Detecting zones of abrupt change in soil data, with an application to an agricultural field

E. Gabriel\textsuperscript{a}, D. Allard\textsuperscript{b}, B. Mary\textsuperscript{c} \& M. Guérif\textsuperscript{d}

\textsuperscript{a}Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, UK, \textsuperscript{b}Unité Biostatistique et Processus Spatiaux, INRA, Agroparc, 84914 Avignon, France, \textsuperscript{c}Unité d’Agronomie, INRA, Rue Fernand Christ, 02007 Laon, France, and \textsuperscript{d}Unité Climat Sol et Environnement, INRA, Agroparc, 84914 Avignon, France

Summary

We propose a new method for estimating and testing the zones where a variable has discontinuities or sharp changes in the mean. Such zones are called Zones of Abrupt Change (ZACs). Our method is based on the statistical properties of the estimated gradient of the variable. The local gradient is first interpolated by kriging. Then we test whether the estimated local gradient exceeds some critical threshold computed under the null hypothesis of a constant mean. The locations where the local test is rejected define the potential ZACs, which are then tested globally. Using this method, we analyse soil data from an agricultural field. The analysis of the main soil components of the ploughed layer (clay, silt and sand particles and calcium carbonate content) reveals the structural variations in the field, linked to boundaries between soil types. Its application to non-permanent variables (soil water and mineral nitrogen content of the [0, 120 cm] soil profile taken at several dates) shows that water content has the same ZACs for all dates, whereas mineral nitrogen has none.

Introduction

\textsuperscript{a}Correspondence: E. Gabriel. E-mail: e.gabriel@lancaster.ac.uk
Many soil, agricultural or environmental problems require knowledge of the spatial variation of the variables under study. Two extreme models of spatial variation are (i) the discrete model used mostly for classification, in which soil variation is explained by a mean effect and uncorrelated noise, and (ii) the continuous model of spatial variation described by a correlated random function, typically addressed by geostatistical techniques. However, reality often lies between these two extremes. Solutions have been proposed to merge the discrete and continuous models, see e.g. Heuvelink & Webster (2001) for a review. Burrough et al. (1997) pointed out that soil differences are sometimes associated with abrupt changes, which they call primary boundaries, and sometimes with smooth transitions, called secondary boundaries.

We propose a unified framework for modelling soil variation that integrates these two concepts. Soil variation is described by the sum of a trend and a correlated random function with zero mean. The trend is generally smooth and slowly varying, but in some places it is discontinuous or varies abruptly; we may recognise these places as primary boundaries. The correlated function may include smooth secondary boundaries. We propose a method for detecting (i.e. estimating and testing) the locations of primary boundaries for a spatially correlated variable. Our method can distinguish between smooth transitions, related to the continuous model, and abrupt transitions, related to the discrete model. We propose to call the latter Zones of Abrupt Change (ZACs). Because the aim of the method is to extract the statistically significant abrupt changes in a variable, ZACs are generally discontinuous and do not constitute boundaries in the everyday sense.

To the best of our knowledge, there is no method for detecting ZACs in soil properties. Webster (1973) described a method for determining boundaries automatically along transects, but only outlined a possible extension of the procedure to two dimensions. More generally, the problem of estimating discontinuity curves in the plane with irregularly located samples has a sparse literature. Womble (1951) illustrated that abrupt changes in allele frequencies can be linked to the boundaries between dif-
Different populations. That method is quite simple: the variable under study is linearly interpolated on a regular grid; one then approximates a gradient vector by computing differences; the top 5% of the gradient values define barriers. With the advent of affordable personal computers, Barbujani et al. (1989) revisited this method, adding the direction of the gradient in the definition of barriers. In Gleyze et al. (2001), a first attempt was made to assess the significance of barriers, but the null hypothesis of no barriers is based on the assumption of uncorrelated spatial variation, which is unrealistic in most environmental applications. Hall et al. (2001) proposed a non-parametric method based on spatial approximations of the local likelihood, assuming that the discontinuity curve passes through a given point in the plane, as a function of this point.

An important limitation of this method is that a starting point is needed in the vicinity of the discontinuity, hence the existence of a discontinuity curve is implicitly assumed without a test of its existence.

Our approach is based on the statistical properties of the estimated gradient of the variable under study. Zones of abrupt change are characterized by sharp variations of the local expectation, estimated by the local gradient of the kriging surface. However, it is important to distinguish between stochastic variations and significant variations of the estimated gradient, i.e. to assess the significance of these variations taking into account the existence of an underlying random function. The method for detecting ZACs therefore includes three stages. The gradient of the variable is first interpolated by geostatistics. In a second stage, a local test based on the estimated gradient is built for rejecting the null hypothesis of no local abrupt change. In the third stage, the local tests are aggregated into a global test for the existence of ZACs.

The paper is organized as follows. We introduce the method first by describing the theory, discussing practical implementation issues and illustrating it on simulated and real data. We then present our application to a set of soil data, in which the main soil components (clay, silt, sand particles and calcium carbonate content), the soil water and mineral nitrogen content were measured in an arable field. Our analysis show that
the same patterns of ZACs are detected for many variables, thereby providing strong evidence that they are linked to boundaries between soil types. Comparison with the soil map demonstrates that it is possible to extract permanent structures of the field from the analysis of non-permanent data.

Theory

General background

Let us consider that the studied variable \( z(\mathbf{x}) \) (e.g., clay or soil water content) is modeled by a random function \( Z(\mathbf{x}) \). Let \( \mathbf{z} = [z(\mathbf{x}_1), z(\mathbf{x}_2), \ldots, z(\mathbf{x}_n)]^T \) be a sample of a realization of \( Z(\mathbf{x}) \) at the locations \( \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \) in a domain \( \mathcal{D} \). We assume that \( Z(\mathbf{x}) \) is second-order stationary with covariance function \( \text{Cov}[Z(\mathbf{x}), Z(\mathbf{x} + \mathbf{h})] = c_Z(\mathbf{h}) \), for all \( \mathbf{x}, \mathbf{h} \) in \( \mathcal{D} \), which is infinitely differentiable when \( ||\mathbf{h}|| > 0 \).

Under a Gaussian assumption for \( Z(\mathbf{x}) \), when \( c_Z(\mathbf{h}) \) is known and \( \mathbb{E}[Z(\mathbf{x})] \) is unknown, the optimal predictor of \( Z(\mathbf{x}) \) at an unsampled location \( \mathbf{x} \) is the ordinary kriging predictor:

\[
\hat{Z}(\mathbf{x}) = c^T(\mathbf{x})C^{-1}\mathbf{z} + \left[1 - c^T(\mathbf{x})C^{-1}\mathbf{1}\right] \frac{1^TC^{-1}\mathbf{z}}{1^TC^{-1}1},
\]

where \( c(\mathbf{x}) = [c_Z(\mathbf{x} - \mathbf{x}_1), c_Z(\mathbf{x} - \mathbf{x}_2), \ldots, c_Z(\mathbf{x} - \mathbf{x}_n)]^T \) is the covariance vector between \( \mathbf{x} \) and the sample locations, \( C \) is the covariance matrix between the sample locations and \( \mathbf{1} \) is the \( n \)-vector \((1, 1, \ldots, 1)^T\).

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, then \( Z(\mathbf{x}) \) has no derivative, but \( \hat{Z}(\mathbf{x}) \) is differentiable, except at the sample locations. In this case, according to Chilès & Delfiner (1999, p. 314), the gradient of the kriging predictor, \( \nabla \hat{Z}(\mathbf{x}) \), can be considered as an estimator of the local gradient. Differentiating Equation (1) with respect to \( \mathbf{x} \) leads to

\[
\nabla \hat{Z}(\mathbf{x}) = \nabla c^T(\mathbf{x})C^{-1}\mathbf{z} - \nabla c^T(\mathbf{x})C^{-1}1 \frac{1^TC^{-1}1z}{1^TC^{-1}1}
\]

\[
= \nabla c^T(\mathbf{x})C^{-1} \left( I - \frac{11^TC^{-1}}{1^TC^{-1}1} \right) \mathbf{z}
\]

\[
= \nabla c^T(\mathbf{x})K^{-1}\mathbf{z},
\]

(2)
where $\nabla c(x) = \begin{pmatrix} \frac{\partial c(x)}{\partial x_1} \\ \frac{\partial c(x)}{\partial x_2} \end{pmatrix}^T$ is the gradient of $c(h)$ at $h = x = (x_1, x_2)^T \in D$ and $I$ is the $n \times n$ identity matrix.

If $c_Z(h)$ has a discontinuity at the origin (the so-called nugget effect in the geostatistical literature), then $\hat{Z}(x)$ is discontinuous at sample locations, but it is still continuous at all other locations. The above theory is thus applicable if there is some proportion of nugget, as illustrated later. It is not applicable however if the variation is a pure nugget, because in this case $\nabla \hat{Z}(x) = 0$ for all $x$ in $D$.

Detection of Zones of Abrupt Change

For detecting the zones where a variable changes abruptly our general model will be that under the null hypothesis of absence of ZACs $Z(x)$ has a constant expectation in the domain. The alternative is that the expectation has sharp changes. For simulations and for the analysis of the power of the method, a model of the alternative hypothesis must be specified. For example Gabriel & Allard (2007) consider that the expectation is discontinuous along a set of curves; detecting ZACs amounts to estimating them.

We consider that there is an abrupt change at a point $x$ of $D$ if the gradient at $x$ is ‘large’. More formally, we first define a local test for deciding if there is an abrupt change at $x$ by testing the local null hypothesis $H_0(x)$ ‘$E[Z(y)]$ is constant for all $y$ in a neighborhood around $x$’ versus $H_1(x)$ ‘$E[Z(x)]$ presents sharp variations at $x$’. Let us denote by $\Sigma(x)$ the covariance matrix of the gradient estimator, $\Sigma(x) = E[\nabla \hat{Z}(x) \nabla \hat{Z}^T(x)] = \nabla c^T(x)K^{-1}\nabla c(x)$. We define the test statistic, $T(x)$, by:

$$T(x) = \nabla \hat{Z}^T(x) \Sigma^{-1}(x) \nabla \hat{Z}(x).$$

(3)

We perform the change of variable $U(x) = \Sigma^{-1/2}(x) \nabla Z(x)$, where $U(x) = [U_1(x), U_2(x)]^T$ and $U_i(x) \sim N(0, 1)$, $i = 1, 2$. This transformation is performed point-wise. It can be shown that $T(x) = U_1^2(x) + U_2^2(x)$ (see Appendix). According to standard results in statistics (von Mises, 1964, p. 406), the test statistic has a marginal $\chi^2_2$ distribution. The null hypothesis $H_0(x)$ is rejected if the test statistic is above the $(1 - \alpha)$-quantile of the $\chi^2_2$ distribution, denoted $t_{1-\alpha}$. For a confidence level $1 - \alpha$, we define the potential
ZACs as the set of points $x$ such that the test statistic is above $t_{1-\alpha}$:

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

If the field is stationary, then we should find a proportion of about $\alpha$ points $x$ for which the local null hypothesis is rejected. Under the alternative hypothesis, the proportion should be larger and we expect the potential ZACs to be spatially structured. In the specific case of discontinuity curves, potential ZACs will be structured along them.

Point-wise tests for each point $x$ of $\mathcal{D}$ now need to be aggregated to test globally the existence of sharp variations. The usual multiple testing techniques (Benjamini & Yekutieli, 2001; Dudoit et al., 2003) cannot be used here because they do not take into account the spatial dependences. Instead, we propose to aggregate the local tests using geometrical properties of the connected clusters of the potential ZACs; see Adler (2000) for a complete review on excursion sets of Gaussian and related random fields. Allard et al. (2005) showed that under the null hypothesis the area of a cluster of the potential ZACs is exponentially distributed:

$$t_{1-\alpha}S_{1-\alpha} \rightarrow \pi \det(\Lambda)^{-1/2}E(2), \text{ as } t_{1-\alpha} \rightarrow \infty,$$  

where $S_{1-\alpha}$ is the area of a cluster of the potential ZACs above the level $t_{1-\alpha}$, $\Lambda$ is the $2 \times 2$ matrix of the curvature of $T(x)$ at its maximum in the cluster, which only depends on the covariance function $c_Z(h)$ and the sampling pattern, and $E(2)$ is an exponential random variable with expectation 2 (see Appendix). This result enables us to compute for each cluster of the potential ZACs its $p$ value under the null hypothesis:

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

We assess the significance of the cluster by comparing this $p$ value to a confidence level $\eta$, for example $\eta = 5\%$. If $P$ exceeds $\eta$, it is considered not significant and is discarded. On the contrary, it is considered significant and defines a ZAC.

The power of the method was assessed by simulations (Gabriel, 2004; Allard et al., 2005). A series of 1000 samples of size $n$ from a standard Gaussian random field with
an exponential covariance function \( c_Z(h) = \exp(-\|h\|/b) \) were randomly located in the unit square. The region was divided into two rectangular subregions. A constant \( a \) was added to the values of the points located in one subregion such that the mean difference between the subregions is \( a \). For \( n = 100 \), \( a = 2.5 \) and \( b = 0.1 \) and \( 0.2 \) the discontinuity was detected on more than 80 \%, and 95 \% respectively, of the simulations. For \( n = 200 \) the discontinuity was detected on more than 90 \% of the simulations for \( b = 0.1 \).

**Implementation**

In practice the method is run on a grid superimposed over the domain. At each grid node \( x_p \), the gradient \( \nabla \hat{Z}(x_p) \), its covariance matrix \( \Sigma(x_p) \) and the test statistic \( T(x_p) \) are computed. Then for a confidence level \( \alpha \), the set of grid nodes whose test statistic \( T(x_p) \) exceeds the \((1-\alpha)-\text{quantile of a } \chi^2_2 \) distribution defines the potential ZACs. For each potential ZAC, the \( p \) value is then computed according to Equation (5).

**Finding the local level \( \alpha \)**

The method needs two levels of significance: a global level \( \eta \), chosen by the user; and a local level \( \alpha \), related to \( \eta \), for detecting potential ZACs. Equation (5) is valid on the plane when \( t_{1-\alpha} \) is very high, i.e. when \( \alpha \) is very small. But as \( \alpha \) tends to 0 the potential ZACs get potentially smaller than the mesh of the interpolation grid. Hence, we have to consider both the mathematical requirement and the ability to detect the potential ZACs in order to determine the level \( \alpha \). The correct level \( \alpha \) thus depends on \( \eta \) and the covariance function through Equation (5), but it also depends on the discretization of the interpolation grid and the sampling pattern. We assess this level by Monte Carlo simulations: we simulate \( M \) realizations of standard Gaussian random fields, corresponding to the null hypothesis of absence of ZACs, on the same sample locations, using the estimated covariance function (see below for estimation of the covariance function). The estimated local level \( 1-\hat{\alpha} \) is such that ZACs are detected on \( \eta M \) simulations. When the discretization of the interpolation grid is sufficiently fine, a level \( \alpha_G \) linked to the integral range of \( c_Z(h) \) is a good approximation to \( \hat{\alpha} \) and needs
fewer computations (Gabriel, 2004). The integral range (Lantuéjoul, 1991) allows one to determine the number of independent local tests that we can consider on \( \mathcal{D} \). The integral range, \( A \), of the covariance function \( c_Z(h) \) is the quantity:

\[
A = \int_{\mathbb{R}^2} \frac{c_Z(h)}{c_Z(0)} \, dh.
\]  

(6)

We denote by \( |\mathcal{D}| \) the area of \( \mathcal{D} \) and by \( \bar{Z}(\mathcal{D}) \) the spatial average of \( Z(\cdot) \) on \( \mathcal{D} \). If the integral range is not null and \( |\mathcal{D}| \) is larger than the integral range, according to Lantuéjoul (1991), the variance of \( \bar{Z}(\mathcal{D}) \) is approximately \( \sigma^2 / N \), with \( N = |\mathcal{D}| / A \).

Then we get an approximation of the global level

\[
\eta = P \left[ \bigcup_{x \in \mathcal{D}} \{ T(x) \geq t_{1-\alpha} \} \right] \approx 1 - \left( P \left[ T(x) \leq t_{1-\alpha} \right] \right)^N,
\]

leading to the definition of the local level:

\[
\alpha_G = 1 - (1 - \eta)^{1/N}.
\]  

(7)

**Estimating the covariance function**

The method is based on the assumption that the covariance function of \( Z(x) \) is known but in practice it must be estimated. For a stationary random field, the variogram

\[
\gamma(h) = \frac{1}{2} E[(Z(x) - Z(x + h))^2]
\]

is related to the covariance function: \( \gamma(h) = c_Z(0) - c_Z(h) \). We estimate the covariance function via the estimation of the variogram, which is widely used in the geostatistical literature. When fitting a variogram, we must choose a parametric family of variograms, and then estimate the parameters within this family.

Simulation studies reported in Gabriel (2004) and Allard et al. (2005) have shown that changing the parametric family of covariance functions from exponential to spherical, or adding a nugget component, did not change significantly the detected ZACs. The same simulations showed that the method is not too sensitive to variations of the range parameter. On the other hand, the method is very sensitive to the sill of the variogram. The method is run under the null hypothesis of the absence of ZACs. In
the presence of a discontinuity the variance $\sigma^2$ of $Z(x)$ is thus overestimated. Since $T(x)$ is inversely proportional to $\sigma^2$, this will result in smaller values of the test statistic $T(x)$, hence to smaller ZACs, higher $p$ values and ultimately to a loss of power, i.e. a loss of ability to detect ZACs. To remove this shortcoming we propose the following iterative procedure.

We first estimate a global variogram (i.e. using all pairs of samples) and a first set of ZACs. If potential ZACs are detected, we re-estimate the variogram by discarding all pairs of samples $\{Z(x_k), Z(x_l)\}$ for which the segment $[x_k, x_l]$ intersects a ZAC.

We estimate of the variogram by a least squares fit of a theoretical function to the experimental variogram $\hat{\gamma}(h) = \frac{1}{2n_h} \sum_{j=1}^{n_h} \{z(x) - z(x + h)\}^2$, where $n_h$ is the number of pairs involved in the sum. If the range of the variogram changes between iterations, the level $\alpha$ must be re-estimated. The ZACs are then re-estimated with the new parameters and the procedure is iterated until the set of ZACs remains unchanged. In all the situations we tested, this happened in less than five iterations. This procedure greatly improves the estimation of the parameters of the variogram in presence of ZACs (Gabriel, 2004) as illustrated below.

Illustrations

Simulation study

The method was implemented in the statistical software R. Source code can be downloaded from E. Gabriel’s web page: www.maths.lancs.ac.uk/~gabriel/. We illustrate the method first on simulated data and then on real data taken from the literature.

A sample of a standard Gaussian random field $Z(x)$ with an exponential covariance function $c_Z(h) = \exp(-||h||/b)$ with $b = 0.1$ is simulated at 100 randomly located points on the unit square, thus defining a vector $z$. We introduced a discontinuity by adding a constant $a$ to all samples located above a sinusoidal curve (dashed line on Figure 1a). The case $a = 0$ thus corresponds to $H_0$. A realization of this process is depicted in Figure 1a for $a = 2.5$ with dots proportional to the value at the sample
location. At the first iteration, the estimated parameters of the variogram $\gamma(h) = c_0 + c_1 \{1 - \exp(-||h||/b)\}$, where $c_0$ denotes the nugget variance and $\sigma^2 = c_0 + c_1$ is the sill, are $\hat{b} = 0.193$, $\hat{c}_0 = 0.547$ and $\hat{c}_1 = 1.602$ (i.e. $\hat{\sigma}^2 = 2.149$). On Figure 1b, each pixel of a $60 \times 60$ grid is coloured if the test statistic $T$ in this pixel is above the threshold $t_{1-\hat{\alpha}}$, where $1 - \hat{\alpha} = 0.9988$ corresponds to the global level $\eta = 0.05$. Detected potential ZACs are structured along the discontinuity. Black pixels correspond to significant potential ZACs, while grey pixels correspond to the non-significant ones. Convergence occurs at the fourth iteration, with parameters $\hat{b} = 0.167$, $\hat{c}_0 = 0.055$ and $\hat{c}_1 = 1.183$ (i.e. $\hat{\sigma}^2 = 1.238$) and $1 - \hat{\alpha} = 0.9989$. The estimated variance $\hat{\sigma}^2$ at convergence decreased by 42% compared to the initial estimate. Figure 1c shows that ZACs are better detected at convergence. Figure 1d represents the lines of maximum rate of change within the detected ZACs. We also applied the iterative procedure for variograms without a nugget effect, i.e. $\gamma(h) = \sigma^2 \{1 - \exp(-||h||/b)\}$. The estimated parameters at the first and last (fourth) iterations are $\hat{b} = 0.137$, $\hat{\sigma}^2 = 2.083$ and $\hat{b} = 0.128$, $\hat{\sigma}^2 = 1.21$ respectively. The detected ZACs at convergence are very similar to the ones observed when the variogram contains a nugget component.

**Figure 1 please here**

Because most variograms with which soil scientists have to work with in practice have nugget components, we performed a similar simulation incorporating a nugget $c_0$ equals to 30% of the total variance (Figure 1e). At the first iteration, the estimated parameters are $\hat{b} = 0.412$, $\hat{c}_0 = 0.645$, $\hat{\sigma}^2 = 3.172$ and $1 - \hat{\alpha} = 0.9984$. Convergence occurred at the third iteration with $\hat{b} = 0.137$, $\hat{c}_0 = 0.311$, $\hat{\sigma}^2 = 1.501$ and $1 - \hat{\alpha} = 0.9940$. Figures 1f and 1g show the detected ZACs at the first iteration and at convergence, and Figure 1d shows the lines of maximum rate of change within the detected ZACs at convergence. In this case, detected ZACs are smaller due to the high fraction of nugget variance.

**Swiss Jura**

The method is now applied to a set of soil data in a $14.5$ km$^2$ region of the Swiss
Jura. These data were obtained and first analysed by Atteia et al. (1994). The concentrations of seven potential soil pollutants, the heavy metal, cadmium, chromium, cobalt, copper, lead, nickel and zinc, were measured at 359 sample locations (Figure 2a). The analysis of these concentrations allowed the authors to evaluate the risk of soil pollution and to determine the sources of pollutants. The region is composed of Jurassic limestones (Figure 2a). Using a one-way analysis of variance, Atteia et al. (1994) showed that geology, and in particular the difference between Argovian and the other formations, is the factor most strongly related to heavy metal concentrations. Here, we focus on the concentrations of cobalt and nickel. Their interpolation by ordinary kriging over a 100 × 100 grid is given on Figures 2b and 2c respectively, indicating low concentrations in the northern, central-eastern and southern parts of the region. We fitted exponential covariance functions to cobalt and nickel. The iterative procedure converged at the second iteration for cobalt and at the third for nickel. The lines of maximum rate of change in the detected ZACs are superimposed on the interpolation maps. Figures 2b and 2c show that the same ZACs are detected for cobalt and nickel in the northern, central-eastern and southern parts of the region. A visual comparison with the geological map indicates that they are located along the boundaries between Argovian and other formations. These results lead to the same conclusion as the one of Atteia et al. (1994): patterns of distribution for cobalt and nickel concentrations are similar to those of the geological map, thus suggesting that these metals mainly derive from the bedrock.

Figure 2 please here

Application to an agricultural field

Data were acquired in an agricultural field of 10 ha in Chambry, northern France, on a regular grid with a distance between nodes of 36 m (Guérif et al., 2001). Figure 3a illustrates the sampling pattern of soil variables. Particle-size fractions (clay, sand, silt) and the calcium carbonate content in soil were measured in the ploughed layer.
(0-30 cm) at one date, and soil water and mineral nitrogen content were measured over a 0-120 cm depth at dates between March 2000 and February 2003.

**Soil types**

Based on the analysis of the soil cores, on additional sampling and on qualitative observations, a soil map was proposed by soil scientists (Nicoullaud *et al.*, 2007). The simplified map includes the four main soil types (S1-S4, see Figure 3b). The central-east part (summit) of the field and its south to south-west slope, both of which are affected by erosion, are occupied by S1: sandy clay loam soils (Calcaric Cambisol; IUSS Working Group, 2006) which are calcareous and lie over soft chalk; and S2: stony, calcareous soils (stony Calcaric Cambisol) lying over hard chalk. The opposite slope (north and north-east) is occupied by S3: loamy clay soils (Luvic Cambisol and Haplic Luvisol) lying over loess; and S4: loamy sandy clay soils (magnesian Calcaric Cambisol) lying over sandstone which are on a transversal axis from the summit to the north-west part of the field.

**Figure 3 please here**

**Soil data**

The concentrations of the main soil components (clay, silt, sand particles and calcium carbonate content, in g kg$^{-1}$) were determined on soil cores of 30 cm depth, taken at the 85 points of the sample grid displayed Figure 3a. Figure 4a depicts the distribution of each of these variables. This figure shows in particular that calcium carbonate content (hereafter called chalk) is the most variable (from almost 0 to more than 400 g kg$^{-1}$) and that sand content may also reach to 400 g kg$^{-1}$.

**Figure 4 please here**

We interpolated values from the data to a $64 \times 98$ grid with a mesh size of 5 m. For conciseness, only clay is illustrated (Figure 5a). Following the soil distribution (Figure 3b), the northern part of the field is clayey and the southern part is chalky. A sandy
and chalky spot appears on the summit of the field, associated with the sandy soil described above.

Soil mineral nitrogen, $\text{NH}_4^+ + \text{NO}_3^- - N$ (in kg ha$^{-1}$), and gravimetric soil water content (in %) were measured on the same grid (Figure 3a), on cores of 120 cm depth and 4 cm diameter. Each soil sample was itself a mixture of three cores taken at 50 cm from one another. We extracted soil mineral nitrogen by mixing 50 g soil with 100 ml of potassium chloride ($0.5 \text{ mol l}^{-1}$), shaking the mixture for half an hour and allowing it to settle for 2 hours. An aliquot of the extract was then taken and analysed for mineral nitrogen (i.e. $\text{NH}_4$ and $\text{NO}_3$ forms) by continuous flow colorimetry. Soil water content was measured by a gravimetric method after 48 hours in an oven at 105 °C. Similar data, collected in a nearby field, have been presented and analysed in Mary et al. (2001). The cropping sequence was the following: sugar beet (2000), winter wheat (2000-2001), spring peas (2002) and winter wheat (2002-2003). Sampling for soil water and mineral nitrogen content was done on eight different dates corresponding to sowing or harvest dates or end of winter for wheat crops (Table 1). The number of sample points was generally around 80, except for October 2000, for which there are only 67 samples due to the difficulty of extracting cores from a very dry soil in the southern part of the field.

**Table 1 please here**

Figure 4b and Figure 4c are the box-plots for these variables. The figures show large variation in water content depending on the sampling dates. At the end of winter (February 2001 and 2003; March 2000 and 2002) the soil water content has to the maximum water holding capacity and its variation is similar for these dates. The other dates correspond to harvest: at this period the soil was drier because the crops had extracted so much water (except for October 2002 where sugar beet was harvested in wet conditions). Figures 5b and 5c display the interpolation map of water content in August 2001 and February 2003 on the $64 \times 98$ grid. The interpolation maps of other dates (not shown) are very similar to those presented here. At the end of the winter,
soil in the southern part of the field contained more water than the northern part did. Spatial variation is less visible at harvest. There is a small dry zone in the eastern part of the field on all dates and in the western part except for March 2002 and February 2003, corresponding to the sandy and chalky summit zone.

In contrast, mineral nitrogen content exhibits greater temporal variation. This is a consequence of the biological processes in the soil-plant system and to cultural practices: mineral nitrogen content is least at harvest (October 2000, August 2001, July 2002). Spatial variation in mineral nitrogen was also smaller on these dates, indicating that crop uptake smoothes the distribution. The greater spatial variation, associated with the largest mean values, occurred in October 2002, i.e. in bare soil 4 months after the pea harvest. These large values and the associated variation are attributed to the important production of NO$_3^-$ by mineralization after harvest, during summer and autumn, which has a strong soil type dependence.

**Parameter estimation**

Exponential covariance functions, $c_Z(h) = \sigma^2 \exp(-\|h\|/b)$, $\sigma^2 > 0$ and $b > 0$, were estimated for all variables and at each step of the iterative procedure. From Equation (6), the corresponding integral range is $A = 2\pi b^2$. The local level derived from Equation (7) is then: $\alpha_G = 1 - (1 - \eta)^{2\pi b^2/|D|}$, with $\eta$ fixed e.g. to 5%.

Table 2 reports the estimated range and sill parameters at the first and last iterations ($k$th, $k \geq 1$). Convergence always occurred in fewer than four iterations. At convergence the estimated variance is usually less than that estimated at the first iteration, showing that the error in estimating the variance arises if we ignore sharp changes. For mineral nitrogen, the iterative procedure always stopped at the first iteration, either because there was no detected ZAC or because the detected ZACs are so small that the estimated covariance function remains unchanged. Note that for chalk the estimated parameters of the covariance function indicate that variograms behave almost linearly. In this case, the integral range and thus $\alpha_G$ could not be computed. We therefore estimated the level of $\alpha$ using $\hat{\alpha}$, i.e. by Monte Carlo simulations.
Detection of ZACs

Table 2 contains the number $s_+$ of significant detected potential ZACs and the number $s_-$ of non-significant ones computed for all variables at each step of the iterative procedure. It indicates that ZACs were always detected for the permanent data and for water content (except for October 2000), and much more rarely for mineral nitrogen content.

Figure 5 displays the lines of the maximum rate of change within the ZACs detected at convergence for clay (Figure 5a), water content in August 2001 (Figure 5b) and water content in February 2003 (Figure 5c). For chalk, clay, sand, silt and water content (except for October 2000), ZACs were detected in the same eastern part of the field. For October 2000, the same ZACs were detected from water content at all sample locations over a 90 cm depth. This ZAC was not found when we considered water content to the full 120 cm depth because some missing data in the southern part of the field led to a loss of power of the detection test (Gabriel & Allard, 2007). We detected ZACs in the western part of the field also for water and mineral nitrogen content in August 2001.

The detected ZACs can be related to the transitions between the main soil types (Figure 3b). We did a one-way analysis of variance with the main soil types as factors. Table 2 gives the proportion of explained variance ($R^2 = \text{ratio between the variance between soil types and the total variance}$) and the average values for each soil type. Soil types explain more than 50 % of the variance of water content for all dates (except October 2000 and August 2001) and the most important effect is attributed to the difference between type S4 (sand) and the other types; water content is strongly related to the water holding capacity of soils, with sandy soil having the smallest water holding capacity. Therefore, ZACs found in the eastern part of the field for water content are
the same as those found for the particle-size fractions (even if they are only in the top soil). They clearly delineate the fourth soil type (summit zone), whereas those identified in the western part (October 2000, February 2001 and August 2001) correspond to the transition between calcareous soil in the southern part and less calcareous and more silty soils in the northern part of the field. Zones of abrupt change were rarely detected in this western part because the transition is smoother.

Concerning mineral nitrogen content, ZACs were detected only for August 2001. The analysis of variance indicates that mineral nitrogen content is only weakly related to the main materials of the soil. This means that biological processes which determine local amount of mineral nitrogen do not respond similarly to soil properties or vary at a smaller scale than the measuring mesh used here (36 m). The spatial variation of cropping practices, particularly the distribution of crop residues and fertilizer application, may contribute to this short range (about 1 m) variability, and might explain why no significant ZAC was detected.

Discussion

The method was used to analyse the variation of soil properties. We have shown that it allowed us to identify abrupt transitions in the main permanent variables of the soil. Remarkably, ZACs in water content were detected at the same locations for all dates (provided that the sampling pattern was appropriate). These structures correspond to the main transition areas on the soil map independently drawn by soil pedologists. It is worth emphasizing that permanent structures could be retrieved from non-permanent data that are highly correlated with the permanent constituents of the soil. None or very few significant structures were found for soil mineral nitrogen content. This is strong evidence that, in this field, mineral nitrogen content is only marginally related to the soil’s constituents.

In an attempt to find ZACs for mineral nitrogen content, one could artificially increase the level $\alpha$. More potential ZACs would then be detected, but at the cost of a larger false discovery rate. However, for large values of $\alpha$ (say, $\alpha = 0.05$) the
convergence given in Equation (4) does not hold, and Equation (5) cannot be used to compute the $p$ values of the potential ZACs. A crude application of the method on mineral nitrogen content, with $\alpha = 0.01$ and $\alpha = 0.05$ has shown that no ZAC was to be found. Actually, the fact that no ZAC was detected for mineral nitrogen content must be considered as a major improvement of our method compared with the pre-existing ones which assume implicitly the existence of ZACs.

Sparse sampling does not allow us to test if the estimated local gradient corresponds to a smooth or a locally sharp transition. This question was addressed in Gabriel & Allard (2007), who investigated the local power of the method. The power at a point $x$ was defined as the probability to detect an abrupt change at this point under the hypothesis of the existence of a discontinuity curve containing $x$. Mapping the local power shows that it is not constant in the domain. Zones with low power indicate that the local sampling pattern is not appropriate for estimating ZACs, in particular because the local sampling is too sparse. Hence, the absence of ZACs in cases similar to that of water content for October 2000, where the sampling pattern is heterogeneous, could be due either to the lack of data or to the true absence of abrupt change. Assessing the power of the detection method allows us to distinguish between these two alternatives and permits us to visualize areas where ZACs cannot be detected because of an inappropriate local sample density. This information can then be used to decide in which area additional samples are necessary. Going back to the analysis of the water content in October 2000, mapping the local power allows us to conclude that the absence of a detected ZAC is due to a lack of samples.

Our Gaussian assumption makes it possible to take into account explicitly the autocorrelation between the samples and to detect sharp variations for a low density of samples in presence of autocorrelation. If the data cannot be considered as Gaussian, they can be transformed to an approximately Gaussian scale, using a parametric form (e.g. log transform, Cox transform) or a non-parametric quantile transform. This does not ensure normality of the random field but should be enough for the method.
to work reasonably well in practice. Transforming the data is not always necessary. Note, however, that even when our Gaussian assumption does hold, the marginal distribution of the data may be far from Gaussian because of the effect of hidden sharp changes.

Conclusions

We have considered that soil variations can be modelled as the sum of a trend, which may include sharp variations, and a stationary spatially correlated residual field. In this framework, we proposed a method for testing whether the data present significant ZACs corresponding to sharp variations or discontinuities of the mean, versus the null hypothesis of a spatially continuous model. As a result, if a variable appears as a correlated random function with a constant or slowly varying mean, as is the case in the standard geostatistical models, then no ZAC will be detected.

This method is an exploratory tool for the researcher. ZACs correspond to the primary boundaries discussed in Burrough et al. (1997). In common parlance, a boundary in space must delineate a region, either entirely or in conjunction with other boundaries (coast, rivers, field margins, etc.). Because ZACs are not necessarily closed curves, they do not define boundaries, but they are very good candidates for the parts of the boundaries corresponding to sharp variations. In our case study, ZACs detected on the water content correspond to the sharp transitions of the variable under study. The boundary between the soil types will thus be a combination of primary boundaries, corresponding to the ZACs, and secondary boundaries within the regions of smooth variations.

A potentially interesting follow-up of this work would be to take into account the estimated ZACs in the interpolation procedure. Heuvelink & Webster (2001) discuss the different models that have been proposed in the literature to bridge the gap between discrete and continuous models: weighted average of kriging, stratified kriging or classification external drift kriging as in Monestiez et al. (2001). A new approach could be proposed: interpolation at a point $x_0$ would be performed using only those
sample points \( x_i \) for which the segment \([x_0, x_i]\) does not intersect a detected ZAC. This neighborhood selection would ensure that all samples used for the kriging would have similar mean.

Several problems remain, among which is the generalization to multivariate data and taking into account temporal information. The first requires a multivariate extension of Theorem 1 and an automatic procedure for estimating the variograms.

Acknowledgements

This research was done as part of E. Gabriel’s doctoral thesis in the Biometry Unit of INRA-Avignon. It was funded by INRA and CETIOM through their project on Precision Agriculture. We are grateful to G. Alavoine, F. Barrois, D. Boitez, O. Delfosse, E. Gréhan, C. Herre, F. Mahu, F. Millon and E. Venet of the Agronomy Unit (INRA-Laon) for soil sampling in the field and analysis in the laboratory. Finally, we thank both Referees, the Associate Editor and R. Webster for helping us to improve the quality of the manuscript.

References


www.maths.lancs.ac.uk/~gabriel/papers/TheseGabriel.pdf


Appendix

For $x = (x^1, x^2)^T \in \mathcal{D}$, we can write the statistic $T(x)$ defined in Equation (3), as the sum of the squares of non-stationary, non-independent Gaussian random fields, $U_1(x)$ and $U_2(x)$:

$$U_1(x) = \frac{1}{\sigma_1(x)} \frac{\partial \hat{Z}(x)}{\partial x^1},$$

$$U_2(x) = \frac{1}{\sqrt{1 - \rho^2(x)}} \left\{ \frac{1}{\sigma_2(x)} \frac{\partial \hat{Z}(x)}{\partial x^2} - \rho(x) \frac{1}{\sigma_1(x)} \frac{\partial \hat{Z}(x)}{\partial x^1} \right\},$$

where $\sigma_i(x) = \sqrt{\Sigma_{[ii]}(x)}$, $i = 1, 2$ and $\rho(x) = \Sigma_{[12]}(x)/\{\sigma_1(x)\sigma_2(x)\}$.

Potential ZACs are defined as the set of points of $\mathcal{D}$ such that $T(x)$ exceeds that for $t_{1-\alpha}, 1-\alpha$-quantile of the $\chi^2_2$ distribution. They are thus excursion sets of $\chi^2$ random fields (Adler, 1981, 2000; Aronowich & Adler, 1988; Worsley, 1994; Cao, 1999). When
$t_{1-\alpha}$ tends to $\infty$, (i.e. when $\alpha$ tends to 0), it can be shown (Allard et al., 2005) that
the area, $S_0$ of a cluster, $C_0$, of this excursion set is related to the curvature of $T(x)$
near its local maximum (shifted at $0$) through the following convergence in law.

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

$$t_{1-\alpha}S_0 \xrightarrow{L} \pi \det(\mathbf{\Lambda})^{-1/2} E(2),$$

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

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

$$\mathbf{\Lambda}_i(x) = E[\nabla U_i(x)\nabla U_i^T(x)] = \text{Var}(\nabla U_i(x)).$$

Write $\sigma_i = \sigma_i(x), i = 1, 2$ and $\rho = \rho(x)$. For notational convenience, also write $D_i = D_i(x) = \partial_i c(x)$ for $i = 1, 2$ where $\partial_i f = \partial f/\partial x^i$. Then, for $k, l = 1, 2$:

$$\mathbf{\Lambda}_1[k,l](x) = \{\partial_k D^T_1 K^{-1} \partial_l D_1 - \partial_k \sigma_1 \partial_l \sigma_1\}/\sigma_1^2,$$

$$\mathbf{\Lambda}_2[k,l](x) = \{\partial_k D^T_2 K^{-1} \partial_l D_2/\sigma_2^2 - \partial_k \sigma_2 \partial_l \sigma_2/\sigma_2^2 + \partial_k \rho \partial_l \rho - \partial_k \rho \partial_l D^T_2 K^{-1} D_1/\sigma_1 \sigma_2$$

$$- \partial_k \rho \partial_l D^T_2 K^{-1} D_1/\sigma_1 \sigma_2 + \rho((\partial_k \sigma_2/\sigma_2)(\partial_l D^T_1 K^{-1} D_2/\sigma_1 \sigma_2 + \partial_l \rho))$$

$$+ (\partial_k \sigma_2/\sigma_2)(\partial_l D^T_1 K^{-1} D_2/\sigma_1 \sigma_2 + \partial_l \rho) + (\partial_k \sigma_1/\sigma_1)(\partial_l D^T_2 K^{-1} D_1/\sigma_1 \sigma_2)$$

$$+ (\partial_k \sigma_1/\sigma_1)(\partial_l D^T_2 K^{-1} D_1/\sigma_1 \sigma_2) - \partial_k D^T_2 K^{-1} \partial_l D_2/\sigma_1 \sigma_2$$

$$- \partial_k D^T_2 K^{-1} \partial_l D_1/\sigma_1 \sigma_2 + \rho^2 [\partial_k D^T_1 K^{-1} \partial_l D_1/\sigma_1^2 - \partial_k \sigma_1 \partial_l \sigma_1/\sigma_1^2$$

$$- \partial_k \sigma_1 \partial_l \sigma_2/\sigma_1 \sigma_2 - \partial_k \sigma_2 \partial_l \sigma_1/\sigma_1 \sigma_2]/(1 - \rho^2) - \rho^2 (\partial_k \rho \partial_l \rho)/(1 - \rho^2)^2,$$

with $\partial_k \rho = (\partial_k D^T_1 K^{-1} D_2 + \partial_k D^T_2 K^{-1} D_1)/\sigma_1 \sigma_2 - \rho (\partial_k \sigma_1/\sigma_1 + \partial_k \sigma_2/\sigma_2)$ and $\partial_k \sigma_i = \partial_k D^T_i K^{-1} D_i/\sigma_i$. Details and a proof of these results can be found in Allard et al. (2005).
Table 1 Crop stage and number of sample points of the non-permanent data.

<table>
<thead>
<tr>
<th>Date</th>
<th>Number of points</th>
<th>Crop stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>March 2000</td>
<td>85</td>
<td>sugarbeet sowing</td>
</tr>
<tr>
<td>October 2000</td>
<td>67</td>
<td>sugarbeet harvest and wheat sowing</td>
</tr>
<tr>
<td>February 2001</td>
<td>85</td>
<td>wheat tillering stage</td>
</tr>
<tr>
<td>August 2001</td>
<td>79</td>
<td>wheat harvest</td>
</tr>
<tr>
<td>March 2002</td>
<td>84</td>
<td>pea sowing</td>
</tr>
<tr>
<td>July 2002</td>
<td>82</td>
<td>pea harvest</td>
</tr>
<tr>
<td>October 2002</td>
<td>83</td>
<td>wheat sowing</td>
</tr>
<tr>
<td>February 2003</td>
<td>84</td>
<td>wheat tillering stage</td>
</tr>
</tbody>
</table>
Table 2 For all variables: parameters estimation (range $\hat{b}$ and sill $\hat{\sigma}^2$) of the exponential covariance functions, number of significant and non-significant potential ZACs, $s_+$ and $s_-$ respectively, at the first iteration (It.=1) and at convergence and results of analysis of variance by soil types (S1: Calcaric Cambisol, S2: stony Calcaric Cambisol, S3: Luvic Cambisol and Haplic Luvisol, S4: magnesian Calcaric Cambisol): proportion of explained variance ($R^2$) and averaged values within each soil type.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Covariance estimation</th>
<th>Potential ZACs</th>
<th>Analysis of variance</th>
<th>Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$b$</td>
<td>$\hat{\sigma}^2$</td>
<td>$s_+$</td>
<td>$s_-$</td>
</tr>
<tr>
<td>Clay</td>
<td>1</td>
<td>79</td>
<td>13937</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>95</td>
<td>1696</td>
<td>1</td>
</tr>
<tr>
<td>Silt</td>
<td>1</td>
<td>19</td>
<td>198</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>28</td>
<td>172</td>
<td>1</td>
</tr>
<tr>
<td>Sand</td>
<td>1</td>
<td>25</td>
<td>1713</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>33</td>
<td>679</td>
<td>1</td>
</tr>
<tr>
<td>Chalk</td>
<td>1</td>
<td>388</td>
<td>35834</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1001</td>
<td>63708</td>
<td>1</td>
</tr>
<tr>
<td>Water</td>
<td>mar-00</td>
<td>1</td>
<td>23</td>
<td>18.9</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>30</td>
<td>19.3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>oct-00</td>
<td>1</td>
<td>26</td>
<td>17.4</td>
</tr>
<tr>
<td></td>
<td>feb-01</td>
<td>1</td>
<td>25</td>
<td>20.3</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>30</td>
<td>11.8</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>aug-01</td>
<td>1</td>
<td>27</td>
<td>21.7</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>39</td>
<td>17.9</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>mar-02</td>
<td>1</td>
<td>23</td>
<td>15.3</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>32</td>
<td>14.4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>jul-02</td>
<td>1</td>
<td>32</td>
<td>20.5</td>
</tr>
<tr>
<td></td>
<td>oct-02</td>
<td>1</td>
<td>27</td>
<td>20.2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>37</td>
<td>19.2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>feb-03</td>
<td>1</td>
<td>25</td>
<td>23.5</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>32</td>
<td>23.8</td>
<td>2</td>
</tr>
<tr>
<td>Nitrogen</td>
<td>mar-00</td>
<td>1</td>
<td>21</td>
<td>246</td>
</tr>
<tr>
<td></td>
<td>oct-00</td>
<td>1</td>
<td>29</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>feb-01</td>
<td>1</td>
<td>26</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>aug-01</td>
<td>1</td>
<td>48</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>mar-02</td>
<td>1</td>
<td>13</td>
<td>323</td>
</tr>
<tr>
<td></td>
<td>jul-02</td>
<td>1</td>
<td>27</td>
<td>222</td>
</tr>
<tr>
<td></td>
<td>oct-02</td>
<td>1</td>
<td>41</td>
<td>719</td>
</tr>
<tr>
<td></td>
<td>feb-03</td>
<td>1</td>
<td>15</td>
<td>136</td>
</tr>
</tbody>
</table>
Figure 1 Detection of the ZACs from 100 randomly located samples of a standard stationary Gaussian random field with the addition of a constant above the dashed curve. The covariance function is exponential (a-d) and includes a nugget component (e-h). (a) and (e) Location of the samples; (b) and (f) Detection of the ZACs at the first iteration; significant potential ZACs are in black; non-significant ones are in grey; (c) and (g) Detection of the ZACs at convergence; (d) and (h) Lines of maximum rate of change in the detected ZACs.

Figure 2 (a) Geological map and sampling pattern; Interpolation map and lines of maximum rate of change in the detected ZACs in the concentration of: (b) cobalt, (c) nickel.

Figure 3 (a) Sample grid; b) Map of the four main soil types: S1 (dark grey): Calcaric Cambisol; S2 (medium grey): stony Calcaric Cambisol; S3 (light grey): Luvic Cambisol and Haplic Luvisol; S4 (white): magnesian Calcaric Cambisol.

Figure 4 Box-plots of (a) clay, silt sand and chalk (g kg$^{-1}$), (b) soil water content (%) and (c) soil mineral nitrogen content (kg ha$^{-1}$). The lower and upper values of the boxes correspond to the first and third quartile, while the bold bar represents the median. Minimum and maximum values are also reported (dotted line).

Figure 5 Interpolation map and lines of maximum rate of change in the detected ZACs for (a) clay; (b) soil water content in August 2001 and (c) soil water content in February 2003.
(a) Argovian
Sequanian
Kimmeridgian
Portlandian
Lehm

1 km

(b) 3% - 32%

(c) 4% - 17%