EVALUATING THE SAMPLING PATTERN WHEN DETECTING ZONES OF ABRUPT CHANGE

BY

Edith Gabriel

Denis Allard

Research Report No. 12

August 2005

Unité de Biométrie

Institut National de la Recherche Agronomique

Avignon, France

http://www.avignon.inra.fr/biometrie

Evaluating the sampling pattern when detecting Zones of Abrupt Change

Edith Gabriel[1] and Denis Allard[2]

[1]*Department of Mathematics and Statistics, Lancaster University, Fylde College, Lancaster LA14YF, United Kingdom*,

[2]*Institut National de la Recherche Agronomique, Unité de biométrie, Domaine Saint-Paul, site Agroparc, 84914 Avignon, France.*

Abstract: We present a method for detecting the zones where a variable irregularly sampled in the plane varies abuptly. Such zones are called Zones of Abrupt Change (ZACs). This method not only allows to estimate ZACs, but also to test their statistical significance against the null hypothesis of a stationary correlated random field. The sampling pattern, in particular its local density is crucial in the detection of potential ZACs. In this paper, we address the problem of evaluating the sampling pattern by assessing the power of the local test used for detecting ZACs. It is shown that mapping the power allows us to identify zones where ZACs may be detected or not. The methodology is illustrated on a soil data set sampled at 8 different dates in an agricultural field. Analyzing the soil water content for ZACs allowed to point out permament structures in the agricultural field related to the boundaries between different soil types. Mapping the power for various sampling density proved to be useful to determine the minimal sampling density necessary for detecting ZACs.

Keywords: Boundary detection, Power, Test, Wombling.

1 Introduction

In a wide class of research area such as ecology, biology or soil science, it is often of interest to map the high spatial variations of the variables under study. A first example can be found in biology, where spatial abrupt variations of allele frequencies of some species may indicate changes in populations (Barbujani, Oden and Sokal, 1989). In the case of human populations, sharp genetic variations can be related to culture and languages (Pagel and Mace, 2004). A second motivating example, used in this paper to illustrate our methodology, originated from

soil sciences: abrupt variations of soil characteristics can be linked to boundaries between soil types.

In this paper we propose a method for detecting the zones where a variable irregularly sampled in the plane changes abruptly and we explore its power, *i.e.* its location dependent ability to detect these zones. In our model, the null hypothesis is "the observed variable has a constant expectation on the study domain $\mathcal{D}$". The alternative is the existence of discontinuities of the expectation along a set of curves Γ. No assumption is made on the shape of Γ. Considering the case of an irregular sparse sampling, we cannot estimate precisely the curves, but rather the zones with sharp variations, *i.e.* with high gradient, hereafter called Zones of Abrupt Change (ZACs).

The estimation of ZACs is an issue related to boundary estimation, edge detection, or change curve estimation. The problem of estimating discontinuity curves from irregularly located samples has not been much addressed. It was first considered by Womble (1951) to illustrate that abrupt changes in allele frequencies can be linked to the boundaries between different populations. The method was quite simple: the variable under study was linearly interpolated on a regular grid. A gradient vector was then approximated by computing differences. The wombling and its extensions (Barbujani, Oden and Sokal, 1989; Oden et al., 1993; Bocquet-Appel and Bacro, 1994; Fortin 1994) define the top 5% (or 10%) of the gradient values as "barriers". A recent review of wombling-based methods is offered in Jacquez, Maruca and Fortin (2000).

Hall and Rau (2001) proposed a tracking method based on spatial approximations of the local likelihood that the discontinuity curve passes through a given point in the plane, as a function of this point. This method presents several drawbacks: since the method is based on tracking, the existence of a discontinuity curve is implicitly assumed without testing its existence; a starting point is needed in the vicinity of the discontinuity; practical implementation requires that the starting point is on the edge of a rectangular study domain, hence implying that the discontinuity intersects it.

More recently, Banerjee, Gelfand and Sirmans (2003), proposed a Bayesian approach to estimate the gradient surface of an irregularly sampled spatial process. It provides local estimates of the gradient along with their posterior predictive intervals. However this method is not specifically designed for detecting boundaries and Zones of Abrupt Change. In particular, it does not provide a way to test globally the presence of ZACs in the study domain.

The aforementioned methods do not test the significance of the detected barriers or discontinuity curves, *i.e.* they do not provide a test to decide if the detected pattern is "due by chance" or if it is because "there is really something there". For example, use of thresholds to identify barriers is subjective, in that, barriers are always found, whether or not their rates of change are statistically significant. In Fortin and Drapeau (1995) and Gleyze, Bacro and Allard (2001), attempts were made to assess the significance of the barriers, but both assumed an uncorrelated variable, an unrealistic hypothesis in environmental sciences. Jacquez and Maruca (1998) involve local and global statistics to determine where statistically significant barriers are and whether the barriers for the entire surface are statistically significant or explained by chance. They proposed several null hypotheses, including absence of correlation and restricted permutations of barriers. Since these methods do not take into account the spatial structure of the variable in the barriers definition, they are of limited interest.

Godtliebsen, Marron and Pizer (2002) proposed a scale-space approach for finding significance in spatial data, with view to clustering. This scale-space approach is a spatial generalization of the method called SiZer developed by Chaudhuri and Marron (1999, 2000) in one dimension. SiZer computes the kernel smoothing for a whole range of bandwidth h, assesses the significance of the slope for all h and proposes a single visualization of the regions of significance in the scale space. This visualization is difficult to extend to the two dimensional case. Visualization is done by movies (time corresponding to bandwidth) and with pictures using streamlines along significant gradient vectors and contours. This method requires heavy computations and visualization is difficult to interpret.

In Allard, Gabriel and Bacro (2005) a method for detecting Zones of Abrupt Change (ZACs) from an irregularly sampled variable is proposed. The variable is modeled as a random field $Z(\cdot)$ defined on $\mathcal{D} \subset \mathbb{R}^2$. ZACs are defined as a discontinuity or a sharp variation of the local expectation of $Z(\cdot)$. This method not only allows to estimate ZACs, but also to test their statistical significance against the null hypothesis of a stationary correlated random field. In order to detect globally the existence of ZACs on $\mathcal{D}$, a two-step method is built. First, local tests are based on the estimated local gradient. The null hypothesis $H_0(\mathbf{x})$: "$\mathbb{E}[Z(\mathbf{y})]$ is constant for all $\mathbf{y}$ in a neighborhood around $\mathbf{x}$", is tested against the alternative, $H_1(\mathbf{x})$: "$\mathbf{x}$ belongs to Γ", where Γ is a curve on which the expectation of the random field is discontinuous. Second, the local tests are aggregated in a global one to determine the significance of the detected ZACs on

$\mathcal{D}$ and hence to discriminate between the two hypothesis H_0: "$Z(\cdot)$ is second order stationary" and H_1: "$\mathbb{E}[Z(\cdot)]$ is piecewise constant, with discontinuities on Γ".

To have a good estimation of the local gradient, the sampling pattern must be locally dense enough. Indeed, a lack of sample points does not allow to test if the estimated local gradient correspond to a smooth transition or a locally sharp transition. In this paper, we address this question by investigating the power of the ZACs detection test, *i.e.* the probability to detect an abrupt change at a point $\mathbf{x}$, under the hypothesis of the existence of a discontinuity curve Γ containing $\mathbf{x}$. To assess the power, the alternative hypothesis must be fully specified: we will assume that Γ has a regular shape, can be locally approximated by its tangent and that the discontinuity is represented by a "buttonhole" around the tangent as precised in Section 3.1. The local power is then calculated at each point $\mathbf{x}$ in $\mathcal{D}$. Several local test statistics are used to assess the local power. Since these local test statistics are highly dependent, we must take these dependencies into account in the local power assessment. Mapping local power shows clearly that the power is not constant on the domain. Zones with low power indicate that the local sampling pattern is not appropriate for estimating ZACs, in particular because the local sample density is too low.

This paper is organized as follows. The method of ZAC detection is described in Section 2. Definition and computation of the power is presented in Section 3. In Section 4, we analyse the soil water content of an agricultural field, for which the method detects ZACs related to sharp transitions between different zones. Then, we show how the sample density directly affects the possibility of detecting ZACs. The paper concludes with a short discussion.

2 Detection of Zones of Abrupt Change

2.1 General background

In this section we only present the main features of the method for detecting ZACs and we refer to Allard, Gabriel and Bacro (2005) and Gabriel (2004) for more details.

Let us denote $Z(\cdot)$ the random field modeling the study variable. Let $\mathbf{Z} = (Z(\mathbf{x}_1), ..., Z(\mathbf{x}_n))'$ be a sample of $Z(\cdot)$ at the locations $\mathbf{x}_1, ..., \mathbf{x}_n$ in $\mathcal{D}$. Under the null hypothesis H_0, we assume second order stationarity of $Z(\cdot) : \mathbb{E}[Z(\mathbf{x})] = m$ for all $\mathbf{x}$ in $\mathcal{D}$ and $\mathrm{Cov}\,(Z(\mathbf{x}), Z(\mathbf{y})) = C_Z(\mathbf{x}-\mathbf{y})$, for all $\mathbf{x}, \mathbf{y}$ in $\mathcal{D}$. For sake of simplicity, we assume that the covariance function $C_Z(\mathbf{h})$ is

continuous everywhere and regular enough (3 times differentiable for all $\|\mathbf{h}\| > 0$). This is for example the case for the exponential covariance function: $C_Z(\mathbf{h}) = \exp(-\|\mathbf{h}\|/b)$, with $b > 0$. In the following, $\mathbf{v}'$ (resp. $\mathbf{M}'$) denotes the transpose vector of any vector $\mathbf{v}$ (resp. matrix $\mathbf{M}$).

Under the gaussian assumption, when m is unknown, the optimal predictor of $Z(\mathbf{x})$ at an unsampled location $\mathbf{x}$ is the ordinary kriging (Chilès and Delfiner, 1999, chap. 3.4),

$$Z^*(\mathbf{x}) = C'(\mathbf{x})\mathbf{C}^{-1}\mathbf{Z} + \big(1 - C'(\mathbf{x})\mathbf{C}^{-1}\mathbf{1}\big)\frac{\mathbf{1}'\mathbf{C}^{-1}\mathbf{Z}}{\mathbf{1}'\mathbf{C}^{-1}\mathbf{1}}, \tag{1}$$

where $C(\mathbf{x}) = \big(C_Z(\mathbf{x} - \mathbf{x}_1), ..., C_Z(\mathbf{x} - \mathbf{x}_n)\big)'$ is the covariance vector between $\mathbf{x}$ and the sample locations, $\mathbf{C}$ is the covariance matrix between the sample locations and $\mathbf{1}$ is the n-vector $(1, \ldots, 1)'$.

Under the regularity conditions of the covariance function, the gradient of the predictor is:

$$W(\mathbf{x}) = \partial Z^*(\mathbf{x}) = \partial C'(\mathbf{x})\mathbf{C}^{-1}\left(\mathbf{Id} - \frac{\mathbf{1}\mathbf{1}'\mathbf{C}^{-1}}{\mathbf{1}'\mathbf{C}^{-1}\mathbf{1}}\right)\mathbf{Z} = \partial C'(\mathbf{x})\mathbf{K}^{-1}\mathbf{Z}, \tag{2}$$

where $\partial C(\mathbf{x}) = \left(\frac{\partial C(\mathbf{x})}{\partial x^1} \quad \frac{\partial C(\mathbf{x})}{\partial x^2}\right)'$ is the gradient of $C(\mathbf{x})$ at $\mathbf{x} = (x^1\ x^2)' \in \mathcal{D}$ and $\mathbf{Id}$ is the $n \times n$ identity matrix. Note that since the derivatives of $C_Z(\mathbf{h})$ are computed away from $\mathbf{h} = \mathbf{0}$, no regularity condition at the origin is imposed on the covariance function. If $C_Z(\mathbf{h})$ is not twice differentiable for $\mathbf{h} = \mathbf{0}$, which is for example the case for an exponential covariance function, $Z(\mathbf{x})$ has no derivative but $Z^*(\mathbf{x})$ is differentiable, except at the sample locations. In this case, we will consider by extension $W(\mathbf{x})$ as a local predictor of the gradient (Chilès and Delfiner, 1999, p. 314).

2.2 Local detection of a discontinuity

We first define a local test to decide if there is an abrupt change at $\mathbf{x}$ by testing the null hypothesis $H_0(\mathbf{x})$ "$\mathbb{E}[Z(\mathbf{y})] = m$ for all $\mathbf{y}$ in a neighborhood around $\mathbf{x}$" versus $H_1(\mathbf{x})$ "$\mathbf{x} \in \Gamma$", where Γ is the discontinuity curves of the piecewise constant function $\mathbb{E}[Z(\cdot)]$. Let us define $\mathbf{\Sigma}(\mathbf{x})$, the covariance matrix of $W(\mathbf{x})$:

$$\mathbf{\Sigma}(\mathbf{x}) = \mathbb{E}[W(\mathbf{x})W'(\mathbf{x})] = \partial C'(\mathbf{x})\mathbf{K}^{-1}\partial C(\mathbf{x}).$$

Under a Gaussian assumption for $Z(\cdot)$ and according to standard results in statistics,

$$T(\mathbf{x}) = W'(\mathbf{x})\mathbf{\Sigma}(\mathbf{x})^{-1}W(\mathbf{x}) \tag{3}$$

has a marginal $\chi^2(2)$ distribution.

The null hypothesis is rejected if $T(\mathbf{x}) \geq t_{1-\alpha}$, where $t_{1-\alpha}$ is the $(1-\alpha)$-quantile of the $\chi^2(2)$ distribution. For a confidence level $1-\alpha$ to be precised in Section 2.4, we define the potential Zones of Abrupt Change as the excursion set of $T(\mathbf{x})$ above $t_{1-\alpha}$:

$$\text{Potential ZACs} = \{\, \mathbf{x} \in \mathcal{D} \,:\, T(\mathbf{x}) \geq t_{1-\alpha} \,\}.$$

If the field is stationary, we expect the potential ZACs to be randomly located or non existent. On the contrary, if there is a discontinuity, potential ZACs are likely to be structured along the discontinuity.

2.3 Global test for potential ZACs

The aim of the global test is to discriminate between the two hypothesis H_0: "$Z(\cdot)$ is second order stationary" and H_1: "$\mathbb{E}[Z(\cdot)]$ is piecewise constant, with discontinuities on Γ". At this stage, we have multiple pointwise tests, one for each point $\mathbf{x}$ in $\mathcal{D}$, that need to be aggregated in order to test globally the existence of Γ. Usual multiple tests techniques (Dudoit, Shaffer and Boldrick, 2003) cannot be used because these local tests are highly dependent. Instead, we propose to aggregate the local tests using geometrical properties of the connected components of the excursion set defining the potential ZACs, see Adler (2000) for a complete review on excursion sets of Gaussian related random fields. Allard, Gabriel and Bacro (2005) have shown that the size $S_{1-\alpha}$ of a connected component $\mathcal{C}_{1-\alpha}$ of a potential ZAC is exponentially distributed:

$$t_{1-\alpha} S_{1-\alpha} \xrightarrow{\mathcal{L}} \pi \det(\mathbf{\Lambda})^{-1/2} E(2), \text{ as } t_{1-\alpha} \to \infty, \tag{4}$$

where $\mathbf{\Lambda}$ is the 2×2 matrix of the curvature of $T(\mathbf{x})$ at the maximum in $\mathcal{C}_{1-\alpha}$, which only depends on the covariance function $C_Z(\mathbf{h})$ and the sampling pattern, and $E(2)$ is an exponential random variable with expectation 2 independent on Z. A more detailed version of this theorem, along with a complete definition of $\mathbf{\Lambda}$ is given in Appendix. From Equation (4), a p-value associated to the test "$\Gamma \cap \mathcal{C}_{1-\alpha} = \emptyset$" versus "$\Gamma \cap \mathcal{C}_{1-\alpha} \neq \emptyset$" can be computed:

$$p = \exp\left(-\frac{1}{2\pi} t_{1-\alpha} S_{1-\alpha} \det(\mathbf{\Lambda})^{1/2} \right). \tag{5}$$

The significance of $\mathcal{C}_{1-\alpha}$ is defined by comparing this p-value to a confidence level η, for example $\eta = 5\%$. If p is above η, $\mathcal{C}_{1-\alpha}$ is considered as not significant and is discarded. On the

contrary, if p is below η, $\mathcal{C}_{1-\alpha}$ is considered as significant and defines a ZAC. Since the probability that there is more than one significant connected component tends to 0 as $t_{1-\alpha} \to \infty$, the global test is thus the following: if all p-values are above η, H_0 is not rejected; otherwise it is rejected.

2.4 Determination of α and covariance estimation

In practice, the method is run on a grid superimposed on the domain. On each grid node $\mathbf{x}_p$, the gradient $W(\mathbf{x}_p)$, the matrix $\mathbf{\Sigma}(\mathbf{x}_p)$ and the field $T(\mathbf{x}_p)$ are computed. Then for a confidence level α to be precised below, the set of grid nodes whose statistic $T(\mathbf{x}_p)$ is above the $(1-\alpha)$-quantile of a $\chi^2(2)$ distribution define the potential ZACs. For each potential ZAC, the p-value is then computed according to Equation (5).

The method needs two levels of significance: a global level η, chosen by the user and a local level, α, related to η, for detecting potential ZACs. Equation (5) is valid on the plane, when $t_{1-\alpha} \to \infty$, *i.e.* when $\alpha \to 0$. But as $\alpha \to 0$, the potential ZACs get potentially smaller than the mesh of the interpolation grid. Hence, the level α must be a trade off between the mathematical requirement and the ability to detect the potential ZACs. The correct level α thus depends on η and the covariance function through Equation (5), but also depends on the discretization of the interpolation grid and the sampling pattern. It is assessed by Monte-Carlo simulations: a set of K standard gaussian random fields, corresponding to the null hypothesis of absence of ZACs, are simulated on the same sample locations with the estimated covariance function (see below for the estimation of the covariance function). The estimated local level $1-\hat{\alpha}$ is such that ZACs are detected on $K\eta$ simulations. In Gabriel (2004), it is shown that when the discretization of the interpolation grid is sufficiently fine, a level α_G linked to the integral range of $C_Z(\mathbf{h})$ is a good approximation of $\hat{\alpha}$ and needs less computations. The integral range (Lantuéjoul, 1991) allows to determine the number of independent local tests that we can consider on $\mathcal{D}$. The integral range A of the covariance function $C_Z(\mathbf{h})$ is the quantity:

$$A = \int_{\mathbb{R}^2} \frac{C_Z(\mathbf{h})}{C_Z(\mathbf{0})} \, \mathrm{d}\mathbf{h}. \tag{6}$$

If $A \neq 0$, there exists a number N such that $|\mathcal{D}|/A \approx N$, where $|\mathcal{D}|$ is the domain area. Hence, for N independent local tests, we have:

$$\eta \approx \mathbb{P}\big[\cup_{i=1}^{N} \big\{T(\mathbf{x}_i) \geq t_{1-\alpha}\big\}\big] = 1 - \Big(\mathbb{P}\big[T(\mathbf{x}_i) \leq t_{1-\alpha}\big]\Big)^N,$$

leading to the following definition for α_G:

$$\alpha_G = 1 - (1 - \eta)^{1/N} . \tag{7}$$

In practice the covariance function is unknown and needs to be estimated. The robustness to the covariance estimation was analyzed by simulations in Gabriel, Allard and Bacro (2004). It was shown that the detection of ZACs is not too sensitive to a misspecification of the covariance function.

The method is run under the null hypothesis of absence of ZACs. In presence of a discontinuity the variance σ^2 of $Z(\mathbf{x})$ is thus overestimated. As $T(\mathbf{x})$ is proportional to σ^{-2}, this will result in lower values of the field $T(\mathbf{x})$, hence to smaller ZACs, higher p-values and ultimately to a loss of power of the method. To solve this difficulty we propose the following iterative procedure. The covariance function is estimated using the variogram function, which is widely used in the geostatistical literature [Cressie (1993), Chilès and Delfiner (1999)]. For a stationary random field $Z(\mathbf{x})$, the variogram $\gamma(\mathbf{h}) = \mathbb{E}[\{Z(\mathbf{x}) - Z(\mathbf{x}+\mathbf{h})\}^2]/2$ is related to the covariance function: $\gamma(\mathbf{h}) = C_Z(\mathbf{0}) - C_Z(\mathbf{h})$. We first estimate a global variogram (*i.e.* using all pairs of samples) and a first set of ZACs. The estimation of the variogram is done by a (possibly weighted) least square fit of the experimental variogram $\hat{\gamma}(\mathbf{h}) = 1/(2n_{\mathbf{h}}) \sum_{\mathbf{x}-\tilde{\mathbf{x}}\sim\mathbf{h}} \{Z(\mathbf{x}) - Z(\tilde{\mathbf{x}})\}^2$, where $n_{\mathbf{h}}$ is the number of pairs involved in the sum. We refer to the literature for a more detailed presentation of the variogram and the practice of its estimation. In presence of ZACs, the variogram is re-estimated by discarding all pairs $\{Z(\mathbf{x}_k), Z(\mathbf{x}_l)\}$ for which the segment $[\mathbf{x}_k, \mathbf{x}_l]$ intersects a ZAC. If the range of the variogram changes, the level α must be re-estimated. ZACs are then re-estimated using the new parameters and the procedure is iterated until convergence occurs. Convergence is reached when the set of ZACs remains unchanged. In all tested situations, this happens in less than 5 iterations. It was shown on simulations (Gabriel, 2004) that this procedure greatly improves the estimation of the parameters of the variogram in presence of ZACs.

3 Power of the local test

3.1 Specifying the alternative

The power of the local test is the probability to reject the null hypothesis of stationarity under the alternative of existence of a discontinuity. In the above described method, the shape of the Zones of Abrupt Change is free. To assess the power of the test the shape of the discontinuity must be specified. We will consider a discontinuity as a perturbation function $f(\mathbf{x})$ of the otherwise constant expectation field. We will require that $f(\mathbf{x})$ must be smooth except along the discontinuity, that the discontinuity has a finite length, that its integral $\int_{\mathbb{R}^2} f(\mathbf{x})\,\mathrm{d}\mathbf{x}$ is equal to zero and that it tends to 0 away from the discontinuity. Under this hypothesis, we can approach a vast class of discontinuity curves by its tangent at the point where the power is assessed (Figure 1.*a*). A good candidate for such a perturbation function is the signed Gaussian density function, nicknamed "bottonhole" (Figure 1.*b*). Let us consider that the perturbation function is anchored at the point $\mathbf{x}_0$ in $\mathcal{D}$, that θ is the angle of the discontinuity with the horizontal axis and that L is the "length" of the discontinuity. Following standard practice, we will set $L = 4\psi$, where ψ^2 is the variance of the Gaussian density. With these notations,

$$f(\mathbf{x}) = a\sqrt{\frac{\pi}{2}}\frac{L}{4}\mathrm{sign}\big(\sin(\phi(\mathbf{x}) - \theta)\big)g\big(4||\mathbf{x} - \mathbf{x}_0||/L\big) \tag{8}$$

where g is the $\mathcal{N}(0,1)$ Gaussian density, $\phi(\mathbf{x})$ is the angle of the line $(\mathbf{x}_0, \mathbf{x})$ with the horizontal axis and sign is the sign function. With this parametrization, the discontinuity jump at $\mathbf{x}_0$ is equal to a. The local alternative hypothesis $H_1(\mathbf{x}_0)$ is thus defined as the presence of a function $f(\cdot; \mathbf{x}_0, a, \theta, L)$.

FIGURE 1, PLEASE HERE

3.2 Assessing the local power

Let us denote $1 - \beta(\mathbf{x}_0)$, the power of the local test, *i.e.* the probability to reject at least one of the local tests $H_0(\mathbf{x})$ when $H_1(\mathbf{x}_0)$ is true, where $H_1(\mathbf{x}_0)$ is the alternative hypothesis specified in Section 3.1. Because the orientation of the discontinuity is unknown, we integrate out the power assuming an uniform orientation:

$$1 - \beta(\mathbf{x}_0) = \frac{1}{\pi}\int_0^\pi \{1 - \beta(\mathbf{x}_0; \theta)\}\,\mathrm{d}\theta \tag{9}$$

where $1-\beta(\mathbf{x}_0;\theta)$ is the power of the alternative hypothesis $H_1(\mathbf{x}_0;\theta)$ presented in Section 3.1, with fixed orientation θ (Figure 1.a). Gabriel (2004) showed by simulations that the integral in Equation (9) is well approximated by the sum:

$$1-\frac{1}{4}\sum_{i=0}^{3}\beta(\mathbf{x}_0;i\pi/4).$$

Under the alternative $H_1(\mathbf{x}_0;\theta)$, the estimated gradient at $\mathbf{x}$ can be written as:

$$W(\mathbf{x};\theta)=W_{H_0}(\mathbf{x})+k_a(\mathbf{x};\theta)=\partial C'(\mathbf{x})\mathbf{C}^{-1}\mathbf{Z}+\partial C'(\mathbf{x})\mathbf{C}^{-1}A(\mathbf{x}_0;\theta), \tag{10}$$

where $A(\mathbf{x}_0;\theta)$ is a n-vector, the elements of which being equal to $f(\mathbf{x}_i;\mathbf{x}_0,a,\theta,L)$, $i=1,\ldots,n$.

Denoting N the number of pixels of the interpolation grid, the power $1-\beta(\mathbf{x}_0;\theta)$ becomes:

$$\begin{aligned}1-\beta(\mathbf{x}_0;\theta) &= \mathbb{P}_{H_1(\mathbf{x}_0;\theta)}\left[\bigcup_{p=1}^{N}\Big\{\text{Rejection of the local null hypothesis } H_0(\mathbf{x}_p)\Big\}\right]\\ &= 1-\mathbb{P}_{H_1(\mathbf{x}_0;\theta)}\left[T(\mathbf{x}_1;\theta)\leq t_{1-\alpha},\ldots,T(\mathbf{x}_N;\theta)\leq t_{1-\alpha}\right],\end{aligned}$$

where $T(\mathbf{x}_p;\theta)=W'(\mathbf{x}_p;\theta)\mathbf{\Sigma}^{-1}(\mathbf{x}_p)W(\mathbf{x}_p;\theta)$ and $W(\mathbf{x}_p;\theta)$ is given by Equation (10). In theory, this power must consider all N local tests in $\mathcal{D}$. But nearby test statistics $T(\mathbf{x}_p;\theta)$ are highly dependent, and $1-\beta(\mathbf{x}_0;\theta)$ cannot be computed as $1-\prod_p \mathbb{P}_{H_1(\mathbf{x}_0;\theta)}\left[T(\mathbf{x}_p;\theta)\leq t_{1-\alpha}\right]$.

In order to evaluate $1-\beta\big(\mathbf{x}_0;\theta\big)$, we will restrict this computation to a moving window $\mathcal{F}_k\subseteq\mathcal{D}$, which will conveniently be chosen as a square containing $N_k=(2k+1)\times(2k+1)$ pixels centered at $\mathbf{x}_0$. Let us denote $1-\beta_{\mathcal{F}_k}(\mathbf{x}_0;\theta)$ the power restricted to the window $\mathcal{F}_k$. Since $\mathcal{F}_k\subseteq\mathcal{F}_{k'}$ when $k\leq k'$, the power $1-\beta_{\mathcal{F}_k}(\mathbf{x}_0;\theta)$ provides a minoration of the true power $1-\beta(\mathbf{x}_0;\theta)$. As $\mathcal{F}_k\to\mathcal{D}$, we have:

$$1-\beta_{\mathcal{F}_k}(\mathbf{x}_0;\theta)\longrightarrow 1-\beta(\mathbf{x}_0;\theta).$$

To compute $1-\beta_{\mathcal{F}_k}(\mathbf{x}_0;\theta)$, we use the decomposition of $T(\mathbf{x};\theta)$ as the sum of two non independent squared gaussian random variables with unit variance, $U_1(\mathbf{x};\theta)$ and $U_2(\mathbf{x};\theta)$ (see Appendix), centered under H_0 and with expectation $\mu_i(\mathbf{x};a,\theta), i=1,2$ in presence of a discontinuity. We use the parametric form of the circle $U_1^2(\mathbf{x};\theta)+U_2^2(\mathbf{x};\theta)=t_{\mathbf{x}}$:

$$U_1(\mathbf{x};\theta)=\sqrt{t_{\mathbf{x}}}\cos(\omega_{\mathbf{x}}), U_2(\mathbf{x};\theta)=\sqrt{t_{\mathbf{x}}}\sin(\omega_{\mathbf{x}}), \tag{11}$$

to assess the power calculated in the center $\mathbf{x}_0$ of $\mathcal{F}_k$ with a fixed direction θ of the discontinuity:

$$\begin{aligned} 1-\beta_{\mathcal{F}_k}(\mathbf{x}_0;\theta) &= 1-\mathbb{P}_{H_1(\mathbf{x}_0;\theta)}\left[T(\mathbf{x}_1;\theta)\leq t_{1-\alpha},\ldots,T(\mathbf{x}_{N_k};\theta)\leq t_{1-\alpha}\right] \\ &= 1-\int_0^{t_{1-\alpha}}\ldots\int_0^{t_{1-\alpha}}\mathbb{P}_{H_1(\mathbf{x}_0;\theta)}\left[T(\mathbf{x}_1;\theta)=t_1,\ldots,T(\mathbf{x}_{N_k};\theta)=t_{N_k}\right]\,\mathrm{d}t_1\ldots\,\mathrm{d}t_{N_k} \\ &= 1-\frac{1}{2^{N_k}}\int_{\mathbf{w}\in[0,2\pi]^{N_k}}\int_{\mathbf{t}\in[0,t_{1-\alpha}]^{N_k}} f_{\mathbf{V}_\theta}(\mathbf{v})\,\mathrm{d}\mathbf{t}\,\mathrm{d}\mathbf{w}, \end{aligned} \tag{12}$$

where

- $\mathbf{V}_\theta$ is a $2N_k$-gaussian vector which elements are $\Big(U_1(\mathbf{x}_1;\theta),U_2(\mathbf{x}_1;\theta),\ldots,U_1(\mathbf{x}_{N_k};\theta),U_2(\mathbf{x}_{N_k};\theta)\Big)'$, with mean $\mathbf{m}_{\mathbf{V}_\theta}=\big(\mu_1(\mathbf{x}_1;a,\theta),\mu_2(\mathbf{x}_1;a,\theta),\ldots,\mu_1(\mathbf{x}_{N_k};a,\theta),\mu_2(\mathbf{x}_{N_k};a,\theta)\big)'$ and covariance matrix:

$$\mathbf{\Sigma}_{\mathbf{V}_\theta} = \begin{pmatrix} 1 & 0 & \ldots & c_{11}(\mathbf{x}_1,\mathbf{x}_{N_k}) & c_{12}(\mathbf{x}_1,\mathbf{x}_{N_k}) \\ 0 & 1 & \ldots & c_{21}(\mathbf{x}_1,\mathbf{x}_{N_k}) & c_{22}(\mathbf{x}_1,\mathbf{x}_{N_k}) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ c_{11}(\mathbf{x}_1,\mathbf{x}_{N_k}) & c_{12}(\mathbf{x}_1,\mathbf{x}_{N_k}) & \ldots & 1 & 0 \\ c_{12}(\mathbf{x}_1,\mathbf{x}_{N_k}) & c_{22}(\mathbf{x}_1,\mathbf{x}_{N_k}) & \ldots & 0 & 1 \end{pmatrix},$$

  where $c_{ij}(\mathbf{x}_k,\mathbf{x}_l)=\mathrm{Cov}\left(U_i(\mathbf{x}_k;\theta),U_j(\mathbf{x}_l;\theta)\right)$, $i,j\in\{1,2\}$, $k,j\in\{1,\ldots,N_k\}$ does not depend on θ,

- $\mathbf{w}=(\omega_1,\ldots,\omega_{N_k})'$ and $\mathbf{t}=(t_1,\ldots,t_{N_k})'$ come from Equation (11),

- $\mathbf{v}=\big(\sqrt{t_1}\cos\omega_1,\sqrt{t_1}\sin\omega_1,\ldots,\sqrt{t_{N_k}}\cos\omega_{N_k},\sqrt{t_{N_k}}\sin\omega_{N_k}\big)'$.

The multiple integral given in Equation (12) is evaluated by Monte-Carlo approximation. A series of n_s independent vectors $\mathbf{V}_\theta$ is generated and

$$1-\beta_{\mathcal{F}_k}(\mathbf{x}_0;\theta)\approx 1-\frac{1}{n_s}\sum_{i=1}^{n_s}\mathbf{1}_{\left\{T(\mathbf{x}_1;\theta)\leq t_{1-\alpha}\right\}}\ldots\mathbf{1}_{\left\{T(\mathbf{x}_{N_k};\theta)\leq t_{1-\alpha}\right\}}.$$

The choice of k is important. On the one hand, k should be as large as possible to approach $1-\beta(\mathbf{x}_0;\theta)$. On the other hand, $\mathbf{\Sigma}_{\mathbf{V}_\theta}$ is a $2(2k+1)^2\times 2(2k+1)^2$ matrix and it becomes quickly intractable to calculate $1-\beta_{\mathcal{F}_k}(\mathbf{x}_0;\theta)$ when the size of $\mathcal{F}_k$ increases (Table 1). However, because of the particular alternative hypothesis model chosen, pixels $\mathbf{x}$ whose distance to $\mathbf{x}_0$ is larger than $L/2$ do not need to be considered. Indeed for these pixels, the elements of $A(\mathbf{x}_0;\theta)$ in the

neighbourhood of $\mathbf{x}$ are small and the probability of the rejection of $H_0(\mathbf{x})$ given that $H_1(\mathbf{x}_0;\theta)$ is true becomes negligible. Empirical evidences on simulations not reported here show that $k = 4$ is a good trade-off between accuracy and computation time.

TABLE 1, PLEASE HERE

Because we consider a subset $\mathcal{F}_k$ of $\mathcal{D}$, we cannot use the above mentioned local levels $\hat{\alpha}$ and α_G. Indeed, $\hat{\alpha}$ depends on the local density on the whole study domain and is mainly influenced by the locations of $\mathcal{D}$ on which the sample points are the most aggregated. Consequently, for any fixed $\mathcal{F}_k$, $\hat{\alpha}$ tends to underestimate the right value of α on this subset. As for α_G, it is an approximation of $\hat{\alpha}$ only if the discretization of the interpolation grid is fine enough, which is not the case any more when considering $\mathcal{F}_k$. To solve this problem, we consider a local level $\alpha_k(\mathbf{x})$ that is estimated in the same way than the global level α, except that $\hat{\alpha}_k(\mathbf{x})$ is such that ZACs are detected on $K\eta$ simulations in the subset $\mathcal{F}_k$.

4 Application to soil data

4.1 Data set

The data considered here were sampled in an agricultural field (10 ha) in Chambry, Northern France, as part of a precision agriculture project (Guérif et al., 2001). Precision agriculture aims at defining a location dependent management within an agricultural field in order to better manage production in quantity and quality and minimize nitrogen losses towards ground water. This implies to consider the soil characteristics and their spatial variability. Detection of ZACs can be an helpful tool for defining boundaries between homogeneous management zones.

Samples were collected on a pseudo-regular grid with distance between nodes equals to 36 m, at 8 different dates from March 2000 to February 2003, see Figure 2. The analysis is performed on soil water content, one of the most important parameters in plant growth models used for managing crops. It was measured on soil cores up to 120 cm. Each soil sample is itself a mixture of three cores taken at 50 cm distance. Similar data, collected in a nearby field are presented and analyzed in Mary et al. (2001).

Figure 2 shows the sample locations and the interpolation map of the soil water content on a 64×98 grid using ordinary kriging. From these interpolated maps, it can be seen that the

Southern part of the field is moister than the Northern part. A small zone with a very low soil water content appears in the Eastern part of the field for all dates and on the Western part for several dates.

FIGURE 2, PLEASE HERE

Figure 3 depicts the boxplot of the soil water content for each date. Figure 2 and Figure 3 illustrate the temporal variability of the soil water content. October 2000, August 2001 and July 2002 are during the dryer season and around periods of harvest; consequently the soil water content is low. Conversely at the end of the winter season, *i.e.* in March 2000 and 2002, and in February 2001 and 2003, soil water content is higher because the water use is low compared to the soil water filling due to the season.

FIGURE 3, PLEASE HERE

4.2 Detection of ZACs

For all dates and at each step of the iterative procedure, exponential covariance functions were estimated, $C_Z(\mathbf{h}) = \hat{\sigma}^2 \exp(-\|\mathbf{h}\|/\hat{b})$, where $\hat{\sigma}^2$ and $\hat{b}$ are the estimated sill and range parameters. A 64×98 grid with a mesh size of 5 m was used. For an exponential covariance function defined on $\mathbb{R}^2$, Equation (6) gives the following integral range: $A = 2\pi b^2$. The local levels α were set from Equation (7): $\alpha_G = 1 - (1-\eta)^{2\pi\hat{b}^2/|\mathcal{D}|}$, with η fixed to 5%. Table 2 gives the range and sill parameters of the exponential covariance functions estimated at the first iteration and at the convergence of the iterative procedure. It shows that convergence occurs in a few iterations (from one to four). At the convergence the estimated variance is usually lower than the one estimated at the first iteration, illustrating the variance estimation error introduced by the existence of discontinuities.

TABLE 2, PLEASE HERE

Figure 4 depicts the detected ZACs of the soil water content: in the Eastern part of the field a ZAC is detected for all dates, except for October 2000; in the Western part a potential ZAC is found for October 2000 and is deemed significant for August 2001. These ZACs can be related to a map of the main soil types (not shown here) drawn independently by soil scientists using

different data. The larger ZAC in the Eastern part of the field delineates the sandy summit zone, while the smaller one in the Western part is related to the transition between chalky and non-chalky soil. In the middle, no ZAC was detected because this area corresponds to a smoother transition. The analysis on a non permanent variable (soil water content) shows permanent structures and indicates that the field could be divided into two management zones corresponding to the top and the bottom of the field. A third zone could be defined, corresponding to the summit zone in the Eastern part of the field. But considering its small size, it would probably not be interesting under a management viewpoint. For October 2000, data were missing in the South-Eastern part of the field. For this reason, the fact that no ZAC was found could be due either to this lack of data or to the true absence of an abrupt change. Assessing the power of the detection method for October 2000 is therefore necessary to distinguish between these two alternatives.

FIGURE 4, PLEASE HERE

4.3 Influence of the sampling pattern on the local power

In this section, we study the influence of the sampling pattern on the detection of ZACs, using the local power presented in Section 3.

Let us first consider the case of a homogeneous sampling pattern, focusing for example on October 2002. There are 83 points forming a regular pattern in the agricultural field (Figure 2); a ZAC is detected in the Eastern part of the field (Figure 4). From this original data set, 10 sample points were randomly discarded, thus providing a restricted sample of 73 points. This operation is then interated three times, to get sample patterns of 63, 53 and 43 points.

The first line of Figure 5 displays the obtained samples. For each point pattern, the covariance function and the ZACs were estimated as presented in Section 4.2. The second line of Figure 5 shows the detected significant potential ZACs in black and the non significant ones in grey. One can clearly see how the detection of ZACs deteriorates as the number of samples decreases. The ZAC in the Eastern part of the field is still partly estimated when 36% of the points are discarded from the original sample, but no ZAC is detected when half of the original sample points are discarded.

For all cases, the local power was computed at each pixel of the interpolation grid using a

9×9 pixels moving window $\mathcal{F}_4$. Because the computation of $\hat{\alpha}_4(\mathbf{x})$ on the whole grid is very long and because the samples have a relatively homogeneous pattern (at least when $n \geq 53$), a uniform level $\bar{\alpha}_4$ was applied, where $\bar{\alpha}_4$ is the average of the level estimated at 10 randomly picked pixels. We got $\bar{\alpha}_4 = 0.00535, 0.00563, 0.00627, 0.00708, 0.00873$ for $n = 83, 73, 63, 53, 43$, respectively. Note that the level $\bar{\alpha}_4$ increases as n decreases, illustrating the loss of power of the method as the number of sample points decreases.

We considered a discontinuity $a = 3\sigma$. From an agronomical viewpoint, the minimum size of a ZAC should be around 70 m to be of interest, *i.e.* twice the mesh size of the original sampling design. For a discontinuity length $L = 140$ m, the section potentially detectable, *i.e.* where the discontinuity jump is larger than 2σ (Gabriel, Allard and Bacro, 2004), has only a 70 m length.

The third line of Figure 5 displays the local power (center of the moving window $\mathcal{F}_4$) of the n-samples. This map shows, for each pixel $\mathbf{x}_p$, an approximation of the probability to detect a "bottonhole" discontinuity anchored in $\mathbf{x}_p$, using all relevant information available in the study domain. The white pixels near the boundary of the agricultural field correspond to pixels where the local power could not be assessed, due to missing values in $\mathcal{F}_4$. These figures illustrate that the local power is strongly linked to the local sampling density; low values of the local power correspond to the zones where sample points were discarded from the regular sampling pattern. For $n \geq 53$, the detected ZACs are always located in areas with a high local power. When $n = 43$, lack of sample points in the vicinity of the ZAC makes it impossible to detect it. In this area, the local power is very low (0.31 on average) and therefore no ZAC can be detected. This results shows that 43 data points is too low a sampling density for detecting ZACs in this 10 ha field.

FIGURE 5, PLEASE HERE

In the light of this exercice, we analyse the results obtained for October 2000, *i.e.* the case of a heterogenous sampling pattern. At this date, there is no sample point in the South-Eastern part of the field and only a small non significant potential ZAC is detected in the Western part of the field (Figure 5).

As discussed in Section 3.3, the local level $\hat{\alpha}_4(\mathbf{x})$ that must be considered to assess the local power depends on the local sample density around $\mathbf{x}$. Consequently, for the sampling pattern for October 2000, the value of $\hat{\alpha}_4(\mathbf{x})$ could not be considered constant on the whole field as in

the previous section. We made a trade off between the computing time and the fact that $\hat{\alpha}_4(\mathbf{x})$ cannot be constant on such a sampling pattern. Hence, the field is divided in three parts: North, South-West and South-East, on which the sample can be considered as homogeneous (Figure 5). The different local levels were evaluated by the mean of $\hat{\alpha}_4(\mathbf{x})$ estimated at 6 randomly located pixels in each part: $\bar{\alpha}_4 = 0.00493, 0.00689, 0.01086$ within respectively the Northern, South-Western and South-Eastern part. Again, we note that $\bar{\alpha}_4$ increases when the density of sample points decreases. The local power, depicted (Figure 5), was assessed letting $a = 3\sigma$ and $L = 140\ m$. This figure again illustrates that a lack of sample points leads to a low local power. This analysis demonstrates that the fact that no ZAC was detected for October 2000 is more probably due to a lack of sample points than due to the genuine absence of a ZAC. The fact that ZACs were detected for all other dates provides further evidence for this conclusion.

5 Discussion

In this paper we have presented a method for detecting Zones of Abrupt Change for spatially correlated data. This method tests locally the existence of abrupt changes and globally the significance of the detected ZACs. Unlike most other methods, it is able to take into account the correlation function of the variable of interest when testing for the significance of the detected ZACs.

No spatial regularity of the sampling pattern is required, but the local sample density has some consequences. In densely sampled zones the estimated gradient may be artificially large and may lead to statistically significant ZACs that are not physically meaningful and usually small. This is also the case for outlierse and our method could thus be used to detect them. Conversely, in an unsampled zone, the estimated gradient is very low and almost constant, thus leading to the impossibility of estimating a ZAC. In view to specify if the absence of ZAC is due to a lack of sample points or to a stationary variable, we assessed the local power of the detection test.

The power at a point $\mathbf{x}$ of the study domain is the probability to reject at least one local test, under the existence of a discontinuity anchored at $\mathbf{x}$. Assessing the power needs to specify an alternative hypothesis, that is, to provide a model of discontinuity. We used a signed Gaussian model to approach a wide class of discontinuities. Other choices are possible but this model

has the advantage of being easy to use and well behaved in terms of regularity and smoothness away from the discontinuity. It has three parameters: its length L, discontinuity jump a and angle with the horizontal axis θ. In the application presented here, we chose to integrate θ with a uniform density to account for all possible orientations. One can alternatively choose to map the mimimum or the maximum of the power at each location. Several power maps can be drawn for different values of L and a, set by the user. For a fixed location $\mathbf{x}$, the power will increase when either L or a increase.

Mapping the power allows us to visualize areas where ZACs cannot be detected because of an unappropriate local sample density. It can be used to evaluate locally the sampling pattern for detecting ZACs and to highlight the areas where additional samples are necessary.

Several hypothesis where made when presenting our method. Second order stationarity was assumed throughout, but intrinsic stationarity is sufficient. In this case, the covariance matrix $\mathbf{C}$ must be replaced by a matrix $\mathbf{\Gamma}$ which elements are $\gamma(x_i - x_j)$ and similarly for the covariance vector $C(\mathbf{x})$. Third order differentiability of the covariance function away from zero is necessary for the matrix $\mathbf{\Lambda}$ to exist. The spherical covariance function does not verify this assumption for $||h|| = b$, where b is its range. However, our method has been succesfully applied with spherical functions on a great number of simulations. Practically speaking, this differentiability assumption is probably not necessary. The random field $Z(\cdot)$ is assumed to be Gaussian. If the data cannot be considered as Gaussian, they can be transformed on a Gaussian scale, using a parametric form (*e.g.* log transform or Cox transform) or a non parametric quantile transform. This of course does not ensure the gaussianity of the random field but should be enough for our method to work well in practice, as illustrated on the soil water content application shown in the paper.

Several problems remain still open at this stage, among which is the generalisation to multivariate data and taking into account temporal information. Generalising our method to multivariate data requires to prove a theorem similar to Theorem 1 in a multivariate framework and also to have a fully automatic procedure for estimating the variogram.

Appendix

From Equation (3), we can write the statistic $T(\mathbf{x})$ of the local test of detection of ZACs as the sum of the square of non-stationary, non-independent gaussian random fields, $U_1(\mathbf{x})$ and $U_2(\mathbf{x})$

which are not identically distributed, $U(\mathbf{x}) = \mathbf{\Sigma}^{-1/2}(\mathbf{x})W(\mathbf{x})$, *i.e.* specifically:

$$\begin{aligned} U_1(\mathbf{x}) &= \frac{W_1(\mathbf{x})}{\sigma_1(\mathbf{x})}, \\ U_2(\mathbf{x}) &= \frac{1}{\sqrt{1-\rho^2(\mathbf{x})}}\left\{\frac{W_2(\mathbf{x})}{\sigma_2(\mathbf{x})} - \rho(\mathbf{x})\frac{W_1(\mathbf{x})}{\sigma_1(\mathbf{x})}\right\}, \end{aligned}$$

where $W_i(\mathbf{x}), i = 1,2$ are the coordinates of $W(\mathbf{x})$, $\sigma_i(\mathbf{x}) = \sqrt{\mathbf{\Sigma}_{[ii]}(\mathbf{x})}, i = 1,2$ and $\rho(\mathbf{x}) = \mathbf{\Sigma}_{[12]}(\mathbf{x})/\sigma_1(\mathbf{x})\sigma_2(\mathbf{x})$.

Potential ZACs are defined as the set of points of $\mathcal{D}$ such that $T(\mathbf{x})$ is above the level $t_{1-\alpha} = \chi^2_{1-\alpha}(2)$. They are thus excursion sets of χ^2 random fields [Adler (2000), Aronowich and Adler (1988), Worsley (1994), Cao (1999)]. When $t_{1-\alpha} \to \infty$, (*i.e.* when $\alpha \to 0$), it can be shown (Allard, Gabriel and Bacro, 2005) that the area $S_{1-\alpha}$ of a connected component $\mathcal{C}_{1-\alpha}$ of this excursion set is related to the curvature of $T(\mathbf{x})$ near its local maximum (shifted at $\mathbf{0}$) through the following convergence in law:

Theorem 1 *Conditional on $T(\mathbf{x})$ having a maximum $T_0 > t_{1-\alpha}$ at $\mathbf{0}$,*

$$t_{1-\alpha}S_{1-\alpha} \xrightarrow{\mathcal{L}} \pi\, det PIST E(\mathbf{\Lambda})^{-1/2} E(2), \tag{13}$$

where $\mathbf{\Lambda} = v\mathbf{\Lambda}_1(\mathbf{0}) + (1-v)\mathbf{\Lambda}_2(\mathbf{0})$, with $v = U_1^2(\mathbf{0})/T_0$.

In this theorem, $\mathbf{\Lambda}_i(\mathbf{x})$ is the covariance matrix of the gradient of $U_i(\mathbf{x})$:

$$\mathbf{\Lambda}_i(\mathbf{x}) = \mathbb{E}\big[\partial U_i(\mathbf{x}) \partial U_i'(\mathbf{x})\big] = Var\big(\partial U_i(\mathbf{x})\big).$$

Let us denote $\sigma_i = \sigma_i(\mathbf{x}), i = 1,2$ and $\rho = \rho(\mathbf{x})$. For the sake of lighter notations, let us further denote $D_i = D_i(\mathbf{x}) = \partial_i C(\mathbf{x})$ for $i = 1,2$ where $\partial_i f = \partial f/\partial x^i$. Then, for $k,l = 1,2$:

$$\begin{aligned} \mathbf{\Lambda}_{1\ [kl]}(\mathbf{x}) &= \{\partial_k D_1' \mathbf{K}^{-1} \partial_l D_1 - \partial_k \sigma_1 \partial_l \sigma_1\}/\sigma_1^2, \\ \mathbf{\Lambda}_{2\ [kl]}(\mathbf{x}) &= \{\partial_k D_2' \mathbf{K}^{-1} \partial_l D_2/\sigma_2^2 - \partial_k \sigma_2 \partial_l \sigma_2/\sigma_2^2 + \partial_k \rho \partial_l \rho - \partial_k \rho \partial_l D_2' \mathbf{K}^{-1} D_1/\sigma_1\sigma_2 \\ &\quad -\partial_l \rho \partial_k D_2' \mathbf{K}^{-1} D_1/\sigma_1\sigma_2 + \rho[(\partial_k\sigma_2/\sigma_2)(\partial_l D_1' \mathbf{K}^{-1} D_2/\sigma_1\sigma_2 + \partial_l\rho) \\ &\quad +(\partial_l\sigma_2/\sigma_2)(\partial_k D_1' \mathbf{K}^{-1} D_2/\sigma_1\sigma_2 + \partial_k\rho) + (\partial_l\sigma_1/\sigma_1)(\partial_k D_2' \mathbf{K}^{-1} D_1/\sigma_1\sigma_2) \\ &\quad +(\partial_k\sigma_1/\sigma_1)(\partial_l D_2' \mathbf{K}^{-1} D_1/\sigma_1\sigma_2) - \partial_k D_1' \mathbf{K}^{-1} \partial_l D_2/\sigma_1\sigma_2 \\ &\quad -\partial_k D_2' \mathbf{K}^{-1} \partial_l D_1/\sigma_1\sigma_2] + \rho^2[\partial_k D_1' \mathbf{K}^{-1} \partial_l D_1/\sigma_1^1 - \partial_k\sigma_1\partial_l\sigma_1/\sigma_1^1 \\ &\quad -\partial_k\sigma_1\partial_l\sigma_2/\sigma_1\sigma_2 - \partial_k\sigma_2\partial_l\sigma_1/\sigma_1\sigma_2]\}/(1-\rho^2) - \rho^2(\partial_k\rho\partial_l\rho)/(1-\rho^2)^2. \end{aligned}$$

Details and proof of these results can be found in Allard, Gabriel and Bacro (2005).

Acknowledgments

This work was done while the first author was at the Unité de Biométrie, Institut National de la Recherche Agronomique, France. The authors want to thank B. Mary and M. Guérif, INRA, for fruitful discussions about the data set, and G. Alavoine, F. Barrois, D. Boitez, O. Delfosse, E. Guérhan, C. Herre, F. Mahu, F. Millon and E. Venet, INRA, for providing the data.

References

Adler, R. (2000) On excursion sets, tube formulas and maxima of random fields. *The Annals of Applied Probability*, **10**, 1–74.

Allard, D., Gabriel, E. and Bacro, J-N. (2005) Estimating and testing Zones of Abrupt Change for spatial data. *Journal of the American Statistician Association* (in revision).

Aronowich, M. and Adler, R. (1988) Sample path behaviour of χ^2 surfaces at extrema. *Advances in Applied Probability*, **18**, 901–920.

Banerjee, S., Gelfand, A., and Sirmans, C. (2003) Directional rates of change under spatial process models. *Journal of the American Statistician Association*, **98**, 946–954.

Barbujani, G., Oden, N. and Sokal, R. (1989) Detecting areas of abrupt change in maps of biological variables. *Systematic Zoology*, **38**, 376–389.

Bocquet-Appel, J-P. and Bacro J-N. (1994) Generalized Wombling. *Systematic Biology*, **43** 442–448.

Cao, J. (1999) The size of the connected components of excursion sets of χ^2, t and F fields. *Advances in Applied Probability (SGSA)*, **31**, 579–595.

Chilès, J-P. and Delfiner, P. (1999) *Geostatistics: modeling spatial uncertainty*, Wiley, New-York.

Cressie, N. (1993) *Statistics for spatial data, Revised Edition.* Wiley, New-York.

Dudoit, S., Shaffer, J-P. and Boldrick, J. (2003) Multiple hypothesis testing in microarray experiments. *Statistical Science*, **18**, 71–103.

Fortin, M-J. (1994) Edge detection algorithms for two-dimensional ecological data. *Ecology*, **75**, 956–965.

Fortin, M-J. and Drapeau, P. (1995) Delineation of ecological boundaries: Comparisons of approaches and significance tests. *Oikos*, **72**, 323–332.

Gabriel, E. (2004) Détection de zones de changement abrupt dans des données spatiales et application à l'agriculture de précision. Ph.D. thesis, University of Montpellier.

Gabriel, E., Allard, D. and Bacro, J-N. (2004) Detecting Zones of Abrupt Change: Application to Soil Data. In *Proceedings of the IV European Conference on Geostatistics for Environmental Applications*, X. Sanchez-Vila, J. Carrera and R. Froidevaux (eds), pp. 437–448.

Gleyze, J-F., Bacro, J-N. and Allard, D. 2001. Detecting regions of abrupt change: Wombling procedure and statistical significance. *geoENV III: Geostatistics for environmental applications.*

Godtliebsen, F., Marron, J. and Pizer, S. (2002) Significance in scale-space for clustering. *Spatial Clustering Modeling*, Boca Raton (eds), Chapman and Hall/CRC, pp. 24–36.

Guérif, M., Beaudoin, C., Durr, V., Houlès, V., Machet, J-M., Mary, B., Moulin, S. and Richard, G. 2001. Designing a field experiment for assessing soil and crop spatial variability and defining site-specific management strategies. In *Proceedings of the third European Conference on Precision Agriculture.* (ed. Grenier and Blackmore), pp. 677–682.

Hall, P. and Rau, C. 2001. Local likelihood tracking of fault lines and boundaries. *Journal of the Royal Statistical Society B*, **63**, 569–582.

Jacquez, G. and Maruca, S. (1998) Geographic boundary detection. In *Proceedings of the 8th International Symposium on Spatial Data Handling.* T.K. Poiker and N. Chrisman (eds) International Geographical Union, pp. 496–509 .

Jacquez, G., Maruca, S. and Fortin, M-J. (2000) From fields to objects: a review of geographic boundary analysis. *Journal of Geographical Systems*, **2**, 221–241.

Lantuéjoul, C. (1991) Ergodicity and integral range. *Journal of Microscopy*, **161**, 387–403.

Mary, B., Beaudoin, N., Machet, J-M., Bruchou, C. and Ariès, F. (2001) Characterization and analysis of soil variability within two agricultural fields: the case of water and mineral N profiles. In *Proceedings of the 3rd European Conference on Precision Agriculture* Grenier and Blackmore (eds), pp. 431–436.

Oden, N., Sokal, R., Fortin, M-J. and Goebl, H. (1993) Categorical Wombling: Detecting regions of significant change in spatially located categorical variables. *Geographical Analysis*, **25**, 315–336.

Pagel, M., and Mace, R. (2004) The cultural wealth of nations. *Nature*, **428**, 275–278.

Womble, W. (1951) Differential systematics. *Science*, **114**, 315–322.

Worsley, K. (1994) Local maxima and the expected Euler characteristic of excursion sets of χ^2, F and t fields. *Advances in Applied Probability*, **26**, 13–42.

Worsley, K. (2001) Testing for signals with unknown location and scale in a χ^2 random field, with application to fMRI. *Advances in Applied Probability (SGSA)*, **33**, 773–793.

Biographical sketches

Edith Gabriel obtained her Ph.D. in Statistics of the Université de Montpellier in December 2004. Currently, she is a Research Associate in the Medical Statistics Unit of Lancaster University and is working on statistical methods in spatial epidemiology.

Denis Allard is a senior scientist working at the Unité de Biométrie, Institut National de la Recherche Agronomique, France. His research interests are focused on spatial statistics, with an emphasis on applications related to environmental data sets. He is an elected member of the board of the Environmental group of the French Statistical Society.

Table 1: Dimension of the covariance matrix $\mathbf{\Sigma}_{\mathbf{V}_\theta}$ as a function of the size of $\mathcal{F}_k$.

k	0	2	4	8	12
Dimension of $\mathcal{F}_k$	1×1	5×5	9×9	17×17	25×25
Dimension of $\mathbf{\Sigma}_{\mathbf{V}_\theta}$	2×2	50×50	162×162	578×578	1250×1250

Table 2: Parameters of the exponential covariances estimated at the first iteration and the convergence of the iterative procedure: range $\hat{b}$ (m) and sill $\hat{\sigma}^2$ (mm^2).

	Iteration 1		Convergence		
Date	$\hat{b}$	$\hat{\sigma}^2$	Iteration	$\hat{b}$	$\hat{\sigma}^2$
March 2000	23	2714	4	30	2781
October 2000	26	2509	1	–	–
February 2001	25	2929	4	30	1704
August 2001	27	3129	3	39	2571
March 2002	23	2209	3	32	2068
July 2002	32	2952	1	–	–
October 2002	27	2906	3	37	2765
February 2003	25	3388	2	32	3420

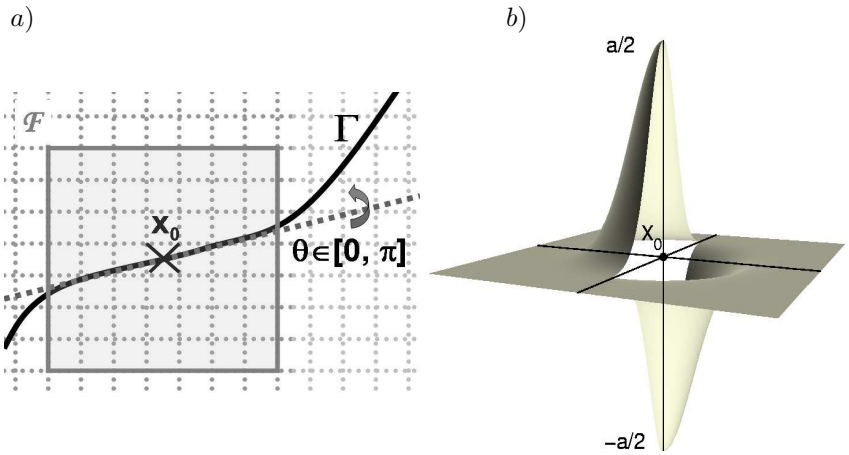

Figure 1: *a*) Approximation of Γ by its tangent at $\mathbf{x}_0$. *b*) Model of alternative hypothesis used to assess the power at $\mathbf{x}_0$.

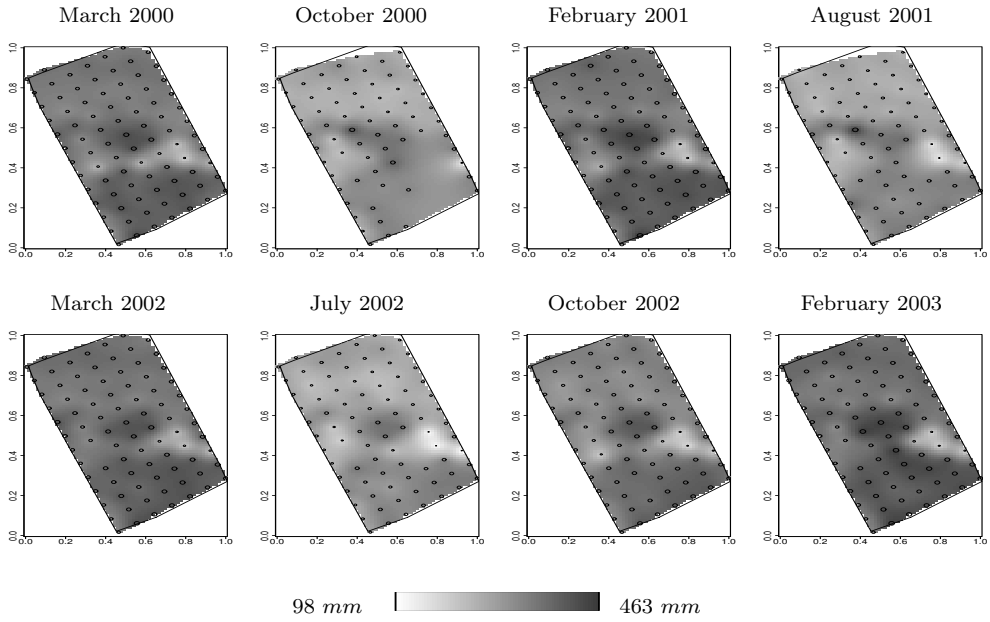

Figure 2: Soil water content: sample points with symbols proportionals to their value and its interpolation map.

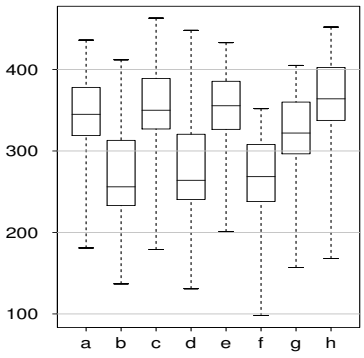

Figure 3: Boxplot of the soil water content. From a to h: March 2000, October 2000, February 2001, August 2001, March 2002, July 2002, October 2002, February 2003.

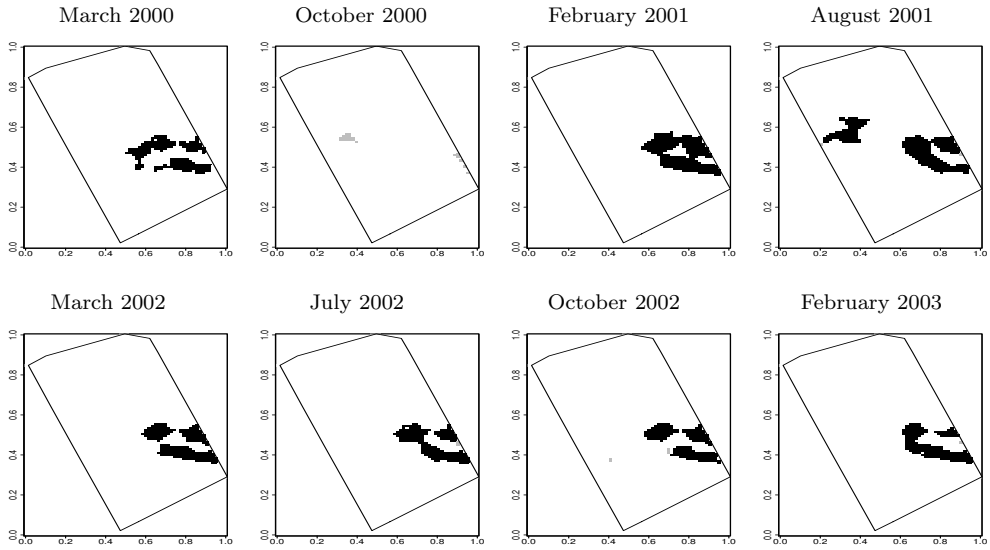

Figure 4: Estimated ZACs (in black) and non significant potential ZACs (in grey) obtained at the convergence of the iterative procedure for the soil water content.

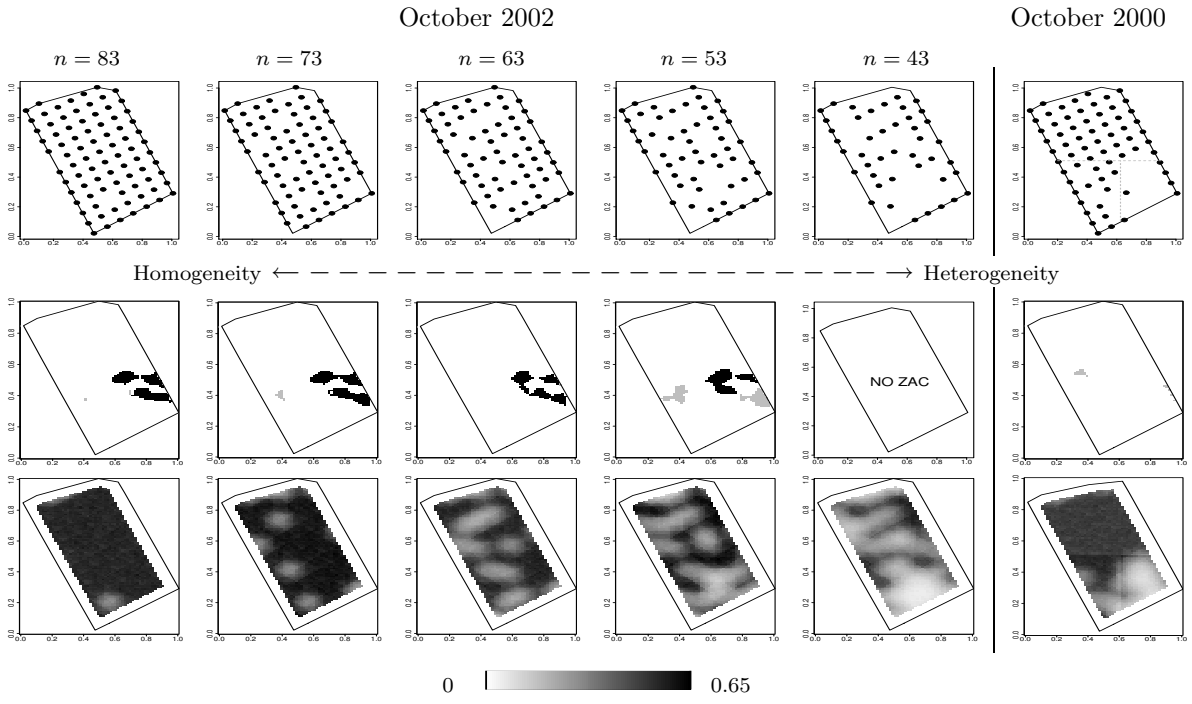

Figure 5: First line: October 2002 sample and sub-samples; October 2000 sample. Second line: estimated ZACs (in black) and non significant potential ZACs (in grey). Third line: Local power calculated on a 9×9-pixels moving window.