal production, have a higher mastitis risk. Introducing the farm density as a covariable in the model could be a way to approximate the specialisation factor.

The focus close to the city of Vannes is more difficult to explain. High farm density could favour microbial spread through passive human transport, but why only in this particular zone? A local analysis would be necessary to explain precisely the factors associated with the foci identified, as only local staff could have accurate and relevant information on local events or singularities having influenced ASCS. The veterinarian of the dairy herd improvement association of the Morbihan suggested that some farms in the department used widely in the nineties a deleterious bull through artificial insemination, which produced cows very susceptible to mastitis. Moreover, the bovine spongiform encephalopathy crisis led to fewer animal reforms in the Morbihan, which involved that cows experiencing high SCS level were kept longer than usual in the herds.

The spatial hazard modelling we developed has many advantages. First, the method applies to continuous variables, such a possibility was lacking in spatial modelling of diseases. Secondly, the proposed model is adjusted for risk factors as in Klassen et al (2005) [25], 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 quantitative, tests of comparative hypothesis on the two components, risk factors and foci presence, can be readily performed.

On the other hand, the present drawback of this method is the choice of the number of foci. This difficulty can be bypassed by introducing in the model not a fixed but a random number of foci, using a BIC or AIC like criteria. A parametric model implies to define the form of the foci. Here we chose a standard Gaussian form characterised by intensity and range parameters and a circular form of the foci, and we observed that the ratio α/ρ^2 were very close (around 6 to 7) for the 3 foci. Another parametric form of the foci component could better fit the problem [26]; the choice of another form must be however supported by epidemiological arguments.

Conclusions

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

The approach of the spatialised risk with a focus component is generic; it is also intended to apply to other diseases, as well as to classical survival models considering the lifetime or occurrence times of infection.

Methods

Data

To evaluate the efficiency of the proposed spatial hazard model, a dataset for which cluster detection had already been performed via a non parametric test procedure [20] was used. The dataset, which concerns the year 1996, contains annual herd SCS (ASCS) for a random sample of 5000 French Holstein dairy herds. For each selected herd, ASCS was computed as the arithmetic mean of all monthly cow SCS values during 1996. The other variables of the dataset, mean parity, number of calvings per season and herd size, had been recognised in various studies as herd factors influencing SCS.

Statistical analysis

Statistical procedures were conducted using R 2.0.1 (R Development Core Team, 2004, R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0). Specific programs, which were implemented to analyse our dataset, are available at request.

After a descriptive analysis of the study variables, the spatial patterns of the data were explored via mapping representations. The interpolation technique of kernel smoothing was used to represent local farm density [27]. Moreover, the presence of spatial correlation was assessed and quantified using a correlogram, which is a graph of autocorrelation of SCS by distance between farms [28].

As a second step, we explored spatial patterns of ASCS from the point of view of survival analysis [29], i.e. following the evolution of the spatial distribution of farms with increasing ASCS. We considered ASCS as "lifetimes" whose hazard function depended on observable local explanatory variables and on the presence of hidden foci. The hazard function *r* at an ASCS level *z* was the probability that a farm ASCS belonged to a small interval [*z,z*+Δ*z*], given that the ASCS was greater or equal to the value *z*: $r = \frac{f}{1-F}$ where *f* was the probability density function, and *F* the cumulative distribution function. To compare to classical survival analysis, it was the hazard of disappearance of points at the level *z*. In our context, the ASCS variable was not censured, and the explanatory variables (risk factors and a foci effect) were not considered as ASCS dependent. The spatial hazard function *r* was defined as follows:

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

where *z* was any ASCS level, *x* the spatial coordinates of the farms, $W^x = (W_1^x, ..., W_J^x)$ the vector of risk factors specific of farms at location *x*, β_j the coefficient for the j[th] risk factor, $r_0(z)$ the underlying hazard function, and $\phi(\gamma, x)$ a potential spatial effect.

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 disappearance of points at any level *z* by a factor $\exp(\beta_j)>1$, i.e. decrease the occurrence of higher ASCS levels.

The log-linearity of continuous variables was checked using the Schoenfeld residuals as recommended in Hill et al (1990) with function "cox.zph" of package "survival".

The spatial effect $\phi(\gamma, x)$, which aimed to take into account the aggregation of the farms, was modelled as the sum of spatial Gaussian foci as follows:

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

where *K* was the number of foci which had to be fixed, and $\gamma = (\alpha, \rho, c)$ the (*K*×4) matrix of parameters. The α_k parameter was the strength of the focus *k*, ρ_k its positive range, and c_k its two geographic coordinates. If a point went close to a focus, the distance $||x\text{-}c_k||$ was low, the exponential increased to the maximum value 1, so the focus effect tended to $\alpha_k / 2\pi\rho_k^2$. Conversely, if a point went far from the focus, the exponential tended to 0 and the focus effect fainted with Gaussian rate until there was no effect of the focus. If α was positive, the global hazard of disappearance of points decreased by a factor $\exp\left(-\alpha_k / 2\pi\rho_k^2\right)$, the focus 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.

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

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

The β coefficients and the vector γ were estimated by maximum likelihood. Calculations were performed using the function “mle” of “stats4” package. We used the Likelihood Ratio Statistic (LRS) to test whether the effects of the covariables were significant [29]. The LRS is the difference in the deviance $D = -2L^*$ for two nested models $\mathcal{M}_i$ and $\mathcal{M}_j$ (with more parameters in $\mathcal{M}_j$ than in $\mathcal{M}_i$):

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

where ν is the difference in the number of parameters between $\mathcal{M}_i$ and $\mathcal{M}_j$.

After the estimation of the parameters β and γ, we estimated the cumulative underlying hazard function $\hat{R}_0(z) = \int_z r_0(u)du$ by the formula:

$$\hat{R}_0(z_i) = \sum_{s=1}^{i} \frac{1}{\sum_{l \geq s} \exp\left(\sum_{j=1}^{J} \hat{\beta}_j W_j^{x_l} - \phi(\hat{\gamma}, x_l) \right)} \quad (5)$$

This estimated underlying hazard function was adjusted for the covariables (risk factors and spatial effect of foci)

Then we compared this adjusted hazard function to the one adjusted only for risk factors, and to the raw hazard function (estimated with all the parameters equal to zero).

As a last step, we mapped the estimated foci and their range and compared them to the results obtained by the non parametric cluster detection method developed by Gay et al (2005). This last method uses a statistic based on the Hellinger distance to compare geographical distribution of the farms $\overline{p}_{z_i}$, at different ASCS levels i, with the global normalised farm density $\overline{p}_{z_0}$:

$$Hel(z_j) = \sqrt{\frac{1}{2} \int_D \left(\sqrt{\overline{p}_{z_0}(x)} - \sqrt{\overline{p}_{z_j}(x)} \right)^2 dx} \quad (6)$$

The significance is tested by permutations, and the progressive concentration of herds with increasing ASCS identifies clusters.

Authors' contributions

EG participated in the epidemiological and statistical analysis, programmed the R code, and drafted the manuscript. JB participated in the epidemiological analysis and helped to draft the manuscript. RS participated in the statistical analysis, assisted with programming the model, and helped to draft the manuscript. All authors participated in the preparation and approved the final version of the manuscript.

References

1. Ward MP, Carpenter TE: **Techniques for analysis of disease clustering in space and in time in veterinary epidemiology**. *Prev Vet Med* 2000, **45**(3-4):257-284.
2. Carpenter TE: **Methods to investigate spatial and temporal clustering in veterinary epidemiology**. *Prev Vet Med* 2001, **48**(4):303-320.
3. Norstrom M, Pfeiffer DU, Jarp J: **A space-time cluster investigation of an outbreak of acute respiratory disease in Norwegian cattle herds**. *Prev Vet Med* 1999, **47**(1-2):107-119.
4. Perez AM, Ward MP, Torres P, Ritacco V: **Use of spatial statistics and monitoring data to identify clustering of bovine tuberculosis in Argentina**. *Prev Vet Med* 2002, **56**(1):63-74.
5. Odoi A, Martin SW, 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.
6. Ozdenerol E, Williams BL, Kang SY, Magsumbol MS: **Comparison of spatial scan statistic and spatial filtering in estimating low birth weight clusters**. *Int J Health Geogr* 2005, **4**:19.
7. Kulldorff M: **A spatial scan statistic**. *Commun Stat-Theor M* 1997, **26**(6):1481-1496.
8. Marshall RJ: **A review of methods for the statistical analysis of spatial patterns of disease**. *J R Statist Soc A* 1991, **154**(part3):421-441.
9. Doran RJ, Laffan SW: **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**(1-2):133-152.
10. Gerbier G, Bacro JN, 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**(1):33-49.
11. Stevenson MA, Benard H, Bolger P, Morris RS: **Spatial epidemiology of the Asian honey bee mite (Varroadestructor) in the North Island of New Zealand**. *Prev Vet Med* 2005, **(in press)**.
12. Yang GJ, Vounatsou P, Zhou XN, 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**(2):155-162.
13. Lawson AB: **Cluster modelling of disease incidence via RJMCMC methods: a comparative evaluation**. *Stat Med* 2000, **19**:2361-2375.
14. Harmon RJ: **Physiology of mastitis and factors affecting somatic cell counts**. *J Dairy Sci* 1994, **77**(7):2103-2112.

15. Romain HT, Adesiyun AA, Webb LA, Lauckner FB: **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**(4):257-271.
16. Busato A, Trachsel P, Schallibaum M, Blum JW: **Udder health and risk factors for subclinical mastitis in organic dairy farms in Switzerland**. *Prev Vet Med* 2000, **44**(3-4):205-220.
17. 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**(11):3989-3999.
18. Oleggini GH, Ely LO, Smith JW: **Effect of region and herd size on dairy herd performance parameters**. *J Dairy Sci* 2001, **84**(5):1044-1050.
19. Ely LO, Smith JW, Oleggini GH: **Regional production differences**. *J Dairy Sci* 2003, **86**:E28-E34.
20. Gay E, Senoussi R, Barnouin J: **A spatial clustering analysis for continuous variables with application to milk somatic cell score in France**. In: *Biometrical Unit Research reports.* Avignon; 2005: [http://www.avignon.inra.fr/stcavignon/centre/unites/biometrie/RR2005_2007.pdf].
21. Laevens H, Deluyker H, Schukken YH, 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**(12):3219-3226.
22. Skrzypek R, Wojtowski J, Fahr RD: **Factors affecting somatic cell count in cow bulk tank milk--a case study from Poland**. *J Vet Med A Physiol Pathol Clin Med* 2004, **51**(3):127-131.
23. 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**(5):1909-1916.
24. Barkema HW, Schukken YH, Lam TJGM, Beiboer ML, Benedictus G, Brand A: **Management practices associated with low, medium, and high somatic cell counts in bulk milk**. *J Dairy Sci* 1998, **81**(7):1917-1927.
25. Klassen AC, 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.
26. Tango T, Takahashi K: **A flexibly shaped spatial scan statistic for detecting clusters**. *Int J Health Geogr* 2005, **4**:11.
27. Silverman BW: **The kernel method for univariate data**. In: *Density estimation for statistics and data analysis.* Edited by Cox D, Hinkley D, Rubin D, Silverman B. London: Chapman and Hall; 1986: 34-94.
28. Cressie NAC: **Geostatistics**. In: *Statistics for spatial data.* Edited by Barnett V, Bradley R, Fisher N, Hunter J, Kadane J, Kendall D, Smith A, Stigler S, Teugels J, Watson G. New York: John Wiley and Sons; 1991: 58-67.
29. Hill C, Com-Nougué C, Kramar A, Moreau T, O'Quigley J, Senoussi R, Chastang C: **Analyse statistique des données de survie**. Paris: INSERM Médecine-Sciences Flammarion; 1990.

Discussion, conclusion et perspectives

V. Discussion sur les méthodes et conclusion

V.1. Méthode non paramétrique basée sur la distance d'Hellinger

La méthode non paramétrique de détection d'agrégats basée sur la distance d'Hellinger présente de nombreux avantages.

Tout d'abord elle s'applique aux variables continues, caractéristique qui faisait défaut aux outils disponibles jusqu'ici. Elle peut aussi s'appliquer aux variables discrètes, et pour les maladies approchées par variable dichotomique (cas/témoins) il suffit de l'appliquer en ne considérant plus que 2 niveaux pour la variable z, 0 et 1. Cependant, même pour ces variables dichotomiques, l'intérêt de cette méthode est qu'il est possible de l'appliquer sur des résidus de modélisation, qui sont des variables continues. Ainsi, les facteurs de risque connus et disponibles de la maladie peuvent être pris en compte, et l'analyse peut porter sur les agrégats non expliqués.

L'hétérogénéité de la population sous jacente est elle aussi prise en compte par cette méthode non paramétrique. La distance d'Hellinger mesure en effet la distance entre densités normalisées de population, l'hétérogénéité est présente dans les 2 niveaux de z, et donc s'annule. De plus, l'homogénéisation de la distribution de population permet parallèlement d'éliminer le problème de l'hétérogénéité dans l'estimation des densités. Elle permet aussi de visualiser les agrégats. En effet, la différence de densités de points aux différents niveaux de score serait difficile à percevoir visuellement si la densité de départ n'était pas quasi-uniforme, et ce, particulièrement dans les zones à faible densité.

Enfin, un des atouts majeurs de la méthode basée sur la distance d'Hellinger est qu'elle ne nécessite aucun paramètre a priori : le nombre, la forme et la distribution des agrégats n'ont pas besoin d'être fixés à l'avance.

Par contre, les résultats obtenus par cette méthode ne sont que qualitatifs. La définition de l'agrégation qui est utilisée s'appuie sur la détection de tendances, donc sur une hétérogénéité d'ordre 1 (moyenne), sans considérer l'hétérogénéité d'ordre 2 (structure de covariance) qui pourrait pourtant être intéressante. De même, telle qu'utilisée dans cette thèse, la méthode ne prend en compte que la forme de la distribution des points, et pas la quantité de points présents à chaque niveau. Afin d'améliorer ce point, une statistique utilisant la distance d'Hellinger, mais considérant l'ensemble des points dépassant un niveau donné, est actuellement en cours d'élaboration.

Enfin, la méthode basée sur la distance d'Hellinger présente le même inconvénient que le test de Tango. En effet, la significativité de l'agrégation peut être testée globalement, et pour chaque niveau de la variable, mais il n'existe pas de test de significativité local pour chaque agrégat détecté visuellement.

Cet outil de détection non paramétrique des agrégats, basé sur la distance d'Hellinger, est flexible et générique. Il peut s'adapter à tout type de variable de mesure, et donc à un panel très large de maladies. De plus, l'extension au temporel, sous réserve de disposer d'un nombre suffisant de répétitions dans le temps, est tout à fait envisageable et ne nécessite pas de traitement à part.

V.2. Méthode paramétrique basée sur le modèle de survie spatialisée

La méthode paramétrique de détection d'agrégats utilisant un modèle de survie spatialisée, permet d'obtenir des résultats quantitatifs et de tester la significativité de chaque foyer. La modélisation intègre par ailleurs les facteurs de risque de la maladie, et tient compte de l'hétérogénéité de la population sous-jacente.

Cependant, cette approche comporte 2 inconvénients importants. Tout d'abord l'utilisation de cette méthode implique de définir a priori une forme d'agrégation, inconvénient commun à toute méthode paramétrique. Nous avons choisi, dans le cadre de cette thèse, une forme gaussienne, ce qui conduit à des paramètres facilement interprétables et à des foyers circulaires (facilement extensible à des foyers ellipsoïdaux). Nous n'avions en effet pas d'informations spécifiques sur le type de propagation spatiale de la maladie, laquelle est a priori non contagieuse inter élevages. Dans le cas de maladies contagieuses, le type de propagation, s'il est connu ou s'il peut être suspecté, permet d'orienter le choix de la forme de l'agrégation (agrégation le long des axes de circulation majeurs…).

Le deuxième inconvénient de cette méthode basée sur la survie spatialisée concerne le nombre de foyers. Plusieurs modèles sont en effet testés, et lorsque l'ajout d'un foyer supplémentaire n'est plus significatif, le modèle précédant est choisi. Il serait néanmoins intéressant de ne plus choisir mais d'estimer ce nombre de foyers, en l'introduisant dans le modèle en tant que paramètre aléatoire estimé à l'aide d'un critère de type AIC ou BIC.

Ce type de modèle de survie avec effet spatial est utilisable pour estimer les foyers d'une maladie mesurée par variable continue. La composante spatiale (effet foyer) est aussi utilisable dans des modèles plus classiques, pour analyser la survenue d'une maladie dans le temps (modèle de Cox) en lui ajoutant une dimension spatiale. Il s'étend aussi facilement au cas dichotomique en raisonnant sur l'odds-ratio spatialisé (modèle logit).

V.3. Conclusion

Les 2 méthodes de détection d'agrégats développées dans cette thèse sont complémentaires. Cette complémentarité est principalement due à la formalisation paramétrique ou pas. L'approche non paramétrique, basée sur la distance d'Hellinger, ne donne que des résultats qualitatifs, mais elle est spatialement flexible et n'impose aucun a priori sur les agrégats (nombre, forme, localisation). L'approche paramétrique par modèle de survie spatialisée permet d'obtenir des résultats quantitatifs, mais cette approche est spatialement plus rigide à cause de la paramétrisation de la forme des agrégats.

Les 2 méthodes sont aussi complémentaires d'un point de vue significativité. La méthode de la distance d'Hellinger teste en effet la significativité de la présence globale d'agrégation, mais pas la significativité locale de chaque agrégat détecté. Par contre la méthode de survie spatialisée teste la significativité de chaque foyer, mais pas la présence globale d'agrégation.

Par ailleurs, l'adaptation des méthodes développées dans la thèse à l'analyse temporelle et spatio-temporelle des agrégats est une perspective intéressante. D'autres développements sont nécessaires, et sont en cours d'élaboration.

VI. Perspectives

VI.1. Quelles autres approches sont envisageables ?

Cette thèse a été l'occasion de nombreux échanges scientifiques, qui ont ouvert de nouvelles pistes de réflexions et de travail. Ainsi nous allons présenter une autre approche de détection qui nous semble originale, que nous avons commencé à explorer. La poursuite de ces mises au point méthodologiques constitue donc une des perspectives intéressantes sur lesquelles pourrait déboucher la thèse.

Modélisation des agrégats par des processus ponctuels marqués

Dans cette thèse, nous avons examiné le processus d'agrégation de façon globale sur la zone d'étude, en considérant l'ensemble des données comme un seul ensemble de points marqués pouvant présenter des agrégats. Mais chaque agrégat potentiel peut aussi être considéré de façon distincte, et être modélisé de façon indépendante. Des modèles sophistiqués de processus ponctuels marqués peuvent être construits pour tenir compte simultanément des agrégats et de leurs interactions. Ils relèvent globalement de modèles de Gibbs pour processus ponctuels.

En s'inspirant d'une méthode de détection de réseaux fins - réseaux routiers dans des images (Stoïca et al., 2004), filaments cosmiques dans des données astronomiques (Stoïca et al., 2005) - un modèle a été développé pour la détection d'agrégats spatiaux, et mis en œuvre sur les données des infections mammaires (Stoïca et Gay, 2005).

Dans cette méthode la structure d'agrégat est considérée comme un ensemble d'objets, ici des disques, gouvernés par un processus ponctuel marqué, avec une densité de probabilité se décomposant en 2 termes :

- une énergie liée aux données, notée $U_d(\cdot)$, correspondant à un processus de Poisson non homogène modélisant la position des objets (les disques) dans l'image (zone d'étude) ;
- une énergie d'interaction entre les objets, notée $U_i(\cdot)$, qui est la superposition d'une interaction par aire et d'une interaction par paires de points.

La définition donnée à l'agrégat impose dans ce cas la détermination d'une valeur seuil, valeur qui est par la suite utilisée dans le terme d'énergie liée aux données du modèle : les points ne sont pris en compte que si la moyenne des valeurs dans le disque dépasse la valeur seuil.
L'estimation des agrégats se fait par maximum de vraisemblance, en ayant recours à des simulations effectuées en utilisant une des techniques de simulations de Monte Carlo, les dynamiques de Metropolis-Hastings.

La méthode de Stoïca et Gay (2005) a été mise en œuvre sur un jeu de données comprenant les SCS annuels pour l'ensemble des élevages bovins laitiers français - de plus de 20 animaux et de race pure Holstein - inscrits au Contrôle Laitier en 1996 (n=33890). Les agrégats principaux ont été détectés dans le centre de la France (Figure 3), résultat qui correspond à une zone détectée par les 2 méthodes développées dans cette thèse. Par contre, cette méthode n'a pas permis de détecter d'agrégat dans le département du Morbihan.

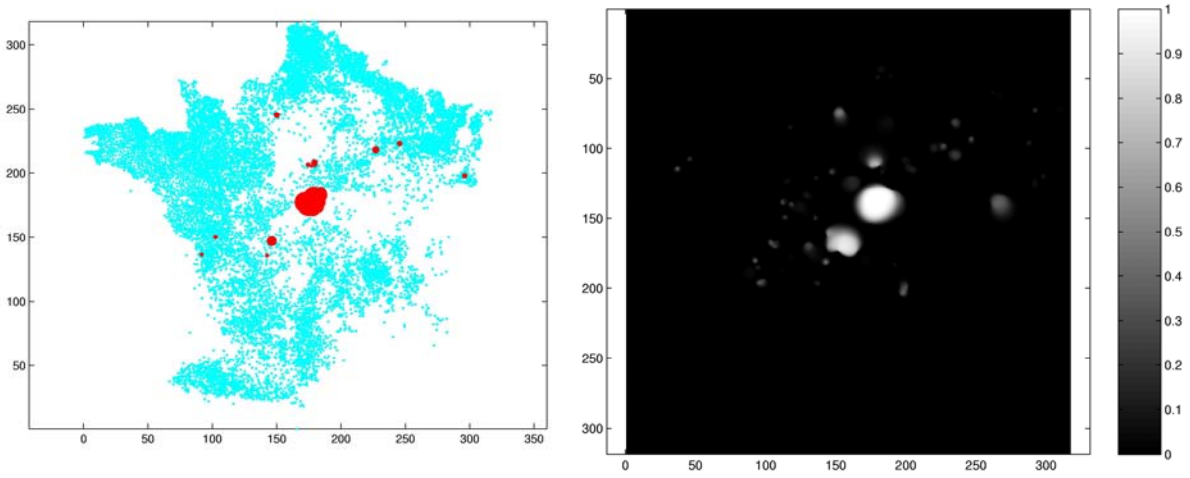


Figure 3 : Agrégats au sein de la variable du score annuel de cellules somatiques du lait, détectés par la méthode de modélisation des agrégats par des processus ponctuels marqués, au sein de la population d'élevages Holstein de race pure inscrits au Contrôle laitier en France (n=33890, année 1996)
A gauche : configuration d'une réalisation ; à droite : configuration moyenne (nombre de visites des disques)

Une analyse plus poussée des résultats et des implications du modèle utilisé en termes biologiques est nécessaire, et sera poursuivie.

VI.2. Interprétation biologique des agrégats

Nous avons, dans cette thèse, concentré notre travail sur les méthodes de détection d'agrégats. Mais l'étape suivante de la réflexion concerne l'interprétation de ces agrégats. Deux types de causes peuvent conduire à l'apparition d'un agrégat :

1) l'hétérogénéité de répartition des valeurs, qui est principalement due à des facteurs de risques de la maladie non connus ou non pris en compte ;
2) la non indépendance des cas, qui sous entend un phénomène de contagion, ou l'émergence d'un nouvel agent pathogène ou d'une nouvelle prédominance d'un agent connu.

L'hypothèse nulle, définie comme le fait que "les valeurs (ou les cas) sont réparties aléatoirement parmi la population à risque", sous entend en fait 2 hypothèses : 1) le fait que tous les individus de la population ont un risque égal face à la maladie, 2) l'indépendance des cas.

Il n'est pas facile de faire la différence entre les 2 types de causes liées à l'apparition d'agrégats. Dans le cas des infections mammaires, l'hétérogénéité due à des facteurs de risque non pris en compte est l'hypothèse explicative la plus probable, et les éléments en cette faveur sont à rechercher en première intention. Cependant, même pour cette maladie a priori non contagieuse inter élevages, nous savons que les mouvements de personnes, de matériel ou d'animaux (plus rares) peuvent être des facteurs de transmission passive de germes. De plus, l'émergence d'un nouvel agent est une hypothèse à considérer aussi, même si la probabilité d'un tel évènement semble assez faible.

Dans le cas de maladies avérées contagieuses, la distinction entre les 2 types de causes d'agrégation est encore plus difficile à faire. Les connaissances acquises sur les maladies étudiées peuvent permettre de formuler des hypothèses pour chaque agrégat détecté, en fonction de sa localisation, de sa forme et de son intensité. L'intégration, dans un deuxième temps, des éventuels facteurs de risque non pris en compte dans l'analyse première peut en outre constituer un moyen de confirmer ou infirmer les hypothèses émises. Enfin, la mise en œuvre d'une enquête épidémiologique locale est susceptible d'aider à tester les hypothèses de contagion.

Un deuxième niveau d'interrogation sur l'interprétation des agrégats concerne leur relation avec l'émergence. Il est certain que dans le cas d'une émergence de maladie, il semble naturel de vouloir détecter rapidement les agrégats de cas pour pouvoir déclarer l'émergence et prendre les mesures adaptées. Or, si un agrégat peut être lié à une émergence, ces 2 notions ne se recouvrent que partiellement. En effet, une émergence ne se traduit pas forcément par un (des) agrégat(s), et, comme nous l'avons déjà vu, un agrégat ne signe pas toujours l'apparition d'une émergence.

Si les méthodes de détection d'agrégats sont reconnues comme des outils utiles dans la détection des émergences (Kulldorff, 1999), le débat reste ouvert quant à la formalisation, à la fois biologique et mathématique, des liens entre agrégats et émergence.

Références

1. Ahrens, C., Altman, N., Casella, G., Eaton, M., Hwang, J.T.G., Staudenmayer, J. et Stefanescu, C., 2001. Leukemia clusters in update New York: how adding covariates changes the story. Environmetrics. 12, 659-672.
2. Ali, A.K.A. et Shook, G.E., 1980. An optimum transformation for somatic cell concentration in milk. J Dairy Sci. 63(3), 487-490.
3. Arnaud, M. et Emery, X., 2000. Estimation et interpolation spatiale : méthodes déterministes et méthodes géostatistiques. Hermes Science, Paris.
4. Banfield, J.D. et Raftery, A.E., 1993. Model-based gaussian and non-gaussian clustering. Biometrics. 49, 803-821.
5. Barkema, H.W., Schukken, Y.H., Lam, T.J.G.M., Beiboer, M.L., Benedictus, G. et Brand, A., 1998. Management practices associated with low, medium, and high somatic cell counts in bulk milk. J Dairy Sci. 81(7), 1917-1927.
6. Barnouin, J., Chassagne, M., Bazin, S. et Boichard, D., 2004. Management practices from questionnaire surveys in herds with very low somatic cell score through a national mastitis program in France. J Dairy Sci. 87(11), 3989-3999.
7. Barnouin, J., Geromegnace, N., Chassagne, M., Dorr, N. et Sabatier, P., 1999. Facteurs structurels de variation des niveaux de comptage cellulaire du lait et de fréquence des mammites cliniques dans 560 élevages bovins répartis dans 21 départements français. Anim Prod. 12, 39-48.
8. Barnouin, J. et Vourc'h, G., 2004. Les maladies émergentes : un défi pour le développement durable des productions animales. INRA Prod Anim. 17(5), 355-363.
9. Besag, J. et Newell, J., 1991. The detection of clusters in rare diseases. J R Statist Soc A. 154(Part 1), 143-155.
10. Brooker, S., Clarke, S., Njagi, J.K., Polack, S., Mugo, B., Estambale, B., Muchiri, E., Magnussen, P. et Cox, J., 2004. Spatial clustering of malaria and associated risk factors during an epidemic in a highland area of western Kenya. Trop Med Int Health. 9(7), 757-766.
11. Busato, A., Trachsel, P., Schallibaum, M. et Blum, J.W., 2000. Udder health and risk factors for subclinical mastitis in organic dairy farms in Switzerland. Prev Vet Med. 44(3-4), 205-220.
12. Carpenter, T.E., 2001. Methods to investigate spatial and temporal clustering in veterinary epidemiology. Prev Vet Med. 48(4), 303-320.
13. Chassagne, M., Barnouin, J. et Le Guenic, M., 2005. Expert assessment study of milking and hygiene practices characterizing very low somatic cell score herds in France. J Dairy Sci. 88(5), 1909-1916.
14. Clark, P.J. et Evans, F.C., 1954. Distance to nearest neighbor as a mesure of spatial relationships in populations. Ecology. 35, 445-453.
15. Clayton, M.K. et Kaldor, J., 1987. Empirical bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics. 43, 671-681.
16. Coulon, J.B., Dauver, F. et Garel, J.P., 1996. Facteurs de variation de la numération cellulaire du lait chez des vaches laitières indemnes de mammites cliniques. INRA Prod Anim. 9(2), 133-139.
17. Coulon, J.B., Gasqui, P., Barnouin, J., Ollier, A., Pradel, P. et Pomies, D., 2002. Effect of mastitis and related-germ on milk yield and composition during naturally-occurring udder infections in dairy cows. Anim Res. 51(5), 383-393.
18. Cressie, N.A.C., 1991. Geostatistics. Dans: Barnett V, Bradley RA, Fisher NIet al (eds), Statistics for spatial data. John Wiley and Sons, New York, pp. 58-67.

19. Cuzick, J. et Edwards, R., 1990. Spatial clustering for inhomogeneous populations. J R Statist Soc B. 52(1), 73-104.
20. David, S., Remontet, L., Bouvier, A.M., Faivre, J., Colonna, M. et Estève, J., 2002. Comment choisir en pratique le modèle permettant de décrire la variation géographique de l'incidence des cancers? Exemple des canccers digestifs de la Côte-d'Or. Rev Epidem et santé publ. 50(413-425).
21. DeGraves, F.J. et Fetrow, J., 1993. Economics of mastitis and mastitis control. Vet Clin North Am Food Anim Pract. 9(3), 421-34.
22. Diggle, P.J., 2000. Overview of statistical methods for disease mapping and its relationship to cluster detection. Dans: Elliott P, Wakefield JC, Best NG and Briggs D (eds), Spatial epidemiology : methods and applications. Oxford University Press, New York, pp. 87-103.
23. Diggle, P.J. et Chetwynd, A.G., 1991. Second-order analysis of spatial clustering for inhomogeneous populations. Biometrics. 47, 1155.
24. Diggle, P.J., Tawn, J.A. et Moyeed, R.A., 1998. Model-based geostatistics. Appl Statist. 47(3), 299-350.
25. Doran, R.J. et Laffan, S.W., 2005. 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. 70(1-2), 133-152.
26. Draper, N.R. et Smith, H., 1996. Applied regression analysis. John Willey & Sons, New York.
27. Droesbeke, J.J., Fichet, B. et Tassi, P., 1989. Analyse statistique des durées de vie : modélisation des données censurées. Economica, Paris.
28. Elliott, P. et Wakefield, J., 2001. Disease clusters: should they be investigated, and, if so, when and how? J R Statist Soc A. 164(part 1), 3-12.
29. Ely, L.O., Smith, J.W. et Oleggini, G.H., 2003. Regional production differences. J Dairy Sci. 86, E28-E34.
30. Ferrandiz, J., Lopez, A. et Sanmartin, P., 1999. Spatial regression models in epidemiological studies. Dans: Lawson A. B., Biggeri A., Böhning D.et al (eds), Disease mapping and risk assessment for public health. John Wiley and sons Ltd, Chichester, pp. 203-215.
31. Fèvre, E.M., Coleman, P.G., Odiit, M., Magona, J.W., Welburn, S.C. et Woolhouse, M.E., 2001. The origins of a new Trypanosoma brucei rhodesiense sleeping sickness outbreak in eastern Uganda. Lancet. 358(9282), 625-628.
32. Gangnon, R.E. et Clayton, M.K., 2000. Bayesian detection and modeling of spatial disease clustering. Biometrics. 56, 922-935.
33. Gangnon, R.E. et Clayton, M.K., 2004. Likelihood-based tests for localized spatial clustering of disease. Environmetrics. 15, 797-810.
34. Gastner, M.T. et Newman, M.E., 2004. Diffusion-based method for producing density-equalizing maps. Proc Natl Acad Sci USA. 101(20), 7499-7504.
35. Gay, E., Senoussi, R. et Barnouin, J., 2005. A spatial clustering analysis for continuous variables with application to milk somatic cell score in France. Biometrical Unit Research reports. Avignon, [http://www.avignon.inra.fr/stcavignon/centre/unites/biometrie/RR2005_7.pdf].
36. Gerbier, G., Bacro, J.N., Pouillot, R., Durand, B., Moutou, F. et Chadoeuf, J., 2002. A point pattern model of the spread of foot-and-mouth disease. Prev Vet Med. 56(1), 33-49.
37. Gibbs, A.L. et Su, F.E., 2002. On choosing and bounding probability metrics. Int Stat Rev. 70(3), 419-435.

38. Hanson, C.E. et Wieczorek, W.F., 2002. Alcohol mortality: a comparison of spatial clustering methods. Soc Sci Med. 55(5), 791-802.
39. Harmon, R.J., 1994. Physiology of mastitis and factors affecting somatic cell counts. J Dairy Sci. 77(7), 2103-2112.
40. Hill, C., Com-Nougué, C., Kramar, A., Moreau, T., O'Quigley, J., Senoussi, R. et Chastang, C., 1990. Analyse statistique des données de survie. INSERM Médecine-Sciences Flammarion, Paris.
41. Huffman, E.M., Mortimer, R., Olson, J.D., Ball, L. et Farin, P.W., 1984. Risk factors for prebreeding pyometra on four Colorado dairy farms. Prev Vet Med. 2(6), 785-790.
42. Kelsall, J.E. et Diggle, P.J., 1998. Spatial variation in risk of disease: a nonparametric binary regression approach. Appl Statist. 47(4), 559-573.
43. Klassen, A.C., Kulldorff, M. et Curriero, F., 2005. Geographical clustering of prostate cancer grade and stage at diagnosis, before and after adjustment for risk factors. Int J Health Geogr. 4, 1.
44. Kulldorff, M., 1997. A spatial scan statistic. Commun Stat-Theor M. 26(6), 1481-1496.
45. Kulldorff, M., 1999. Statistical evaluation of disease cluster alarms. Dans: Lawson A. B., Biggeri A., Böhning D.et al (eds), Disease mapping and risk assessment for public health. John Wiley and sons Ltd, Chichester, pp. 143-149.
46. Kulldorff, M. et Nagarwalla, N., 1995. Spatial disease clusters: detection and inference. Stat Med. 14(8), 799-810.
47. Laevens, H., Deluyker, H., Schukken, Y.H., De Meulemeester, L., Vandermeersch, R., De Muelenaere, E. et De Kruif, A., 1997. Influence of parity and stage of lactation on the somatic cell count in bacteriologically negative dairy cows. J Dairy Sci. 80(12), 3219-3226.
48. Lawson, A.B., 1995. MCMC methods for putative pollution source problems in environmental epidemiology. Stat Med. 14(21-22), 2473-2485.
49. Lawson, A.B., 2000. Cluster modelling of disease incidence via RJMCMC methods: a comparative evaluation. Stat Med. 19, 2361-2375.
50. Lawson, A.B., 2001. Statistical methods in spatial epidemiology. John Wiley & Sons Ltd, Chichester.
51. Lederberg, J., Shope, R.E. et Oaks, S.C.J., 1992. Emerging infections: Microbial threats to health in the United States. National Academy Press, Washington DC.
52. Manly, B.F.J., 1991. Randomization and Monte Carlo methods in biology. Chapman and Hall, London.
53. Marshall, R.J., 1991. A review of methods for the statistical analysis of spatial patterns of disease. J R Statist Soc A. 154(part3), 421-441.
54. Moran, P.A.P., 1950. Notes on continuous stochastic phenomena. Biometrika. 37, 17-23.
55. Morris, S.E. et Wakefield, J.C., 2000. Assessment of disease risk in relation to a pre-specified source. Dans: Elliott P, Wakefield JC, Best NG and Briggs D (eds), Spatial epidemiology: methods and applications. Oxford University Press, New York, pp. 153-184.
56. Norstrom, M., Pfeiffer, D.U. et Jarp, J., 1999. A space-time cluster investigation of an outbreak of acute respiratory disease in Norwegian cattle herds. Prev Vet Med. 47(1-2), 107-119.
57. Odoi, A., Martin, S.W., Michel, P., Middleton, D., Holt, J. et Wilson, J., 2004. Investigation of clusters of giardiasis using GIS and a spatial scan statistic. Int J Health Geogr. 3, 11.
58. Olea-Popelka, F.J., Griffin, J.M., Collins, J.D., McGrath, G. et Martin, S.W., 2003. Bovine tuberculosis in badgers in four areas in Ireland: does tuberculosis cluster? Prev Vet Med. 59(1-2), 103-111.

59. Oleggini, G.H., Ely, L.O. et Smith, J.W., 2001. Effect of region and herd size on dairy herd performance parameters. J Dairy Sci. 84(5), 1044-1050.
60. Openshaw, S., Charlton, M., Wymer, C. et Craft, A.W., 1987. A mark I geographical analysis machine for the automated analysis of point data sets. Int J Geogr Inf Syst. 1(335-358).
61. Ozdenerol, E., Williams, B.L., Kang, S.Y. et Magsumbol, M.S., 2005. Comparison of spatial scan statistic and spatial filtering in estimating low birth weight clusters. Int J Health Geogr. 4, 19.
62. Perez, A.M., Ward, M.P., Torres, P. et Ritacco, V., 2002. Use of spatial statistics and monitoring data to identify clustering of bovine tuberculosis in Argentina. Prev Vet Med. 56(1), 63-74.
63. Reneau, J.K., 1986. Effective use of dairy herd improvement somatic cell counts in mastitis control. J Dairy Sci. 69(6), 1708-1720.
64. Richardson, S., 1992. Modélisation statistique des variations géographiques en épidémiologie. Rev Epidem et santé publ. 40, 33-45.
65. Romain, H.T., Adesiyun, A.A., Webb, L.A. et Lauckner, F.B., 2000. 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. 47(4), 257-271.
66. Ross, A. et Davis, S., 1990. Point pattern analysis of the spatial proximity of residences prior to diagnosis of persons with Hodgkin's disease. Am J Epidemiol. 132(1 Suppl), S53-62.
67. Rupp, R., Boichard, D., Bertrand, C. et Bazin, S., 2000. Bilan national des numérations cellulaires dans le lait des différentes races bovines françaises. INRA Prod Anim. 13(4), 257-267.
68. Seegers, H., Fourichon, C. et Beaudeau, F., 2003. Production effects related to mastitis and mastitis economics in dairy cattle herds. Vet Res. 34(5), 475-491.
69. Selvin, S., Merrill, D., Schulman, J., Sacks, S., Bedell, L. et Wong, L., 1988. Transformations of maps to investigate clusters of disease. Soc Sci Med. 26(2), 215-221.
70. Senoussi, R., Chadoeuf, J. et Allard, D., 2000. Weak homogenization of points processes by space deformations. Adv Appl Prob (SGSA). 32, 948-959.
71. Silverman, B.W., 1986. The kernel method for univariate data. Dans: Cox DR, Hinkley DV, Rubin D and Silverman BW (eds), Density estimation for statistics and data analysis. Chapman and Hall, London, pp. 34-94.
72. Skrzypek, R., Wojtowski, J. et Fahr, R.D., 2004. Factors affecting somatic cell count in cow bulk tank milk--a case study from Poland. J Vet Med A Physiol Pathol Clin Med. 51(3), 127-131.
73. Stevenson, M.A., Benard, H., Bolger, P. et Morris, R.S., 2005. Spatial epidemiology of the Asian honey bee mite (Varroadestructor) in the North Island of New Zealand. Prev Vet Med. (in press).
74. Stoïca, R.S., Descombes, X. et Zerubia, J., 2004. A Gibbs point process for road extraction in remotely sensed images. Int J Comp Vision. 57(2), 121-136.
75. Stoïca, R.S. et Gay, E., 2005. Cluster detection in spatial data using marked point processes. Rapport de recherche n°11. Unité de Biométrie INRA Avignon.
76. Stoïca, R.S., Martinez, V.J., Mateu, J. et Saar, E., 2005. Detection of cosmic filaments using the Candy model. Astron Astrophys. 434, 423-432.
77. Tango, T., 1995. A class of tests for detecting 'general' and 'focused' clustering of rare diseases. Stat Med. 14(21-22), 2323-34.

78. Tango, T., 1999. Comparison of general tests for spatial clustering. Dans: Lawson A. B., Biggeri A., Böhning D.et al (eds), Disease mapping and risk assessment for public health. John Wiley and sons Ltd, Chichester, pp. 111-117.
79. Tango, T., 2000. A test for spatial disease clustering adjusted for multiple testing. Stat Med. 19(2), 191-204.
80. Tango, T. et Takahashi, K., 2005. A flexibly shaped spatial scan statistic for detecting clusters. Int J Health Geogr. 4, 11.
81. Thomas, A. et Carlin, B.P., 2003. Late detection of breast and colorectal cancer in Minnesota counties: an application of spatial smoothing and clustering. Stat Med. 22(1), 113-127.
82. Timander, L.M. et McLafferty, S., 1998. Breast cancer in West Islip, NY: a spatial clustering analysis with covariates. Soc Sci Med. 46(12), 1623-1635.
83. Turnbull, B.W., Iwano, E.J., Burnett, W.S., Howe, H.L. et Clark, L.C., 1990. Monitoring for clusters of disease: application to leukemia incidence in upstate New York. Am J Epidemiol. 132(1 Suppl), S136-143.
84. Wakefield, J.C., Kelsall, J.E. et Morris, S.E., 2000. Clustering, cluster detection and spatial variation in risk. Dans: Elliott P, Wakefield JC, Best NG and Briggs D (eds), Spatial epidemiology: methods and applications. Oxford University Press, New York, pp. 128-152.
85. Waller, L.A., 2000. A civil action and statistical assessments of the spatial pattern of disease: do we have a cluster? Regul Toxicol Pharmacol. 32(2), 174-183.
86. Ward, M.P. et Carpenter, T.E., 2000. Techniques for analysis of disease clustering in space and in time in veterinary epidemiology. Prev Vet Med. 45(3-4), 257-284.
87. Wartenberg, D., 2001. Investigating disease clusters: why, when and how? J R Statist Soc A. 164(part 1), 13-22.
88. Yang, G.J., Vounatsou, P., Zhou, X.N., Tanner, M. et Utzinger, J., 2005. A Bayesian-based approach for spatio-temporal modeling of county level prevalence of Schistosoma japonicum infection in Jiangsu province, China. Int J Parasitol. 35(2), 155-62.

Annexes

Programmes sous R

En gris clair : commentaires des programmes

Fonctions de lissage par noyau

```
# Estimateur densité de points par noyau
####################################

  kde2D <- function(x,y,c,r,xg,yg){

      # x,y coordonnées spatiales centrées normées
      # c constante de fenêtrage a  n^(-1/6) près
      # r côte de la grille (grille r*r)
      # xg, yg coordonnées des points de la grille

   m <- length(x)
   if(length(y)!=m)
     stop("Data vectors must be the same length")

   ha <- c*(m^(-1/6))

   # calcul des distances entre chaque point et chaque point de la grille,
   # divisées par la constante de fenêtrage (matrice 150*5000)
   xa <- outer(xg,x,"-")/ha
   ya <- outer(yg,y,"-")/ha

   # noyau gaussien
   d <- matrix(dnorm(xa),r,m)%*%
        t( matrix(dnorm(ya),r,m))

   d <- d/ha^2
   d <- d/sum(d)
   return(d)

   }

#########################################################
```

```
# Estimateur intensité des marques par noyau
########################################

   wkde2D <- function(x,y,z,h,r,d0) {

      # x,y coordonnées spatiales centrées normées
      # z marque du point
      # h constante de fenêtrage
      # r taille de la grille
      # d0 le seuillage du numérateur

   nx <- length(x)
   if((length(y)!=nx) & (length(z)!= nx))
      stop("Data vectors must be the same length")

   # préparation de la grille
   lims <- c(range(x),range(y))
   xgm <- seq(lims[1]-4*h,lims[2]+4*h,length=r)
   ygm <- seq(lims[3]-4*h,lims[4]+4*h,length=r)

   # distance entre les points et les points de la grille
   ax <- outer(xgm,x,"-")/h
   ay <- outer(ygm,y,"-")/h

   # noyau gaussien
   num <- matrix(dnorm(ax),r,nx) %*%  diag(z) %*%
       t(matrix(dnorm(ay),r,nx))
   den <- matrix(dnorm(ax),r,nx) %*%
       t(matrix(dnorm(ay),r,nx))

   gz <- num/den
   gz[den < d0] <- 0
   return(list(x=xgm,y=ygm,q=gz))

   }

################################################
```

Fonction de déformation radiale de l'espace

```
# Déformation radiale
 ##################

   radial <- function(V,p){
      # V coordonnées géographiques centrées et normées
      # p grille de densité (par noyau) du nuage de points V
   n <- length(V[,1])
   # détermination des indices à l'origine de la grille
   l0 <- min(which(xg >= 0)) ; k0 <- min(which(yg >= 0))

   # calculs des nouvelles coordonnées des points de V
   for (i in 1:n) {
      # repérage des points de la grille entre 0 et X
      l <- min(which(xg >= V[i,1]))
      if (V[i,1] <= 0){ l <- l-1; u <- xg[(l0-1):l]}
         else {u <- xg[l0:l]}
      k <- min(which(yg >= V[i,2]))
      if (V[i,2]<= 0){ k <- k-1; v <- yg[(k0-1):k]}
         else { v <- yg[k0:k]}
      c <- V[i,2]/V[i,1]      # coefficient de la droite rayon issue du centre de gravité
      # intersections du rayon OX avec la grille P1xP2
      uv <- c*u    # ordonnées correspondant aux u
      vu <- v/c    # abscisses correspondant aux v
      uu <- c(u,vu);    vv <- c(v,uv);          # toutes les intersections du rayon avec la grille
      L <- length(uu)
      if (V[i,1] >= 0){ uu <- sort(uu) }
         else { uu <- -sort(-uu)}
      if (V[i,2] >= 0){ vv <- sort(vv)}
         else { vv <- -sort(-vv)}
      S <- sqrt((V[i,1])^2 +(V[i,2])^2)        # longueur du segment
      # première valeur doublée
      s <- 0;  tau <- 0
      for (j in 2:L) {
         if ((abs(uu[j]) <= abs(xg[l])) && (abs(vv[j])<= abs(yg[k]))){
            ds <- sqrt((uu[j]-uu[j-1])^2 +(vv[j]-vv[j-1])^2 )/S
            s <- s+ds
            # éléments de la grille correspondant aux uu, vv
            f <- max(which(xg <= uu[j]))
            h <- max(which(yg <= vv[j]))
            tau <- tau+p[f,h]*s*ds
            }
         }
      V[i,] <- V[i,]*sqrt(2*tau)
   }
   return(V)
}
############################################################
```

Fonction de calcul de la distance d'Hellinger

```
#  Distance d'Hellinger entre 2 niveaux de la marque
##############################################

  Hell <- function(x,y,z,a,c,r,xg,yg,p) {

   # x,y coordonnées
   # z marque du processus ponctuel initial (niveau de base)
   # a niveau du score z que l'on compare au niveau de base
   # c constante  de fenêtrage a  n^(-1/6) près
   # r côte de la grille (grille r*r)
   # xg, yg coordonnées des points de la grille
   # p résultat de l'interpolation au niveau de base z=0
   # nl nombre de lignes de la matrice

  I <- which(z >= a)    # choix du niveau de z
  na <- length(I)    # coordonnées des points avec niveau z > a
  nl <- length(x)

  if (na==0){E[3,]<-1} else {

    # application de la fonction de lissage
    q <- kde2D(x[I],y[I],c,r,xg,yg)

    # option d'affichage des intensités de points à chaque niveau a de z
      colori <- sort(heat.colors(30),decreasing=T)   # gamme de couleurs
      titre <- c("Densité d'élevages au niveau de score :",as.character(round(a,digits=2)))
      image(xg,yg,q,col=colori,main=titre,
         xlab="abscisse après déformation", ylab="ordonnée après déformation")
      contour(xg,yg,q,add=T,nlevels=6)

    # fonction d'Hellinger  entre Pa et P initiale
    He <- sqrt(sum((sqrt(p)-sqrt(q))^2)/2)
    E <- c(a,na/nl,He)}
    return(E)   # la matrice E a 3 lignes :
                # E[1,] = niveau de z où la distance est calculée
                # E[2,] = pourcentage de points à ce niveau (effectif)
                # E[3,] = distance d'Hellinger
}

#######################################################
```

Programme d'application du calcul de la distance d'Hellinger

```
# Calcul de la distance d'Hellinger sur fichier d'1 année et permutations dans l'espace
#############################################################################

        # e96 - e00 : fichiers de travail pour les années 96 à 2000
        # z : marque des points pour l'année étudiée

# Distance d'Hellinger observée

      # s niveaux (sans le min et le max)
       # s+1 intervalles
       # s+2 valeurs
       s<-15
       D<-cbind(e96$res,e97$res,e98$res,e99$res,e00$res)
       dz<-(max(D)-min(D))/(s+1)   # taille des (s+1)intervalles
       a<-min(D)+(0:(s+1))*dz    # (s+2) valeurs des niveaux de score

   # Calcul de la distance d'Hellinger selon la cote z observée et graphique

       E<-rep(1,(s+2))   # 1 pour imposer dernier niveau=1
       pp<-rep(0,(s+2))   # vecteur pourcentage de points aux (s+2) valeurs
       for(i in 1:(s+1)){
          E[i] <- Hell(xr,yr,z,a[i],cr,rr,xgr,ygr,pr)[3]
          pp[i] <- length(which(z >= a[i]))/nl
          }

       x11()
       plot(a,E,pch="*",col="blue",ylim=c(0,1),xlab="niveau de score", ylab="distance d'Hellinger",
          main="Distance d'Hellinger aux différents niveaux de score")
       points(a,pp,pch="+")  # pourcentage de points à ce niveau z

   # Calcul du Hellinger total

   ds<-pp[1:(s+1)]-pp[2:(s+2)]    # pourcentage de points spécifiques au niveau
   Hel<-sum(E[2:(s+2)]*ds)
   print(Hel)

############################

# Tests de permutations dans l'espace et graphique avec bornes de confiance

   t <- 1000   # nombre de permutations
   alpha<-0.05  # risque alpha pour la significativité
   ialpha<-round(alpha*t/2,0)  # borne minimum avec risque bilatéral
   Ialpha<-t-ialpha    # borne maximum avec risque bilatéral
```

```
# préparation des matrices
EE<-matrix(1,nr=t,nc=(s+2))
ppp<-rep(0,(s+2))
m<-rep(0,(s+2)) ; M<-rep(0,(s+2))
Hel2<-rep(0,t)

# permutations et calculs de la distance d'Hellinger
for(i in 1:t){
zz <- z[sample(seq(1,nl,1),nl)];    # permutations dans l'espace
   for(j in 1:(s+1)){
   EE[i,j]<-Hell(xr,yr,zz,a[j],cr,rr,xgr,ygr,pr)[3]    # calculs d'Hellinger
   ppp[j] <- length( which( z >= a[j] ))/nl
   }
print(i)
ds<-ppp[1:(s+1)]-ppp[2:(s+2)]   # pourcentage de points spécifiques au niveau
Hel2[i]<-sum(EE[i,2:(s+2)]*ds)  # calcul du Hellinger total pour chaque permutation
}

# Calcul de l'intervalle de confiance
for (j in 1:(s+2)) {
EE[,j]<-sort(EE[,j])
m[j]<-EE[ialpha,j]  # points de confiance inférieure
M[j]<-EE[Ialpha,j]  # points de confiance supérieure
}

# Hellinger total de l'ensemble des permutations
Hel2<-sort(Hel2)
Hel2min<-Hel2[ialpha]
Hel2max<-Hel2[Ialpha]
print(Hel2min)
print(Hel2max)

# Nouveau data frame
R<-list(data.frame(pp,E,m,M),data.frame(Hel,Hel2min,Hel2max),EE,Hel2)
   # R[[1]]: % de points, Hellinger pour l'année, He_max par points, He_min par points
   # R[[2]]: valeurs globales : He_année, He_min, He_max
   # R[[3]]: Hellinger pour les simulations
   # R[[4]]: Hellinger total pour les simulations

# graphique
x11()
plot(a,R[[1]][,2],pch=20,col="black",ylim=c(0,1), xlab="level", ylab="Hellinger distance",
   main="Distance d'Hellinger aux différents niveaux de score")
points(a,R[[1]][,3], type="l", col="gray40", lty=2)
points(a,R[[1]][,4], type="l", col="gray40", lty=2)
points(a,R[[1]][,1], pch="*",col="gray")
title(sub="15 niveaux, 1000 permutations, alpha=5%, * = % de points", cex.sub=0.8,
   adj=1,font.sub=3)

#################################################################
```

Programme du modèle de survie spatialisée

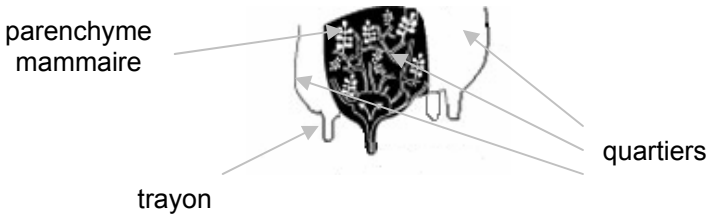
```
# Modélisation spatio-temporelle : K foyers
####################################

# Sélection des données
   z <- D$score  # variable à modéliser
   D1<-cbind(rgmoy,nbvelete,nbvelaut,nbvelhiv,nbvelprint,nbvl2,nbvl3,nbvl4) # FR
   D2 <- cbind(x,y)  # coordonnées géographiques
   nl <- dim(D)[1]

# préparation des données
   # ordonner les tableaux selon z croissant
   I  <-  order(z,c(1:nl))
   z  <-  z[I]
   D1 <-  D1[I,] ; D2 <- D2[I,]
   # traitement des valeurs z identiques
   Delta <- rank(z, ties.method = "min" )

#################

# calcul du -log de la fonction de vraisemblance
   lmv  <- function(alpha,beta){
   k <- length(beta)
   # calcul de sommation
   aa <- c(D1 %*% matrix(beta,nrow=k,ncol=1)) # beta W (partie FR)
    # partie foyer
   norm<-matrix(0,nr=nl,nc=K)
   bb<-matrix(0,nr=nl,nc=K)
   for (i in 1:K){
         norm[,i]<-(D2[,1]-alpha[3,i])^2 +(D2[,2]- alpha[4,i])^2  # distance ||x-c||²
         # effet foyer : (alpha/(2*pi*rho²))*exp(-||x-c||²/2rho²)
         bb[,i]<-(alpha[1,i]/(2*pi*(alpha[2,i]^2)))*exp(-norm[,i]/(2*(alpha[2,i]^2)))
         }
   # beta W - effet foyer
   ab<-aa-apply(bb,1,sum)
   expab <- exp(ab)
   # calcul des ensembles a risque (inverser,cumuler inverser)
   expab <- cumsum(expab[nl:1])
   expab <- expab[nl:1]
   # tenir compte des multiplicité de points (Delta >0)
   cc <- log(expab[Delta])      # partie 2 de la log vraisemblance partielle
   # log de la vraisemblance partielle
   mv <- sum(ab) - sum(cc)
   return(-mv)
}

 ######################
```

```
# Nombre de foyers
K<-3

# Algorithme  de recherche de MV
library(stats4)

   # Transformation entre l'écriture avec tous les paramètres et l'écriture avec alpha beta
   ll <- function(a1,a2,a3,a4,a5,a6,a7,a8,a9,a10,a11,a12,a13,a14,a15,a16,a17,a18,a19,a20){
         # beta : ensemble des FR
         beta <- c(a1,a2,a3,a4,a5,a6,a7,a8)
         # alpha : K fois (alpha, rho, x du foyer, y du foyer)
         alpha  <- cbind(c(a9,a10,a11,a12),c(a13,a14,a15,a16),c(a17,a18,a19,a20))
         lmv(alpha,beta)
   }

   # Optimisation de la vraisemblance
   V <- mle(ll,
      method="Nelder-Mead",
      start=list(a1=-0.5,a2=0.003,a3=0.006,a4=-0.01,a5=-0.01,a6=0.02,a7=-0.01,a8=0.01,
        a9=0,a10=0.1,a11=0,a12=0,a13=0,a14=0.1,a15=0,a16=0,a17=0,a18=0.1,a19=0,a20=0),
      control=list(trace=1, maxit=15000))

   print(summary(V))
   print(logLik(V))

##############################################
```

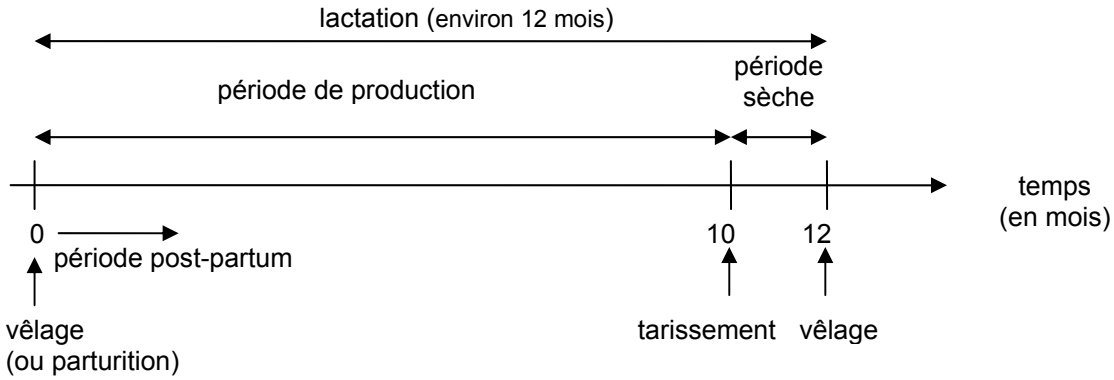
Glossaire

Termes biologiques relatifs aux infections mammaires

Anatomie de la mamelle de la vache laitière



Cycle de production de la vache laitière



Comptage de cellules somatiques : numération des leucocytes présents dans le lait, variable utilisée pour mesurer de degré d'inflammation mammaire, et donc dépister les mammites.

Fibrose : transformation fibreuse des formations tissulaires.

Mammite : inflammation de la glande mammaire.

Mammites subcliniques : forme d'infection mammaire inapparente, non détectée par le seul examen clinique, dépistée par le comptage de cellules somatiques du lait.

Primipare : animal ayant vêlé une seule fois, et donc en rang de lactation 1 (ou en 1ère lactation).

Rang de lactation : nombre de lactations effectuées par une vache à un moment donné (environ 4 à 5 lactations durant la carrière d'une vache laitière).

Score de cellules somatiques (SCS) : transformation logarithmique du comptage de cellules somatiques (CCS) :

$$SCS = \log_2\left(CCS / 100\,000\right) + 3$$

Stade de lactation : positionnement temporel (jours ou mois) au sein de la lactation.

Termes statistiques

Approche bayésienne : démarche statistique qui intègre de l'a priori, sous forme de lois de probabilité, sur les paramètres statistiques que l'on cherche à estimer.

Autocorrélation spatiale : fonction estimant les corrélations (indice statistique indiquant le degré de liaison) entre les différentes valeurs d'une même variable selon la distance qui sépare les points où elle est mesurée.

Champ aléatoire : ensemble des valeurs d'une variable aléatoire définie en tout point d'un domaine spatial. Il est dit homogène (ou stationnaire) si les distributions de tout ensemble de valeurs du champ est invariant par translation. Un champ stationnaire au 2ème ordre a une moyenne constante et une covariance ne dépendant que de l'écart entre les points.

Corrélogramme : graphe représentant l'autocorrélation en fonction de la distance. Le corrélogramme ρ est un rapport de covariogrammes, le covariogramme C mesurant la covariance spatiale entre les valeurs d'une même variable Z prises en x et $x+h$:

$$C(h) = \mathrm{cov}\left[Z(x), Z(x+h)\right]$$

$$\rho(h) = \frac{C(h)}{C(0)}$$

Le covariogramme pour une distance nulle $C(0)$ est la covariance entre un point et lui-même, c'est-à-dire la variance locale.

Développement asymptotique : Obtention de lois limites quand le nombre d'observations est jugé suffisamment grand.

Effet pépite : se produit quand au champ aléatoire se superpose un bruit blanc (avec variance σ_0 mais de covariance nulle), dont on ne peut tirer aucune information en dehors de son amplitude σ_0. La fonction de semi-variance γ est alors discontinue en 0 et on a :

$$\sigma_0 = C(o) - \lim_{h \to 0^+} C(h)$$

Fenêtres mobiles : zones de travail délimitées dans l'espace d'étude et balayant ce dernier.

Méthodes MC (Monte Carlo) : méthodes statistiques de simulation d'une distribution. Les méthodes MCMC (Monte Carlo Markov Chain) s'appliquent en particulier quand la distribution n'est que partiellement connue, en utilisant des techniques de chaînes de Markov.

Moment d'ordre k : un moment d'ordre k d'une variable Z relative à une constante donnée c est défini comme la valeur espérée de la variable aléatoire $(Z- c)^k$, c'est-à-dire $\mathrm{E}(Z- c)^k$. Si c est égal à la moyenne de population, le moment est dit centré. L'espérance est le moment d'ordre 1, la variance un moment d'ordre 2.

Procédures de randomisation : techniques de ré-échantillonnages répétés à partir de l'ensemble des données disponibles pour évaluer la variabilité de l'estimation de paramètres. Les techniques les plus classiques sont les permutations et le bootstrap.

Proches voisins : technique d'analyse considérant les voisins d'un point c, classés selon leur distance par rapport à ce point c. L'ordre de voisinage est soit défini par le nombre de voisins, soit par une distance maximale au point c.

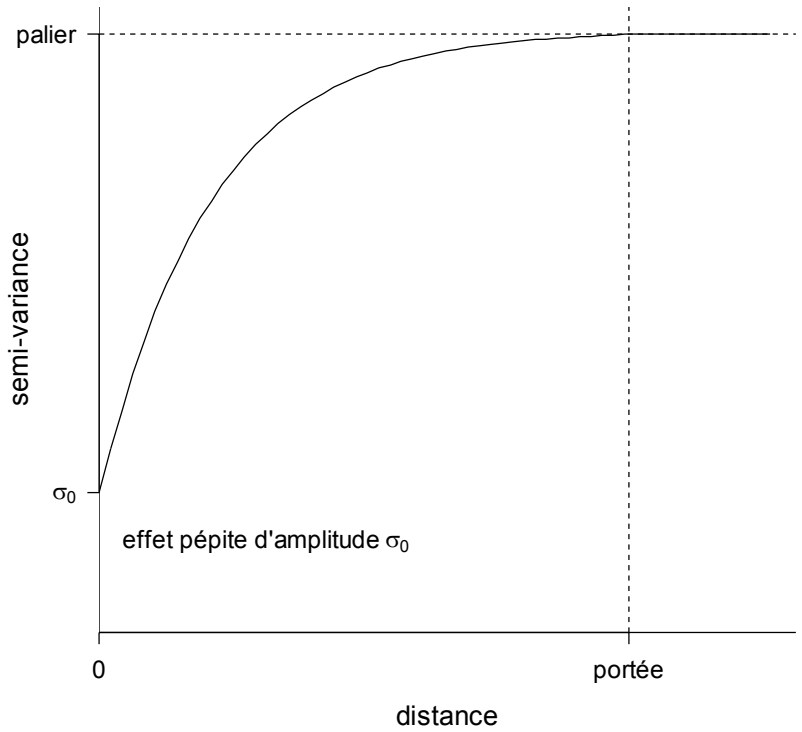
Sur dispersion : on dit qu'il y a sur dispersion quand il subsiste des écarts importants entre estimations du modèle et valeurs observées (ces valeurs observées sont plus dispersées que les valeurs estimées).

Variogramme (ou semi-variogramme, mais le préfixe est souvent omis) : graphe représentant la semi-variance en fonction de la distance. La semi-variance est un moment d'ordre 2, qui mesure la dissemblance entre les valeurs d'une même variable Z prises en x et $x+h$:

$$\gamma(h) = \frac{1}{2}\mathrm{var}\left[Z(x) - Z(x+h)\right]$$

Il est à noter que : $\gamma(h) = C(0) - C(h)$, C étant le covariogramme (cf. corrélogramme).

Exemple de variogramme :



Spatial cluster detection with continuous variables: application to an indicator of bovine mastitis

Cluster detection is an important field of investigation in epidemiology. Methods of cluster detection had mainly been developed for diseases described through dichotomic variables (case/control). But cattle intramammary infections are evaluated by the somatic cell score, a continuous variable expressing milk leukocyte count. Consequently, the thesis objective was to develop methods of spatial cluster detection, dealing with continuous variables, to analyse the spatial distribution of somatic cell score in French bovine dairy herds.

The first approach developed, which is non parametric, is based on the Hellinger distance between spatial distributions. The measure of the distance between farm distribution, at different levels of somatic cell score, and the global farm density allowed to highlight a progressive concentration in some specific areas, and thus to detect clusters. The second approach, which is parametric, uses a model quantifying spatialised survival risk for increasing score levels. It integrates known risk factors for the cellular score, as well as a "foci-effect" with a Gaussian form. The clusters detected with this method were the same as the ones highlighted by the method based on the Hellinger distance.

The two methods are new flexible and generic tools for spatial cluster detection. They show several advantages: 1) they deal with diseases measured by continuous or discrete variables, 2) they account for known risk factors of the disease, and focus on unexplained clusters, and 3) they take into account the spatial heterogeneity of background population.

Keywords: spatial statistics, cluster detection, Hellinger distance, survival analysis, udder infection

Détection d'agrégats spatiaux dans le cas d'une variable continue : application à un indicateur de l'infection mammaire chez les bovins

---

La détection des agrégats est une problématique d'importance en épidémiologie. Les méthodes de détection d'agrégats ont essentiellement été développées pour des maladies mesurées par des variables dichotomiques (cas/témoins). Or, dans le cas des infections mammaires des bovins, la maladie est évaluée par le score cellulaire, variable continue qui est une expression de la numération leucocytaire du lait. En conséquence, l'objectif de la thèse a consisté à développer des méthodes de détection d'agrégats spatiaux adaptées aux variables continues, afin d'analyser la répartition spatiale du score cellulaire dans les élevages bovins laitiers français.

La première approche qui a été développée, non paramétrique, est basée sur la distance d'Hellinger entre distributions spatiales. La mesure de la distance entre la distribution des élevages à différents niveaux de score cellulaire et la densité globale d'élevages a permis de mettre en évidence une concentration progressive dans certaines zones, et donc de détecter des agrégats. Une seconde approche, paramétrique, utilise un modèle quantifiant la survie spatialisée, pour des niveaux de score croissants, en intégrant les facteurs de risque connus du score cellulaire et un "effet-foyers" de forme gaussienne. Les agrégats détectés par cette méthode ont été identiques à ceux obtenus avec la méthode basée sur la distance d'Hellinger.

Les deux méthodes constituent de nouveaux outils flexibles et génériques pour la détection d'agrégats spatiaux. Elles offrent en effet plusieurs avantages : 1) elles sont applicables à des maladies mesurées par une variable continue ou discrète, 2) elles prennent en compte les facteurs de risque connus de la maladie pour ne s'intéresser qu'aux agrégats inexpliqués, 3) elles tiennent compte de l'hétérogénéité spatiale de la population sous-jacente.

Mots clés : statistiques spatiales, détection d'agrégats, distance d'Hellinger, analyse de survie, infection mammaire

---

Discipline : Epidémiologie

Institut National de la Recherche Agronomique

Unité d'Epidémiologie Animale
INRA Clermont-Ferrand Theix
63122 St-Genès Champanelle

Unité de Biométrie
INRA Domaine St-Paul, Agroparc
84914 Avignon Cedex 9