DÉTECTION DE CHANGEMENTS ABRUPTS DANS LE GRADIENT D'UN CHAMP GAUSSIEN ET APPLICATION AUX SCIENCES DE L'ENVIRONNEMENT

Edith GABRIEL [1]

RÉSUMÉ

Ce papier propose une méthode pour estimer et tester les zones où une variable échantillonnée dans le plan varie brusquement. Ces zones sont appelées Zones de Changement Abrupt (ZCAs). La méthode repose sur les propriétés statistiques du prédicteur du gradient local de la variable. Une statistique de test local définie à partir de celui-ci est comparée à un seuil critique calculé sous l'hypothèse nulle d'une moyenne constante sur le domaine d'étude. Cela permet de définir les ZCAs potentielles comme l'ensemble des points où le test local rejette l'hypothèse nulle. Afin de tester la significativité globale des ZCAs potentielles détectées, les tests locaux sont ensuite agrégés en utilisant les propriétés géométriques des composantes connexes des ZCAs potentielles. Le schéma d'échantillonnage, et en particulier sa densité locale, détermine la puissance du test local. La cartographie de cette puissance est illustrée et permet d'identifier les zones où d'éventuelles ZCAs peuvent être détectées ou non. La méthodologie est appliquée à des données de sol prélevées dans une petite région du Jura suisse. L'analyse des teneurs en métaux lourds tels que le nickel et le cobalt révèle les principales structures géologiques de la région.

Mots-clés : Champs gaussiens, Champs de χ^2, Ensemble d'excursion, Estimation de gradient, Géostatistique, Puissance.

ABSTRACT

We propose a method for estimating and testing the zones where a variable presents discontinuities or sharp variations in the mean. Such zones are called Zones of Abrupt Change (ZACs). Our method is based on the statistical properties of the predictor of the local gradient of the variable under study. A local test statistic is defined from it and is compared to some critical threshold computed under the null hypothesis of a constant mean. The locations where the null hypothesis is rejected define the potential ZACs. Then, to assess their significance, we aggregate the local tests using geometrical properties of the connected components in the potentials ZACs. The sampling pattern, in particular its local density, is crucial in the power of the local

1. Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, United Kingdom. E-mail : e.gabriel@lancaster.ac.uk

Ce travail constitue une partie du travail de thèse de l'auteur ; thèse réalisée au sein de l'unité de biométrie de l'INRA d'Avignon, sous la direction de D. Allard et M. Guérif et la collaboration de J-N. Bacro (Université Montpellier 2) et qui a obtenu le prix de thèse Marie-Jeanne Duhamel décerné par la SFdS.

test used for detecting ZACs. It is shown that mapping the power allows us to identify zones where ZACs may or may not be detected. The methodology is applied to a soil data set sampled in a small part of the Swiss Jura. Analyzing heavy metal concentrations for ZACs allowed us to identify the main geological structures of the region.

Keywords : Excursion set, Gaussian random fields, Geostatistics, Gradient estimation, Power, χ^2 random fields.

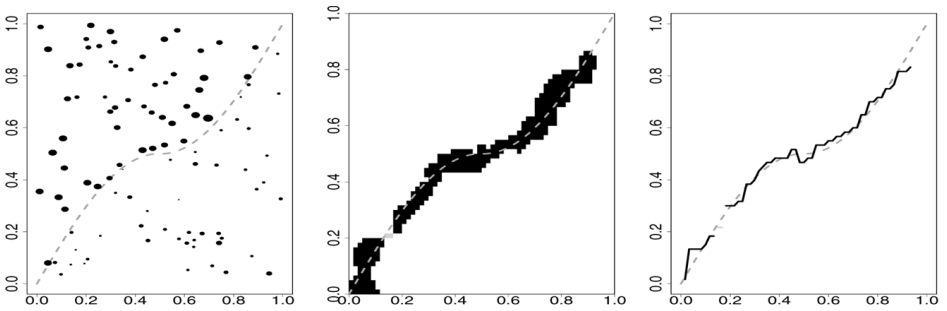
1. Introduction

De nombreuses études en biologie, en sciences de l'environnement ou en sciences du sol requièrent la connaissance des variations spatiales des variables d'étude. Cela se traduit souvent par la nécessité de cartographier les zones où les variables présentent des changements abrupts. Un premier exemple s'inscrit dans le cadre de l'étude des populations en biologie et en écologie (Womble, 1951 ; Barbujani *et al.*, 1989) : les changements abrupts dans les fréquences d'allèles peuvent être associés aux frontières entre différentes populations. Un deuxième exemple, utilisé ici pour illustrer la méthodologie (section 5), est issu des sciences de l'environnement. La variation d'éléments potentiellement toxiques dans le sol, associée à la complexité de la nature du sol, implique des concentrations très hétérogènes. Afin d'expliquer certaines pollutions, il est important de connaître non seulement les sources, mais aussi la distribution spatiale des polluants et donc de déterminer les zones de forte variation de la concentration.

Dans ces exemples, la variable d'étude n'est connue qu'en un nombre limité de points et les méthodes de détection de zones de changement abrupt doivent être basées sur l'estimation du champ aléatoire sous-jacent. Dans ce travail, le modèle général est que, sous l'hypothèse nulle, la variable d'intérêt est une réalisation d'un processus stationnaire d'ordre deux, noté $Z(\cdot)$. L'hypothèse alternative est que l'espérance de $Z(\cdot)$ présente des changements abrupts, ici modélisés par des discontinuités le long d'un ensemble de courbes notées Γ. Sous ce modèle, nous proposons une méthode de détection des lieux où $Z(\cdot)$ varie brusquement. Ces lieux sont appelés Zones de Changement Abrupt (ZCAs). Comme l'échantillon de points est d'assez faible densité, il n'est pas possible d'estimer précisément les courbes Γ, mais plutôt les zones de forte variation.

Avant de présenter de façon détaillée la méthode, nous l'illustrons sur un exemple. Nous avons simulé dans le carré unitaire un échantillon $\mathbf{Z}$ de 100 points répartis aléatoirement (réalisation d'un processus de Strauss), issus d'un champ aléatoire gaussien centré réduit de fonction de covariance exponentielle $C_Z(\mathbf{h}) = \exp(-\|\mathbf{h}\|/b)$, de portée $b = 0.1$. Une discontinuité a été introduite le long d'une courbe sinusoïdale (Figure 1*a* courbe en pointillés) en ajoutant une constante a aux échantillons situés au-dessus de cette courbe. Le cas $a = 0$ correspond à l'absence de discontinuité. La figure 1*a* représente

une réalisation pour $a = 2.5$, la taille des disques étant proportionnelle à la valeur représentée. La figure 1*b* représente les Zones de Changement Abrupt obtenues par notre méthode. Celles-ci se situent le long de la discontinuité. Notons que la même procédure appliquée pour $a = 0$ n'a détecté aucune ZCA. La figure 1*c* représente les lieux de variation maximale dans les ZCAs.



(a) *(b)* *(c)*

FIG 1. — (*a*) 100 points répartis aléatoirement issus d'un champ gaussien centré réduit, une constante $a = 2.5$ a été ajoutée aux échantillons situés au-dessus de la courbe en pointillés. (*b*) Détection des ZCAs. (*c*) Lieux des variations maximales dans les ZCAs.

La méthode repose sur les propriétés locales de l'estimation du gradient du champ $Z(\cdot)$. Le principe est le suivant. Dans un premier temps le gradient est interpolé de façon optimale en utilisant des techniques de la géostatistique. Dans un deuxième temps, nous définissons une statistique de test local, $T(\mathbf{x})$, comme étant la norme standardisée du gradient local. L'ensemble des points pour lesquels $T(\mathbf{x})$ est au-dessus d'un seuil $t_{1-\alpha}$ (que nous déterminons) définit des zones de changement abrupt potentielles. Cet ensemble, noté $\mathcal{A}_{t_{1-\alpha}}$, est appelé ensemble d'excursion du champ $T(\cdot)$ au-dessus de $t_{1-\alpha}$. Enfin dans un troisième temps, afin de tester la significativité globale des ZCAs potentielles, les tests locaux sont agrégés en un test global en utilisant la distribution asymptotique de la surface des composantes connexes de $\mathcal{A}_{t_{1-\alpha}}$ sous l'hypothèse nulle. Nous verrons comment il est possible de déterminer le niveau local α contrôlant le niveau global η. La théorie liant les niveaux α et η dans le plan est basée sur les ensembles d'excursion de champs de χ^2, voir Adler (1981 and 2000), Aronowich and Adler (1988), Worsley (1994) et Cao (1999) pour le cas de champs de χ^2 standards (somme du carré de deux champs gaussiens indépendants et identiquement distribués). La variable étant irrégulièrement échantillonnée, le gradient estimé est non stationnaire et nous verrons à la section 3 que la statistique $T(\mathbf{x})$ est un champ de χ^2 non standard. Aussi nous ne pouvons pas utiliser les résultats existants sur les champs de χ^2 standards et développons de nouveaux résultats théoriques sur les champs de χ^2 non stationnaires afin de déterminer la distribution asymptotique de la surface d'une composante connexe lorsque $t_{1-\alpha} \to \infty$.

La qualité de l'estimation du gradient étant liée à la densité d'échantillonnage et la variable pouvant être irrégulièrement échantillonnée, nous devons regarder s'il est pertinent d'appliquer la méthode sur l'ensemble du domaine ou s'il existe des zones sur lesquelles elle ne peut être appliquée. Une réponse à ce problème repose sur le calcul de la puissance du test local de détection de changements abrupts, *i.e.* la probabilité de rejeter l'hypothèse nulle de stationnarité de la variable, sous l'hypothèse alternative d'existence de discontinuités. La puissance est ainsi calculée en chaque point du domaine d'étude. Sa cartographie permet ensuite d'identifier les zones où l'échantillonnage n'est pas approprié à l'estimation des ZCAs.

La deuxième section de ce papier est consacrée à l'émergence du problème de détection de ZCAs dans différents contextes, aux méthodes qui en ont découlé et à la présentation des notions nécessaires sur les champs aléatoires et la géostatistique. La méthode et le calcul de la puissance du test de détection sont ensuite présentés dans la troisième section. Dans la quatrième section, nous verrons comment nous avons pu résoudre les problèmes liés à l'implémentation de la méthode : approximation du domaine par une grille, détermination du niveau local α et estimation de la fonction de covariance du champ $Z(\cdot)$, mais aussi comment le schéma d'échantillonnage affecte la détection de ZCAs. L'application de la méthode à des données de sciences de l'environnement est présentée dans la cinquième section. L'analyse des concentrations de métaux lourds dans une petite région du Jura suisse montre que des mêmes ZCAs sont détectées pour différentes variables et qu'elles sont fortement liées aux frontières entre différentes zones géologiques.

2. Contexte

2.1. Pourquoi et comment détecter des ZCAs ?

L'estimation de courbes de discontinuités à partir d'un échantillon de points aléatoirement répartis dans le plan est un problème qui a été peu abordé en dépit de son importance concrète évidente. Le problème a été introduit par Womble en 1951 lors de l'étude des populations en biologie et en écologie. Dans ce contexte les changements abrupts dans des fréquences de gènes peuvent être liés aux frontières entre différentes populations. Par exemple des études sur la diversité génétique montrent que les différences de langues réduisent les échanges génétiques entre populations (Pagel and Mace, 2004). La procédure proposée par Womble (1951), connue sous le nom de wombling, est assez simple. La variable d'étude est interpolée linéairement sur une grille régulière. Un vecteur de gradient est ensuite calculé par différences. Le wombling et ses extensions (Barbujani *et al.*, 1989 ; Oden *et al.*, 1993 ; Bocquet-Apple and Bacro, 1994 ; Fortin, 1994) définissent les barrières comme étant les 5% ou 10% des valeurs du gradient les plus élevées. Une synthèse des méthodes basées sur le wombling est proposée dans Jacquez *et al.* (2000).

Hall and Rau (2001) ont développé une méthode basée sur les approximations spatiales de la vraisemblance locale qu'une courbe de discontinuité passe par un point donné, comme fonction de ce point. Il s'agit d'une méthode de

tracking, *i.e.* elle avance pas à pas le long de la discontinuité en prenant, à chaque pas, la direction qui maximise la vraisemblance locale. Cette méthode présente plusieurs inconvénients. Etant basée sur une méthode de tracking, elle suppose implicitement l'existence d'une courbe de discontinuité et elle nécessite un point de départ situé près de la courbe de discontinuité, de préférence le long du bord d'un domaine rectangulaire de sorte que la discontinuité l'intersecte.

Plus récemment, Banerjee *et al.* (2003) ont proposé une approche bayésienne d'estimation du gradient d'un processus spatial irrégulièrement échantillonné. Cependant cette approche n'est pas spécifiquement vouée à détecter des barrières ou des ZCAs. En particulier, elle ne permet pas de tester globalement la présence de ZCAs dans le domaine d'étude.

Le point commun, et principal inconvénient, à toutes ces méthodes est qu'elles ne testent pas la significativité des barrières ou ZCAs détectées, *i.e.* elles ne permettent pas de décider si celles-ci sont dues au hasard ou si elles traduisent un réel effet. Dans le wombling, l'utilisation de seuils fixes pour identifier des barrières entraîne une détection systématique, que leur existence soit significative ou non. Des tests de permutation ont été proposés afin de tester la significativité des barrières (Fortin and Drapeau, 1995 ; Jacquez and Maruca, 1998 ; Gleyze *et al.*, 2001). Ces permutations déstructurent les données et ne peuvent donc s'appliquer que dans le cas où la variable est spatialement non corrélée. Cette hypothèse étant le plus souvent irréaliste en sciences de l'environnement, l'approche par tests de permutation est d'un intérêt limité.

La méthode de détection de ZCAs proposée dans la section 3 répond à ces problèmes ; en particulier elle permet de tester la significativité des ZCAs détectées. Cette méthode se situant dans le cadre des champs aléatoires (Adler, 1981 ; Yaglom, 1986) et de la géostatistique (Cressie, 1993 ; Chilès et Delfiner, 1999), les sous-sections suivantes rappellent quelques résultats théoriques sur lesquels elle repose.

2.2 Rappels sur la théorie des champs aléatoires

Soient $\mathcal{D}$ un sous-ensemble de $\mathbb{R}^2$ et $(\Omega, \mathcal{B}, P)$ un espace de probabilité. Un champ aléatoire est une fonction de deux variables $Z(\mathbf{x}, \omega)$, $\omega \in \Omega$, telle que pour tout $\mathbf{x} \in \mathcal{D}$, $Z(\mathbf{x}, \cdot) = Z(\mathbf{x})$ est une variable aléatoire sur $(\Omega, \mathcal{B}, P)$. Chaque fonction $Z(\cdot, \omega)$ est une réalisation de la variable aléatoire. Pour simplifier les notations, le champ aléatoire sera dans la suite noté $Z(\cdot)$.

Afin de détecter des zones de changement abrupt, nous supposons qu'en l'absence de ZCA, le champ aléatoire $Z(\cdot)$ est un champ gaussien stationnaire d'ordre 2.

Stationnarité

Un champ aléatoire est dit *stationnaire d'ordre 2* (ou faiblement stationnaire) s'il satisfait les conditions :

$$(i) \quad \mathbb{E}\left[Z(\mathbf{x})\right] = m,$$
$$(ii) \quad \mathrm{Cov}\big(Z(\mathbf{x}), Z(\mathbf{x}+\mathbf{h})\big) = C_Z(\mathbf{h}), \forall \mathbf{h} \in \mathbb{R}^2.$$

L'*hypothèse intrinsèque* suppose que, pour tout $\mathbf{h}$, les accroissements $Z(\mathbf{x}+\mathbf{h})-Z(\mathbf{x})$ sont de moyenne nulle et que leur variance ne dépend que du vecteur $\mathbf{h}$:

$$\mathbb{E}\left[Z(\mathbf{x}+\mathbf{h})-Z(\mathbf{x})\right]=\mathbf{0} \text{ et } \mathrm{Var}\big(Z(\mathbf{x}+\mathbf{h})-Z(\mathbf{x})\big)=2\gamma(\mathbf{h}).$$

La fonction $\gamma(\mathbf{h})$ est appelée variogramme. L'estimateur classique du variogramme proposé par Matheron (1962) est :

$$\hat{\gamma}(\mathbf{h})=\frac{1}{2|N(\mathbf{h})|}\sum_{N(\mathbf{h})}\big(Z(\mathbf{x}_i)-Z(\mathbf{x}_j)\big)^2, \tag{1}$$

où la somme est prise sur $N(\mathbf{h})=\{(i,j):\mathbf{x}_i-\mathbf{x}_j\simeq\mathbf{h}\}$ et $|N(\mathbf{h})|$ est le nombre d'éléments distincts de $N(\mathbf{h})$.

Un champ aléatoire stationnaire d'ordre 2 est un champ aléatoire intrinsèque et par conséquent possède un variogramme. Dans ce cas variogramme et covariance sont liés par la relation :

$$\gamma(\mathbf{h})=C_Z(\mathbf{0})-C_Z(\mathbf{h}). \tag{2}$$

Continuité et différentiabilité

Rappelons que pour $(\mathbf{x}_n)_n$ une suite de points et $\mathbf{x}=(x^1,x^2)$ fixé dans $\mathbb{R}^2$ tel que $\|\mathbf{x}_n-\mathbf{x}\|\to 0$ quand $n\to\infty$, si $\mathbb{E}\left[\,\|Z(\mathbf{x}_n)-Z(\mathbf{x})\|^2\,\right]\to 0$ quand $n\to\infty$, alors $Z(\cdot)$ est *continu en* $\mathbf{x}$ *en moyenne quadratique*.

Théorème 1 (Adler, 1981). — *Un champ aléatoire $Z(\cdot)$ est continu en moyenne quadratique en $\mathbf{x}^*$, si et seulement si sa fonction de covariance $(\mathbf{x},\mathbf{y})\mapsto C_Z(\mathbf{x},\mathbf{y})$ est continue pour $\mathbf{x}=\mathbf{y}=\mathbf{x}^*$. Si $C_Z(\cdot,\cdot)$ est continue sur la diagonale, alors elle est continue partout.*

Théorème 2 (Adler, 1981). — *Si $\partial^2 C_Z(\mathbf{x},\mathbf{y})/\partial x^i\partial y^i$ existe et est finie au point $(\mathbf{x},\mathbf{x})\in\mathbb{R}^4$, alors :*

$$\frac{\partial Z(\mathbf{x})}{\partial x^i}=\lim_{\epsilon\to 0}\frac{Z(\mathbf{x}+\epsilon\delta_i)-Z(\mathbf{x})}{\epsilon} \tag{3}$$

existe et est appelée dérivée en moyenne quadratique de $Z(\cdot)$ en $\mathbf{x}$, où δ_i désigne le $i$ème vecteur unitaire de $\mathbb{R}^2$. Si cette dérivée existe $\forall\mathbf{x}\in\mathbb{R}^2$, alors $Z(\cdot)$ possède une dérivée en moyenne quadratique. La fonction de covariance de $\partial Z(\mathbf{x})/\partial x^i$ est $\partial^2 C_Z(\mathbf{x},\mathbf{y})/\partial x^i\partial y^i$.

Lorsque le champ aléatoire est stationnaire d'ordre 2, les conditions assurant la continuité en moyenne quadratique et l'existence de dérivées en moyenne quadratique se simplifient : si $C_Z(\cdot)$ est continue à l'origine alors le champ $Z(\cdot)$ est continu en moyenne quadratique et si $C_Z(.)$ est dérivable deux fois à l'origine, alors $Z(\cdot)$ est dérivable en moyenne quadratique en tout point.

Champs gaussiens et champs de χ^2

Un champ gaussien est un champ aléatoire tel que pour tout échantillon $\mathbf{x}_1, \ldots, \mathbf{x}_n$ de points de $\mathbb{R}^2$ la distribution jointe de $\big(Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_n)\big)'$ est une gaussienne multivariée. Il est entièrement spécifié par sa fonction moyenne $m(\mathbf{x}) = \mathbb{E}\,[Z(\mathbf{x})]$ et sa fonction de covariance $C_Z(\mathbf{x}, \mathbf{y}) = \mathbb{E}\left[\big(Z(\mathbf{x}) - m(\mathbf{x})\big)\big(Z(\mathbf{y}) - m(\mathbf{y})\big)\right]$. Si un champ gaussien est de moyenne constante et de fonction de covariance dépendant seulement de $\mathbf{x} - \mathbf{y}$, alors le champ est stationnaire d'ordre 2.

PROPOSITION 1. — *Si $Z(\cdot)$ est un champ aléatoire gaussien possédant une dérivée en moyenne quadratique, alors $\partial Z(\cdot)/\partial x^i$ est aussi un champ aléatoire gaussien.*

Cela vient directement du théorème 2 et du fait qu'un vecteur aléatoire est gaussien si et seulement si toute combinaison linéaire de ses composantes est une variable aléatoire gaussienne.

Pour construire un champ de χ^2 à p degrés de liberté, Adler (1981) considère p champs gaussiens, $Z_1(\mathbf{x}), \ldots, Z_p(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^2$, indépendants et stationnaires. Supposons que chaque $Z_i(\cdot)$ soit de moyenne nulle et que tous les $Z_i(\cdot)$ aient la même fonction de covariance, $C_Z(\mathbf{h})$, telle que $C_Z(\mathbf{0}) = 1$. Alors, pour tout $\mathbf{x} \in \mathbb{R}^2$, le champ $Y(\mathbf{x}) = [Z_1(\mathbf{x})]^2 + \cdots + [Z_p(\mathbf{x})]^2$ définit un champ de χ^2 à p degrés de liberté, avec $\mathbb{E}\,[Y(\mathbf{x})] = p$ et $\mathrm{Var}(Y(\mathbf{x})) = 2p$. Comme les $Z_i(\cdot)$, $i = 1, ..., p$ sont stationnaires, il en va de même pour le champ $Y(\cdot)$.

PROPOSITION 2 (Adler, 1981). — *Soit $C_Z(\mathbf{x}, \mathbf{y})$ la fonction de covariance commune des $Z_i(\cdot)$. La fonction de covariance, $C_Y(\mathbf{x}, \mathbf{y})$, du champ $\chi^2(p)$, $Y(\cdot)$, défini précédemment est : $C_Y(\mathbf{x}, \mathbf{y}) = 2p\,[C_Z(\mathbf{x}, \mathbf{y})]^2$.*

La distribution de $Y(\cdot)$, ainsi que toutes ses propriétés statistiques, sont entièrement déterminées lorsque p et $C_Y(\cdot)$ (ou de façon équivalente $C_Z(\cdot)$) sont connus.

2.3. Le krigeage

Soit $\mathbf{Z} = \big(Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_n)\big)'$ un échantillon de valeurs réelles d'un champ gaussien $Z(\cdot)$ aux points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ du domaine $\mathcal{D}$. Lorsque $Z(\cdot)$ est stationnaire d'ordre 2, de fonction de covariance connue et de moyenne inconnue, le krigeage ordinaire est un prédicteur de la valeur du champ en un point $\mathbf{x}$ utilisant les valeurs $\mathbf{Z}$. Celui-ci est défini comme le meilleur prédicteur linéaire sans biais au sens des moindres carrés (ou BLUP, *Best Linear Unbiased Predictor*), et s'écrit :

$$Z^*(\mathbf{x}) = C'(\mathbf{x})\mathbf{C}^{-1}\mathbf{Z} + \big(1 - C'(\mathbf{x})\mathbf{C}^{-1}\mathbf{1}\big)\frac{\mathbf{1}'\mathbf{C}^{-1}}{\mathbf{1}'\mathbf{C}^{-1}\mathbf{1}}\mathbf{Z}, \tag{4}$$

où $\mathbf{1}$ est le vecteur de longueur n d'éléments 1, $C(\mathbf{x})$ est le vecteur de covariance entre $Z(\mathbf{x})$ et les $Z(\mathbf{x}_i)$, $i = 1, \ldots, n$, et $\mathbf{C}$ est la matrice de covariance entre les $Z(\mathbf{x}_i)$. Notons que le champ $Z^*(\cdot)$ est non-stationnaire.

Supposons maintenant que le champ aléatoire vérifie les conditions de la stationnarité intrinsèque et non plus celle de la stationnarité d'ordre 2. Cela signifie que l'incrément $Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})$ est d'espérance nulle et que $\mathrm{Var}\left(Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})\right) = 2\gamma(\mathbf{h})$. Ces hypothèses ne permettent pas de travailler directement sur $Z(\cdot)$ mais seulement sur des accroissements de $Z(\cdot)$. On peut montrer que le prédicteur de krigeage intrinsèque a la même forme que celui du krigeage ordinaire, mais dans lequel on remplace terme à terme la fonction de covariance C par le variogramme γ (Chilès et Delfiner, 1999).

3. Détection de Zones de Changement Abrupt

Nous rappelons le modèle général sous lequel nous nous plaçons. Sous l'hypothèse nulle le champ aléatoire $Z(\cdot)$ est d'espérance constante sur le domaine $\mathcal{D}$. L'alternative est que $\mathbb{E}\left[Z(\cdot)\right]$ présente des discontinuités le long d'un ensemble de courbes notées Γ. Aucune hypothèse n'est faite sur la forme de ces courbes. A ce stade il n'est pas réellement nécessaire de considérer que l'espérance présente des discontinuités ; il suffirait de considérer des variations localement fortes. Ce modèle nous sera toutefois utile lorsque nous aborderons la question de la puissance du test à la section 3.4. Nous faisons les hypothèses suivantes :

$\mathcal{H}_1$: la fonction de covariance de $Z(\cdot)$ est stationnaire :
$C_Z(\mathbf{x},\mathbf{y}) = C_Z(\mathbf{x}-\mathbf{y}), \forall \mathbf{x},\mathbf{y} \in \mathcal{D}$,

$\mathcal{H}_2$: $C_Z(\mathbf{h})$ est indéfiniment différentiable pour tout $\mathbf{h}$ tel que $\|\mathbf{h}\| > 0$.

L'hypothèse $\mathcal{H}_1$ est usuelle en géostatistique. Elle est nécessaire pour l'estimation de la fonction de covariance lorsqu'il n'y a pas de répétition des données et que la densité d'échantillonnage n'est pas très élevée. Sous l'hypothèse nulle, $\mathcal{H}_1$ entraîne que le prédicteur optimal de $Z(\mathbf{x})$ en un point non échantillonné $\mathbf{x}$ est le krigeage ordinaire (Equation 4). L'hypothèse $\mathcal{H}_2$ est moins usuelle car elle porte sur la régularité de la fonction de covariance en dehors de $\mathbf{0}$. Elle est vérifiée par un grand nombre de fonctions de covariance. L'hypothèse $\mathcal{H}_2$ entraîne le résultat suivant :

Proposition 3. — *Sous les hypothèses $\mathcal{H}_1$ et $\mathcal{H}_2$, le champ $Z^*(\cdot)$ est indéfiniment différentiable presque sûrement et en moyenne quadratique pour tout $\mathbf{x} \in \mathcal{D}$, sauf éventuellement aux points d'échantillonnage.*

En effet, d'après la relation (4), le champ aléatoire $Z^*(\cdot)$ hérite des propriétés de différentiabilité de la fonction $C(\mathbf{x})$, *i.e.* de $C_Z(\mathbf{x}-\mathbf{x}_i), \forall i = 1,\ldots,n$. Ainsi, si la fonction de covariance est supposée indéfiniment différentiable pour $\|\mathbf{h}\| > 0$, $Z^*(\cdot)$ sera indéfiniment différentiable pour tout $\mathbf{x}$, sauf peut-être aux points d'échantillonnage. Cela implique que $Z^*(\cdot)$ est presque partout indéfiniment différentiable. Si $Z(\cdot)$ est un champ gaussien de variance finie, il en découle l'indéfinie différentiabilité en moyenne quadratique. Le champ $Z^*(\cdot)$ est continu aux points d'échantillonnage si $C_Z(\mathbf{h})$ est continue à l'origine. De

façon plus générale, $Z^*(\cdot)$ est k-fois différentiable aux points d'échantillonnage si $C_Z(\mathbf{h})$ est $2k$-fois différentiable à l'origine. Si $C_Z(\mathbf{h})$ possède un effet de pépite, *i.e.* si $C_Z(\mathbf{h})$ est discontinue en $\mathbf{h} = \mathbf{0}$, alors $Z^*(\cdot)$ est discontinu aux points d'échantillonnage.

Nous verrons dans la section 6 comment ces hypothèses peuvent être affaiblies. Remarquons qu'aucune hypothèse n'est exigée en $\mathbf{h} = \mathbf{0}$. En particulier la continuité et la différentiabilité à l'origine ne sont pas requises. En effet, la discontinuité à l'origine de la fonction de covariance n'entraîne la discontinuité du champ interpolé qu'aux points d'échantillonnage, qui constituent un ensemble de mesure nulle dans $\mathcal{D}$.

3.1. Estimation du gradient

D'après le théorème 2, la condition mathématique de différentiabilité d'un champ aléatoire est la double différentiabilité à l'origine de la fonction de covariance du champ. Les modèles de covariance les plus utilisés en pratique (modèles sphérique et exponentiel) ne sont pas différentiables à l'origine. Les dérivées du champ ne sont par conséquent pas définies. Par contre, nous pouvons définir un taux de variation. Soit $\delta_i,\ i = 1, 2$ les vecteurs unitaires de $\mathbb{R}^2$. Par linéarité du système de krigeage, le prédicteur optimal du taux de variation $\big(Z(\mathbf{x} + \epsilon\delta_i) - Z(\mathbf{x})\big)/\epsilon$ est $\big(Z^*(\mathbf{x} + \epsilon\delta_i) - Z^*(\mathbf{x})\big)/\epsilon$. Le passage à la limite $\epsilon \to 0$ est autorisé pour le champ $Z^*(\cdot)$. La quantité $\partial Z^*(\cdot)/\partial x^i$ existe et nous la considérons comme une estimation de la variation locale du champ $Z(\cdot)$ (Chilès et Delfiner, 1999).

D'après la proposition 3, le gradient $W(\cdot)$ de $Z^*(\cdot)$ existe : $W(\mathbf{x}) = \big(W_1(\mathbf{x}), W_2(\mathbf{x})\big)' = \big(\partial_1 Z^*(\mathbf{x}), \partial_2 Z^*(\mathbf{x})\big)'$, avec

$$\partial_i Z^*(\mathbf{x}) = \partial_i C(\mathbf{x})' \left(\mathbf{C}^{-1} + \frac{\mathbf{C}^{-1}\mathbf{1}\mathbf{1}'\mathbf{C}^{-1}}{\mathbf{1}'\mathbf{C}^{-1}\mathbf{1}} \right) \mathbf{Z} = \partial_i C(\mathbf{x})' \mathbf{K}^{-1} \mathbf{Z},$$

où ∂_i dénote la dérivée partielle le long de la $i$ème coordonnée, $i \in \{1, 2\}$. Banerjee *et al.* (2003) proposent une théorie complète sur la distribution de champ de gradient pour des processus gaussiens stationnaires.

3.2. Test local de détection

Nous allons dans un premier temps tester la présence d'une discontinuité en un point $\mathbf{x} \in \mathcal{D}$. Pour cela nous définissons un test local entre les hypothèses $H_0(\mathbf{x})$: $\mathbb{E}\left[Z(\mathbf{y})\right] = m$, pour tout $\mathbf{y}$ dans un petit voisinage de $\mathbf{x}$ et $H_1(\mathbf{x})$: $\mathbf{x} \in \Gamma$, où Γ est la courbe de discontinuité.

Soit $\mathbf{\Sigma}(\mathbf{x})$ la matrice de covariance du champ interpolé $W(\mathbf{x})$, $\mathbf{\Sigma}(\mathbf{x}) = \mathbb{E}\left[W(\mathbf{x})W'(\mathbf{x})\right] = \partial C(\mathbf{x})' \mathbf{K}^{-1} \partial C(\mathbf{x})$, où $\partial C(\mathbf{x})$ est la $n \times 2$ matrice $\big(\partial_1 C(\mathbf{x}), \partial_2 C(\mathbf{x})\big)$. Comme $Z(\mathbf{x})$ est un champ gaussien, $W(\mathbf{x})$ est un vecteur gaussien et $\mathbf{\Sigma}(x)^{-1}$ est sa matrice de précision. Nous définissons la statistique de test $T(\mathbf{x})$ par :

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

Sous l'hypothèse locale nulle, $T(\mathbf{x})$ a une distribution de χ^2 à deux degrés de liberté. L'hypothèse locale nulle sera rejetée si le gradient local de la surface estimée est supérieur à une certaine valeur critique, *i.e.* lorsque $T(\mathbf{x})$ sera supérieure au quantile d'ordre $1-\alpha$ de la distribution $\chi^2(2)$, que nous noterons $t_{1-\alpha}$. Nous définissons les ZCAs potentielles comme étant l'ensemble d'excursion de la statistique de test local au dessus du niveau $t_{1-\alpha}$: $\mathcal{A}_{t_{1-\alpha}} = \{\mathbf{x} : T(\mathbf{x}) > t_{1-\alpha}\}$.

Notre prochaine étape consiste à proposer un test global afin de rejeter H_0 contre H_1. Pour cela nous allons suivre une approche similaire à celle usuellement utilisée en fMRI (*fonctional Magnetic Resonance Imaging*) et proposée dans Worsley (2001) et Cao (1999). Cette approche nous permet d'obtenir la distribution asymptotique de la surface des composantes connexes des ZCAs afin d'en tester la significativité.

3.3. Test global de significativité

Afin d'agréger les tests locaux pour déterminer si $\mathbf{Z}$ est issu d'un champ aléatoire stationnaire (hypothèse nulle) ou d'un champ présentant des courbes de discontinuité (hypothèse alternative), nous travaillons sur les composantes connexes des ZCAs potentielles pour lesquelles nous pouvons établir un test. Le test global consiste à déterminer la significativité de chaque composante connexe de $\mathcal{A}_{t_{1-\alpha}}$, indépendamment les unes des autres. Cela revient à tester simultanément $H_0(\mathbf{x}_{\mathcal{C}_j})$ versus $H_1(\mathbf{x}_{\mathcal{C}_j})$ pour tous les $\mathbf{x}_{\mathcal{C}_j}$ appartenant à la $j$ème composante connexe. Notons que, sous l'hypothèse nulle, lorsque le seuil $t_{1-\alpha}$ est élevé il y a, avec une très forte probabilité, au plus une composante connexe sur $\mathcal{D}$. Ainsi tester sur chaque composante connexe est équivalent à tester sur le domaine entier.

Notons $\mathcal{C}_{t_{1-\alpha}}$ une composante connexe de $\mathcal{A}_{t_{1-\alpha}}$. Afin de tester la significativité de $\mathcal{C}_{t_{1-\alpha}}$, nous prenons comme statistique de test sa surface, $S_{t_{1-\alpha}}$. Dans le cadre de la théorie des ensembles d'excursion de champs aléatoires (Adler, 2000 ; Cao, 1999), nous avons établi le théorème suivant sur la distribution asymptotique de $S_{t_{1-\alpha}}$ pour un seuil élevé.

Théorème 3. — *Sous l'hypothèse nulle, conditionnellement à l'événement "$T(\mathbf{x}_0)$ est un maximum en $\mathbf{x}_0$ de hauteur supérieure à $t_{1-\alpha}$",*

$$t_{1-\alpha} S_{t_{1-\alpha}} \xrightarrow{\mathcal{L}} \pi \det\left(\mathbf{\Lambda}(\mathbf{x}_0)\right)^{-1/2} E(2), \text{ quand } t_{1-\alpha} \to \infty, \tag{5}$$

où $E(2)$ est une variable aléatoire exponentielle d'espérance 2 indépendante de $T(\cdot)$ et $\mathbf{\Lambda}(\mathbf{x}_0)$ est une matrice associée à la courbure du champ $T(\cdot)$ autour de son maximum dans $\mathcal{C}_{t_{1-\alpha}}$.

Eléments de preuve : (Le lecteur peut se référer à Gabriel (2004) et Allard *et al.* (2005) pour une démonstration complète de ce théorème et des résultats intermédiaires permettant d'y aboutir.) Comme nous nous intéressons à la structure du champ $T(\cdot)$ autour d'un maximum local, nous utilisons la notion de champ conditionnel. Dans ce qui suit, nous allons adopter les notations suivantes :

- Le champ $T(\cdot)$ est un champ de χ^2 à deux degrés de liberté. Il se décompose en la somme du carré de deux champs gaussiens $U_1(\cdot)$ et $U_2(\cdot)$, dont les expressions sont données en annexe : $T(\mathbf{x}) = U_1(\mathbf{x}) + U_2(\mathbf{x})$.
- $T_{T_0}(\cdot)$ est le champ $T(\cdot)$ conditionnel à l'événement $\mathcal{E}_0$: "$T(\cdot)$ possède un maximum local en $\mathbf{x}_0$ de hauteur T_0".
- $\mathbf{\Lambda}_i(\mathbf{x})$ est la variance des dérivées premières du champ $U_i(\mathbf{x})$. Les expressions des matrices $\mathbf{\Lambda}_i(\mathbf{x})$ ne dépendent que de la fonction de covariance du champ $Z(\cdot)$ et du schéma d'échantillonnage. Celles-ci sont données de façon explicite en annexe.

Comme nous considérons le gradient du champ interpolé, les champs $U_1(\cdot)$ et $U_2(\cdot)$ sont non stationnaires et non indépendants. Aussi nous n'avons pas pu utiliser les résultats existants sur les champs de χ^2 standards (Worsley, 1994 ; Cao, 1999). La distribution asymptotique de la surface d'une composante connexe de $\mathcal{A}_{t_{1-\alpha}}$ s'obtient en trois étapes :

1. Dans un premier temps nous avons pu montrer que conditionnellement à l'événement $\mathcal{E}_0$, $T(\mathbf{x})$ est asymptotiquement approché par un paraboloïde elliptique dans le voisinage de $\mathbf{x}_0$:

$$T_{T_0}(\mathbf{x}) = T_0\big(1 - \mathbf{x}'\mathbf{\Lambda}(\mathbf{x}_0)\mathbf{x}\big) + o_p(1) \quad \text{quand} \quad T_0 \to \infty,$$

où $\mathbf{\Lambda}(\mathbf{x}_0) = v_0\mathbf{\Lambda}_1(\mathbf{x}_0) + (1 - v_0)\mathbf{\Lambda}_2(\mathbf{x}_0)$, avec $v_0 = U_1^2(\mathbf{x}_0)/T_0$.

La surface $S_{t_{1-\alpha}}$ de la composante connexe contenant $\mathbf{x}_0$ est la section horizontale du paraboloïde elliptique à la distance $T_0 - t_{1-\alpha}$ à partir du maximum.

2. Conditionnellement à $\mathcal{E}_0$, la surface $S_{t_{1-\alpha}}$ et la distance $T_0 - t_{1-\alpha}$ sont asymptotiquement reliées par :

$$S_{t_{1-\alpha}} = \frac{T_0 - t_{1-\alpha}}{T_0}\pi\left(\mathbf{\Lambda}(\mathbf{x}_0)\right)^{-1/2} + o_p\left(\frac{1}{T_0}\right), \text{ quand } t_{1-\alpha} \to \infty.$$

En effet, $S_{t_{1-\alpha}}$ correspond à la mesure de Lebesgue de la composante connexe contenant $\mathbf{x}_0$:

$$\begin{aligned} S_{t_{1-\alpha}} &= \int \mathbf{1}_{\left\{\mathbf{x}\,:\,T_{T_0}(\mathbf{x}) > t_{1-\alpha}\right\}}(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_{T_0\left(1-\mathbf{x}'\mathbf{\Lambda}(\mathbf{x}_0)\mathbf{x}\right)+o_p(1)>t_{1-\alpha}} \mathrm{d}\mathbf{x}, \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{d'après le point 1} \\ &= \det\left(\mathbf{\Lambda}(\mathbf{x}_0)\right)^{-1/2} \int_{\frac{T_0 - t_{1-\alpha}}{T_0}+o_p\left(\frac{1}{T_0}\right)>\mathbf{y}'\mathbf{y}} \mathrm{d}\mathbf{y}. \end{aligned}$$

3. Enfin, on peut montrer que conditionnellement à $\mathcal{E}_0$ et $T_0 > t_{1-\alpha}$:

$$T_0 - t_{1-\alpha} = E(2) + o_p(1), \text{ quand } t_{1-\alpha} \to \infty,$$

où $E(2)$ est une variable aléatoire exponentielle d'espérance 2.

Ceci est dû au fait que les maxima du champ $T(\cdot)$ peuvent être considérés comme un processus ponctuel localement fini et qu'un champ de χ^2 à deux degrés de liberté est un champ exponentiel d'espérance 2. Aussi nous pouvons écrire $\mathbb{P}\left[T(\mathbf{x}) > t_{1-\alpha}\right] = e^{-t_{1-\alpha}/2} = \frac{1}{|\mathcal{D}|}\int_{t_{1-\alpha}}^{+\infty} \lambda(z) S_{t_{1-\alpha}}\,\mathrm{d}z$, où $\lambda(z)\,\mathrm{d}z$ représente l'espérance du nombre de maxima locaux et $|\mathcal{D}|$ la surface du domaine $\mathcal{D}$. Le point 2 permet alors d'en déduire l'expression de $\lambda(z)$: $\lambda(z) = \frac{1}{4\pi}|\mathcal{D}| \det\left(\mathbf{\Lambda}(\mathbf{x}_0)\right)^{1/2} z e^{-z/2}$ et il est possible d'établir la loi des excès conditionnelle à l'existence d'un maximum :

$$\mathbb{P}\left[T_0 > t_{1-\alpha} + u \mid T_0 > t_{1-\alpha}\right] = \frac{\displaystyle\int_{t_{1-\alpha}+u}^{+\infty} \lambda(z)\,\mathrm{d}z}{\displaystyle\int_{t_{1-\alpha}}^{+\infty} \lambda(z)\,\mathrm{d}z} = \exp(-u/2).$$

Finalement, lorsque $t_{1-\alpha} \to \infty$, les points 2 et 3 permettent d'obtenir la distribution asymptotique de la surface $S_{t_{1-\alpha}}$. □

Ce théorème de convergence permet de calculer une p-valeur asymptotique associée à chaque composante connexe :

$$p = \exp\left(-\frac{1}{2\pi} t_{1-\alpha} S_{t_{1-\alpha}} \det\left(\mathbf{\Lambda}(\mathbf{x}_0)\right)^{1/2}\right). \tag{6}$$

Cette p-valeur est ensuite comparée à un niveau de référence η, par exemple 5%. Si p est inférieure à η, la composante connexe est jugée significative. Nous définissons alors les ZCAs comme étant l'ensemble des ZCAs potentielles significatives.

3.4. Puissance du test local

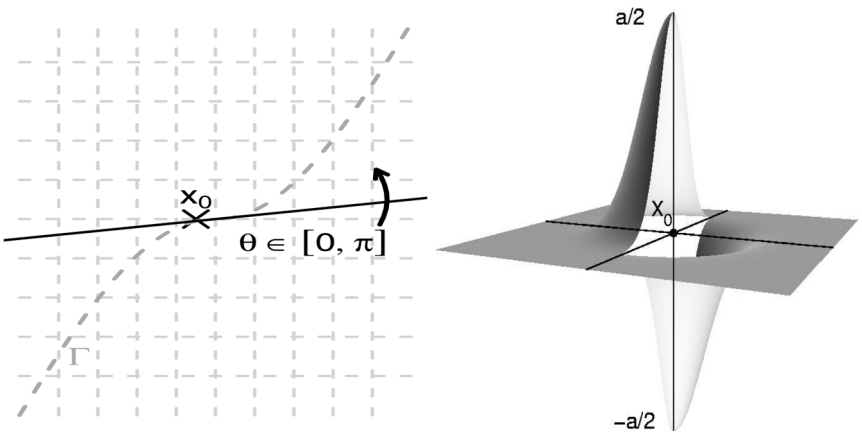
Rappelons que la méthode est basée sur l'étude du gradient local. Celui-ci devant être bien estimé, il est nécessaire d'avoir un échantillon suffisamment dense en tout point. En effet la densité locale d'échantillonnage a une conséquence directe sur la puissance de la méthode : il est difficile de détecter des changements abrupts lorsque l'échantillon est trop clairsemé. La rareté des données ne permet pas de tester si le gradient local prédit correspond à une transition régulière ou localement forte. Cela mène à nous poser la question de la puissance du test. Cependant considérer la puissance globale en fonction de la densité de points d'échantillonnage n'est pas pleinement satisfaisant car cela fournit la probabilité qu'une discontinuité soit détectée sachant son existence, mais cela ne nous indique pas si toutes les zones du domaine sont suffisamment échantillonnées. Aussi nous calculons la puissance du test de détection en chaque point du domaine $\mathcal{D}$. Le calcul de cette puissance est complexe d'une part parce qu'il faut se placer sous l'hypothèse alternative d'existence d'une discontinuité alors que l'on ne connaît pas la forme de cette discontinuité, et d'autre part parce que les tests locaux ne sont pas indépendants.

Spécification de l'hypothèse alternative

La puissance du test local au point $\mathbf{x}_0$ est la probabilité de rejeter l'hypothèse nulle $H_0(\mathbf{x}_0)$, lorsque l'hypothèse alternative $H_1(\mathbf{x}_0) : \mathbf{x}_0 \in \Gamma$ est vérifiée. Nous avons vu que la méthode laisse libre la forme des ZCAs. Afin de calculer une puissance du test de détection, nous allons nous donner un modèle de discontinuité. Nous supposons que la discontinuité est une perturbation du champ moyen, notée $f(\mathbf{x})$. Nous supposons également que $f(\mathbf{x})$ est lisse, sauf le long de la discontinuité, que $\int_{\mathbb{R}^2} f(\mathbf{x})\,\mathrm{d}\mathbf{x} = 0$ et que $f(\mathbf{x})$ tend vers 0 lorsque l'on s'éloigne de la discontinuité. Sous cette hypothèse, nous pouvons approcher une large classe de courbes de discontinuité par leur tangente au point $\mathbf{x}_0$ où la puissance est calculée (Figure 2*a*). La densité gaussienne signée, aussi appelée "boutonnière", est un bon candidat pour une telle perturbation (Figure 2*b*) :

$$f(\mathbf{x}) = a\sqrt{\frac{\pi}{2}}\frac{L}{4}\mathrm{sign}\left(\sin\left(\phi(\mathbf{x}) - \theta\right)\right) g\big(4||\mathbf{x} - \mathbf{x}_0||/L\big), \tag{7}$$

où L est la "longueur" de la discontinuité, sign est la fonction signe, $\phi(\mathbf{x})$ représente l'angle entre l'axe des abscisses et la droite $(\mathbf{x}, \mathbf{x}_0)$, θ donne l'angle de la tangente en $\mathbf{x}_0$ et l'axe des abscisses et g est la densité gaussienne standard. La perturbation $f(\mathbf{x})$ ainsi définie est telle que de part et d'autre de la tangente, le champ ait pour moyenne $\pm a/2$, *i.e.* le saut de discontinuité en $\mathbf{x}_0$ est égal à a. L'hypothèse alternative locale est alors définie par l'existence d'une fonction $f(\cdot; \mathbf{x}_0, a, \theta, L)$.



(a) *(b)*

FIG 2. — (*a*) Approximation de Γ par sa tangente en $\mathbf{x}_0$. (*b*) Modèle de perturbation permettant de spécifier une hypothèse alternative pour calculer la puissance en $\mathbf{x}_0$.

Calcul de la puissance locale

Notons $1-\beta(\mathbf{x}_0)$ la puissance du test local en $\mathbf{x}_0$. Comme nous ne connaissons pas l'orientation de la discontinuité nous dirons que le taux de détection

d'une discontinuité passant par $\mathbf{x}_0$ est l'intégrale sur toutes les orientations θ possibles de la discontinuité :

$$1-\beta(\mathbf{x}_0) = \frac{1}{\pi}\int_0^{\pi} \{1-\beta(\mathbf{x}_0;\theta)\}\, \mathrm{d}\theta, \tag{8}$$

où $1-\beta(\mathbf{x}_0;\theta)$ représente la puissance sous l'alternative $H_1(\mathbf{x}_0;\theta)$ définie ci-dessus, à orientation θ fixée (Figure 2*a*). Des simulations (Gabriel, 2004) ont montré que $1-\frac{1}{8}\sum_{i=0}^{7}\beta(\mathbf{x}_0; i\pi/8)$ fournit une bonne approximation de l'intégrale (8).

Sous $H_1(\mathbf{x}_0;\theta)$ le gradient estimé en $\mathbf{x}$ s'écrit :

$$W(\mathbf{x};\theta) = W_{H_0}(\mathbf{x}) + k_a(\mathbf{x};\theta) = \partial C'(\mathbf{x})\mathbf{K}^{-1}\mathbf{Z} + \partial C'(\mathbf{x})\mathbf{K}^{-1}A(\mathbf{x}_0;\theta), \tag{9}$$

où $A(\mathbf{x}_0;\theta)$ est un vecteur de longeur n, d'éléments $f(\mathbf{x}_i;\mathbf{x}_0,a,\theta,L)$, $i=1,\ldots,n$.

Afin de calculer la puissance en $\mathbf{x}_0$, nous allons utiliser la forme paramétrique de l'équation d'un cercle $U_1^2(\mathbf{x};\theta)+U_2^2(\mathbf{x};\theta)=t_{\mathbf{x}}$:

$$U_1(\mathbf{x};\theta) = \sqrt{t_{\mathbf{x}}}\cos(\omega_{\mathbf{x}}), U_2(\mathbf{x};\theta) = \sqrt{t_{\mathbf{x}}}\sin(\omega_{\mathbf{x}}). \tag{10}$$

Soit $T(\mathbf{x};\theta) = W'(\mathbf{x};\theta)\mathbf{\Sigma}^{-1}(\mathbf{x})W(\mathbf{x};\theta)$, où $W(\mathbf{x};\theta)$ est donné par l'équation (9). La puissance $1-\beta(\mathbf{x}_0;\theta)$ s'écrit :

$$\begin{aligned}
&1-\beta(\mathbf{x}_0;\theta)\\
&= \mathbb{P}_{H_1(\mathbf{x}_0;\theta)}\left[\bigcup_{p=1}^{N}\left\{\text{Rejet de l'hypothèse locale nulle } H_0(\mathbf{x}_p)\right\}\right]\\
&= 1-\mathbb{P}_{H_1(\mathbf{x}_0;\theta)}\left[T(\mathbf{x}_1;\theta)\leqslant t_{1-\alpha},\ldots,T(\mathbf{x}_N;\theta)\leqslant t_{1-\alpha}\right],\\
&= 1-\int_0^{t_{1-\alpha}}\cdots\int_0^{t_{1-\alpha}}\mathbb{P}_{H_1(\mathbf{x}_0;\theta)}\left[T(\mathbf{x}_1;\theta)=t_1,\ldots,T(\mathbf{x}_N;\theta)=t_N\right]\mathrm{d}t_1\ldots\mathrm{d}t_N\\
&= 1-\frac{1}{2^N}\int_{\mathbf{w}\in[0,2\pi]^N}\int_{\mathbf{t}\in[0,t_{1-\alpha}]^N} f_{\mathbf{V}_\theta}(\mathbf{v})\,\mathrm{d}\mathbf{t}\,\mathrm{d}\mathbf{w},
\end{aligned} \tag{11}$$

où

- N est le nombre de pixels de la grille sur laquelle la méthode et le calcul de la puissance sont implémentés (voir section 4),
- $\mathbf{V}_\theta = \Big(U_1(\mathbf{x}_1;\theta),U_2(\mathbf{x}_1;\theta),\ldots,U_1(\mathbf{x}_N;\theta),U_2(\mathbf{x}_N;\theta)\Big)'$ est un vecteur gaussien, de moyenne $\mathbf{m}_{\mathbf{V}_\theta} = \big(\mu_1(\mathbf{x}_1;a,\theta),\mu_2(\mathbf{x}_1;a,\theta),\ldots,\mu_1(\mathbf{x}_N;a,\theta),\mu_2(\mathbf{x}_N;a,\theta)\big)'$ et de matrice de covariance :

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

avec $c_{ij}(\mathbf{x}_k, \mathbf{x}_l) = \mathrm{Cov}\left(U_i(\mathbf{x}_k;\theta), U_j(\mathbf{x}_l;\theta)\right)$, $i,j \in \{1,2\}$, $k,j \in \{1,\ldots,N\}$ indépendant de θ,

- $\mathbf{w} = (\omega_1,\ldots,\omega_N)'$ et $\mathbf{t} = (t_1,\ldots,t_N)'$ proviennent de l'équation (10),
- $\mathbf{v} = \left(\sqrt{t_1}\cos\omega_1, \sqrt{t_1}\sin\omega_1, \ldots, \sqrt{t_N}\cos\omega_N, \sqrt{t_N}\sin\omega_N\right)'$.

En pratique, l'intégrale multiple (11) est évaluée par une approximation de type Monte-Carlo : pour une série de n_s simulations du vecteur $\mathbf{V}_\theta$,

$$1-\beta(\mathbf{x}_0;\theta) \approx 1 - \frac{1}{n_s}\sum_{i=1}^{n_s} \mathbf{1}_{\{T(\mathbf{x}_1;\theta)\leqslant t_{1-\alpha}\}} \ldots \mathbf{1}_{\{T(\mathbf{x}_N;\theta)\leqslant t_{1-\alpha}\}}.$$

4. Implémentation et illustrations

En pratique la méthode est appliquée sur une grille superposée sur le domaine d'étude. En chaque pixel $\mathbf{x}_p$, nous calculons dans un premier temps le gradient $W(\mathbf{x}_p)$, sa matrice de covariance $\mathbf{\Sigma}(\mathbf{x}_p)$ et le champ $T(\mathbf{x}_p)$. Pour ces calculs nous avons besoin de la matrice de covariance entre les données, $\mathbf{C}$, et en chaque pixel du vecteur de covariance $C_Z(\mathbf{x}_p)$ et de ses dérivées. Ces éléments sont obtenus à partir de la fonction de covariance $C_Z(\cdot)$ du champ $Z(\cdot)$ qui doit être estimée (voir section 4.3). Nous définissons ensuite les ZCAs potentielles au niveau $1-\alpha$ comme l'ensemble des pixels pour lesquels la statistique $T(\mathbf{x}_p)$ est au-desssus de $t_{1-\alpha}$. Pour chaque ZCA potentielle de surface $S_{t_{1-\alpha}}$, nous calculons la valeur critique associée au test de significativité : $p = \exp\left(-t_{1-\alpha}\det\left(\mathbf{\Lambda}\right)^{1/2} S_{t_{1-\alpha}}/2\pi\right)$. Cette relation est une approximation numérique du théorème 3 dans laquelle la surface de la composante connexe est approchée par la surface des pixels qui la composent.

Nous allons donc maintenant successivement traiter le problème de la discrétisation du domaine par une grille, de la détermination du niveau local α et de l'estimation de la fonction de covariance, et voir comment nous avons implémenté le calcul de la puissance du test de détection.

4.1. Discrétisation du plan par une grille

Nous avons montré que dans le plan et sous H_0 la statistique $X = t_{1-\alpha}S_{t_{1-\alpha}}\det\left(\mathbf{\Lambda}\right)^{1/2}/\pi$ utilisée dans le test global de significativité a pour distribution une loi exponentielle d'espérance 2 lorsque $t_{1-\alpha}\to\infty$. Cependant en pratique X dépend de la discrétisation. Pour un ensemble de 1000 simulations sous H_0, nous avons calculé la statistique X obtenue pour $1-\alpha = 0.995$ et 0.999 et selon deux tailles de grilles : 30×30 et 60×60. Ces résultats ont montré que, si la discrétisation n'est pas assez fine, les petites ZCAs, *i.e.* les faibles valeurs de X sont sous-représentées. Cela vient du fait que par construction les ZCAs dont la surface est inférieure à la maille de la grille ne sont pas détectées. Ce phénomène ré-apparaît lorsque le niveau $1-\alpha$ augmente car dans ce cas les ZCAs ont tendance à être petites et surtout peu

nombreuses. Ainsi, le niveau $1-\alpha$ doit varier avec la taille de la grille, et nous dirons que la discrétisation est raisonnablement fine dès que l'ensemble des valeurs de X peut être considéré comme une réalisation d'une variable exponentielle d'espérance 2. En pratique une grille 100×100 s'est toujours révélée largement suffisante.

4.2. Détermination du niveau local α

La méthode nécessite deux niveaux de significativité : un niveau global η, fixé par l'utilisateur, et un niveau local α, lié à η, pour la détection de ZCAs. L'équation (6) est vérifiée dans le plan lorsque $t_{1-\alpha} \to \infty$, *i.e.* lorsque $\alpha \to 0$. Mais lorsque $\alpha \to 0$, les ZCAs potentielles tendent à devenir très petites et potentiellement de surface inférieure à la maille de la grille. Aussi, le niveau α doit permettre d'établir un certain compromis entre les contraintes mathématiques et l'aptitude à détecter des ZCAs potentielles. Le niveau α dépend donc, de par l'équation (6), de η et de la fonction de covariance, mais aussi de la grille d'interpolation et du schéma d'échantillonnage. Il est estimé par simulation Monte-Carlo : pour une série de M simulations sous H_0, correspondant à la configuration d'échantillonnage et à la fonction de covariance de la variable, le niveau $\hat{\alpha}$ correspond à la plus grande valeur pour laquelle ηM simulations exactement présentent des ZCAs.

Dans Gabriel (2004), il a pu être montré que, lorsque la discrétisation est suffisamment fine, un niveau α_G, lié à la portée intégrale de la fonction de covariance, d'une part fournissait une bonne approximation du niveau α et d'autre part nécessitait beaucoup moins de calculs. La portée intégrale (Lantuéjoul, 1991) permet de déterminer une approximation du nombre d'hypothèses locales indépendantes que l'on peut considérer dans $\mathcal{D}$. La portée intégrale, A, de $C_Z(\mathbf{h})$ est :

$$A = \int_{\mathbb{R}^2} \frac{C_Z(\mathbf{h})}{C_Z(\mathbf{0})} \, \mathrm{d}\mathbf{h}. \tag{12}$$

Soit $\bar{Z}(\mathcal{D})$ la moyenne spatiale de $Z(\cdot)$ sur $\mathcal{D}$. Si $A \neq 0$ et $|\mathcal{D}| >> A$, d'après Lantuéjoul (1991), $\mathrm{Var}\big(\bar{Z}(D)\big) \approx \sigma^2/N$, avec $N = |\mathcal{D}|/A$. Aussi nous avons :

$$\eta = \mathbb{P}\Big[\cup_{\mathbf{x}\in\mathcal{D}} \big\{T(\mathbf{x}) > t_{1-\alpha}\big\}\Big] \approx 1 - \Big(\mathbb{P}\big[T(\mathbf{x}) \leqslant t_{1-\alpha}\big]\Big)^N,$$

ce qui mène à la définition suivante de α_G :

$$\alpha_G = 1 - (1-\eta)^{1/N}. \tag{13}$$

4.3 Estimation de la fonction de covariance

La méthode suppose que la fonction de covariance de $Z(\cdot)$ est connue, mais en pratique elle doit être estimée. Sous l'hypothèse de stationnarité, d'après l'équation (2), nous estimons la fonction de covariance via l'estimation du variogramme (Equation 1). Il y a trois paramètres à estimer.

Le premier paramètre est la forme paramétrique (modèle exponentiel, sphérique, ...). Des simulations (Gabriel *et al.*, 2004) ont montré qu'elle est d'importance secondaire pourvu que les conditions sur la régularité de la fonction de covariance de $Z(\cdot)$ soient remplies.

Le deuxième paramètre à estimer est la portée (distance à partir de laquelle il n'y a plus de corrélation notable entre les données). Des résultats de simulations montrent que les ZCAs sont mieux détectées pour une surestimation de la portée et sont dans ce cas plus grandes, mais au prix d'un taux de fausses alarmes plus élevé (détection de ZCAs ne correspondant à aucune discontinuité réelle). Ce phénomène s'explique par une régularité plus importante du champ aléatoire lorsque la portée est grande, ce qui a pour conséquence de rejeter plus souvent l'hypothèse de stationnarité en présence d'une discontinuité. Dans le cas d'une sous-estimation le résultat inverse est observé.

Enfin le dernier paramètre à estimer est la variance. La méthode est appliquée sous l'hypothèse nulle d'absence de ZCAs. Aussi en présence d'une discontinuité, la variance σ^2 du champ $Z(\cdot)$ est surestimée. Comme $T(\mathbf{x})$ est proportionnelle à σ^{-2}, cela entraîne de faibles valeurs du champ $T(\cdot)$, des petites ZCAs, de grandes p-valeurs et par conséquent une perte de puissance de la méthode. Afin de surmonter cette difficulté, nous proposons une procédure itérative dans laquelle nous estimons à chaque itération alternativement les paramètres du variogramme et les ZCAs. Dans un premier temps nous estimons le variogramme global, *i.e.* en utilisant tous les couples de points. En présence de ZCAs, le variogramme est ré-estimé en éliminant tous les couples de données $\{Z(\mathbf{x}_k), Z(\mathbf{x}_l)\}$ tels que le segment $[\mathbf{x}_k, \mathbf{x}_l]$ intersecte une ZCA. Si la portée change, le niveau α doit être recalculé. Les ZCAs sont alors réestimées à l'aide des nouveaux paramètres. La procédure est ré-itérée jusqu'à convergence, *i.e.* jusqu'à ce que l'ensemble de ZCAs détectées soit identique pour deux itérations successives. Malgré l'absence de preuve de la convergence de cette procédure, son application aussi bien sur des simulations que sur des données réelles a montré que la convergence était toujours atteinte en moins de 5 itérations. Cela s'explique par le fait que la pluplart des couples de points écartés lors de la ré-estimation du variogramme sont éliminés dès la première ré-estimation. Les paramètres de la fonction de covariance sont ensuite assez proches lors des itérations suivantes.

4.4. Illustration

La méthode a été implémentée avec le logiciel **R**. Le code est accessible sur la page web de l'auteur : www.maths.lancs.ac.uk/∼gabriel.

Nous allons illustrer cette procédure sur l'exemple présenté dans l'introduction. Nous avions simulé 100 points issus d'un champ gaussien centré-réduit, de fonction de covariance exponentielle de portée $b = 0.1$. Une discontinuité avait été introduite en ajoutant aux points situés au-dessus de la courbe en pointillés (figure 1*a*) une constante $a = 2.5$. Pour cet exemple, une grille 60 $\times$ 60 s'est révélée suffisante. Les paramètres du modèle de covariance estimés

à la première itération sont : $\hat{b} = 0.137$ et $\hat{\sigma}^2 = 2.083$, loin des valeurs du modèle simulé.

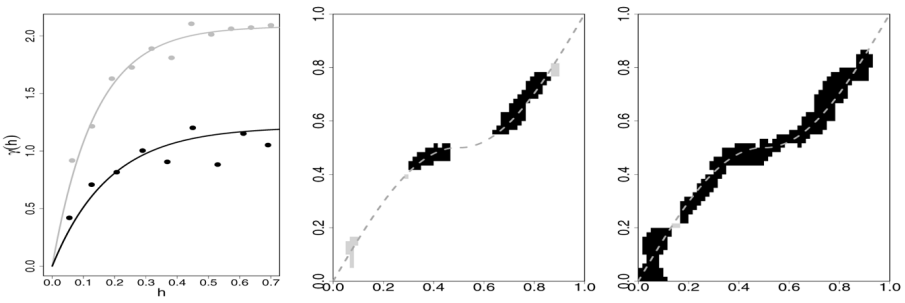


FIG 3. — (*a*) Variogrammes empiriques (points) et estimés (courbes) à la première itération (gris) et à la convergence (noir). (*b*) Détection des ZCAs à la première itération. (*c*) Détection des ZCAs à la convergence ; les ZCAs significatives sont représentées en noir et les non significatives en gris.

La Figure 3*b* illustre les tests locaux au niveau $1 - \hat{\alpha} = 0.9939$ correspondant au niveau global $\eta = 0.05$ pour $\hat{b} = 0.137$. Les pixels sont colorés en noir si ils appartiennent à une composante connexe significative et sont colorés en gris dans le cas contraire. La procédure itérative converge à la quatrième itération. Les paramètres obtenus à la convergence sont : $\hat{b} = 0.128$, $\hat{\sigma}^2 = 1.21$ et $1-\hat{\alpha} = 0.9973$. La figure 3*a* représente les variogrammes empiriques (points) et estimés (courbes) à la première itération (gris) et à la convergence (noir). Nous pouvons constater que la variance estimée à la convergence a diminué de 42% par rapport à celle de la première itération. Les figures 3*b* et 3*c* montrent que la discontinuité est beaucoup mieux détectée à la convergence.

4.5. Approximation de la puissance du test local

La puissance $1 - \beta(\mathbf{x}_0, \theta)$ du test de détection de ZCAs est calculée en chaque pixel de la grille d'interpolation. Dans le calcul présenté dans la section 3.4, la puissance tient compte de toute l'information du domaine et cela rend son évaluation très lourde. En effet, pour calculer l'intégrale multiple (11), nous devons connaître les éléments de la matrice de covariance $\mathbf{\Sigma}_{\mathbf{V}_\theta}$. Cette matrice étant de dimension $2N^2 \times 2N^2$, elle est difficile voire impossible à inverser lorsque N devient grand. Comme $f(\mathbf{x})$ tend vers 0 lorsque l'on s'éloigne de la discontinuité, les pixels se situant loin de la discontinuité contribuent peu à l'intégrale (11). Nous calculons donc la puissance en un point $\mathbf{x}_0$ en nous limitant à une fenêtre $\mathcal{F}_k \subseteq \mathcal{D}$ centrée en $\mathbf{x}_0$ que nous imposerons carrée et contenant $N_k = (2k+1) \times (2k+1)$ pixels. Notons $1 - \beta_{\mathcal{F}_k}(\mathbf{x}_0, \theta)$ cette puissance restreinte à la fenêtre $\mathcal{F}_k$. Puisque $\mathcal{F}_k \subseteq \mathcal{F}_{k'}$ lorsque $k \leqslant k'$, nous obtenons : $1 - \beta_{\mathcal{F}_k}(\mathbf{x}_0; \theta) \leqslant 1 - \beta_{\mathcal{F}_{k'}}(\mathbf{x}_0; \theta) \leqslant 1 - \beta(\mathbf{x}_0; \theta)$. Lorsque $\mathcal{F}_k \to \mathcal{D}$, nous avons : $1 - \beta_{\mathcal{F}_k}(\mathbf{x}_0; \theta) \longrightarrow 1 - \beta(\mathbf{x}_0; \theta)$, avec $1 - \beta_{\mathcal{F}_k}(\mathbf{x}_0, \theta) = 1 - \frac{1}{2^{N_k}} \int_{\mathbf{w} \in [0,2\pi]^{N_k}} \int_{\mathbf{t} \in [0, t_{1-\alpha}]^{N_k}} f_{\mathbf{V}_\theta}(\mathbf{v}) \, \mathrm{d}\mathbf{t} \, \mathrm{d}\mathbf{w}$. Le choix de k est important. D'une part k doit être le plus grand possible afin d'approcher $1 - \beta(\mathbf{x}_0; \theta)$.

Mais d'autre part la matrice $\mathbf{\Sigma}_{\mathbf{V}_\theta}$ étant de dimension $2(2k+1)^2 \times 2(2k+1)^2$, elle devient difficile à calculer lorsque la taille de $\mathcal{F}_k$ augmente (Table 1). D'après le modèle d'hypothèse alternative choisi, les pixels $\mathbf{x}$ distants de plus de $L/2$ de $\mathbf{x}_0$ peuvent ne pas être considérés. En effet pour ces pixels les éléments de $A(\mathbf{x}_0;\theta)$ sont très faibles dans le voisinage de $\mathbf{x}$ et la probabilité de rejet de $H_0(\mathbf{x})$ sachant que $H_1(\mathbf{x}_0;\theta)$ est vraie devient négligeable.

TABLEAU 1. — Dimension de la matrice de covariance $\mathbf{\Sigma}_{\mathbf{V}_\theta}$ comme fonction de la taille de $\mathcal{F}_k$.

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

Comme nous considérons un sous-ensemble $\mathcal{F}_k$ de $\mathcal{D}$, nous ne pouvons pas utiliser les niveaux $\hat{\alpha}$ et α_G définis ci-desssus. En effet, $\hat{\alpha}$ dépend de l'échantillonnage sur l'ensemble du domaine et est essentiellement influencé par les zones de $\mathcal{D}$ pour lesquelles l'échantillon de points est le plus agrégé. Par conséquent, pour toute fenêtre $\mathcal{F}_k$, $\hat{\alpha}$ tend à sous-estimer la vraie valeur de α dans cette fenêtre. En ce qui concerne α_G, il s'agit d'une approximation de $\hat{\alpha}$ dans le cas où la discrétisation de la grille d'interpolation est assez fine, ce qui n'est plus le cas dans le sous-ensemble $\mathcal{F}_k$. Afin de calculer la puissance $1-\beta_{\mathcal{F}_k}(\mathbf{x}_0;\theta)$, nous évaluons pour chaque fenêtre $\mathcal{F}_k$ de centre $\mathbf{x}_0$, un niveau local $\alpha_k(\mathbf{x}_0)$. Le niveau $\alpha_k(\mathbf{x}_0)$ est déterminé de la même manière que le niveau α, sauf que $\hat{\alpha}_k(\mathbf{x}_0)$ est tel que des ZCAs sont détectées pour ηM simulations dans la fenêtre $\mathcal{F}_k$.

4.6. Effet du schéma d'échantillonnage sur la détection de ZCAs

Afin d'illustrer l'effet du schéma d'échantillonnage sur la détection de ZCAs, nous reprenons l'exemple vu ci-dessus. Nous avons éliminé aléatoirement 20 points de l'échantillon complet et ainsi obtenu un sous-échantillon de $n = 80$ points. Cette opération a été itérée deux fois afin d'obtenir des sous-échantillons de 60 et 40 points. La première ligne de la figure 4 représente la répartition spatiale des différents échantillons. Notons que les sous-échantillons ne sont pas nécessairement spatialement homogènes, par exemple dans la partie $y < 0.4$ de l'échantillon de 40 points.

La deuxième ligne de la figure 4 montre les ZCAs détectées (ZCAs significatives en noir et non significatives en gris). Cette figure illustre la détérioration de la détection de ZCAs lorsque le nombre de points d'échantillonnage diminue. On constate en particulier que la ZCA détectée en bas à gauche de la discontinuité s'affine, puis devient non significative et enfin disparaît. Seule la ZCA au centre du carré est détectée pour tous les échantillons. Dans le cas $n = 40$, elle ne correspond qu'à un petit fragment de la discontinuité.

La puissance locale a été calculée en chaque pixel de la grille d'interpolation, en utilisant une fenêtre glissante de taille 9×9 pixels pour $a = 2.5$ et $L = 0.3$. La troisième ligne de la figure 4 donne la cartographie de la puissance calculée pour chaque échantillon de points. L'implémentation de la puissance ne permet pas de calculer la puissance à proximité des bords du domaine, d'où la présence d'une bande de pixels blancs autour du carré. Ces figures montrent que la puissance dépend de la densité locale d'échantillonnage : plus l'échantillon de points est localement dense, plus la puissance est localement élevée. Au contraire, plus l'échantillon est épars, plus la puissance est faible. Pour les quatre échantillons de points la cartographie de la puissance montre des zones de forte puissance situées autour de points non isolés et des zones de plus faible puissance entre les "clusters" lorsque n diminue. Nous constatons également que la discontinuité introduite dans notre exemple se situe dans une zone de forte puissance pour les échantillons de 100, 80 et 60 points. Ce n'est plus le cas pour l'échantillon de 40 points.

5. Une application aux sciences de l'environnement

5.1. Présentation des données

La méthode présentée dans la section 3 est illustrée sur des données de sol prélevées près de La Chaux de Fonds (figure 5*a*), dans une région de 14.5 km^2. La stratégie d'échantillonnage, spécialement mise en œuvre pour cartographier la distribution des polluants sur l'ensemble de la région, est la suivante. Un premier ensemble de points d'échantillonnage correspond aux 214 nœuds d'une grille régulière superposée sur le domaine d'étude. La distance entre nœuds est de 250 m. Un cinquième de ces points sont le point de départ d'une progression géométrique de distances 100 m, 40 m, 15 m et 6 m. La figure 5*b* illustre les 359 points d'échantillonnage. Ces données ont été analysées par Atteia *et al.* (1994) et publiées dans Goovaerts (1997). L'analyse des concentrations de plusieurs métaux lourds potentiellement toxiques (cadmium, chrome, cobalt, cuivre, nickel, plomb et zinc) a permis d'étudier les risques de pollution et de déterminer les sources des polluants. Nous connaissons également la géologie de la région (figure 5*b*). Cette région date du Jurassique supérieur (ère secondaire) et est constituée de marnes noires (argovien, kimmeridgien et sequanien), de calcaires et de grès (portlandien) et de lehm (zones décalcifiées).

Nous allons indépendamment nous intéresser aux concentrations de cobalt et de nickel. Le résumé statistique de ces concentrations est donné dans la table 2.

TABLEAU 2. — Résumé statistique des concentrations de cobalt et de nickel.

	Min.	1er Qu.	Med.	Moy.	3ème Qu.	Max.	σ
Cobalt	1.55	6.66	9.84	9.44	12.10	20.60	3.57
Nickel	1.98	14.60	20.68	20.02	25.38	53.20	8.09

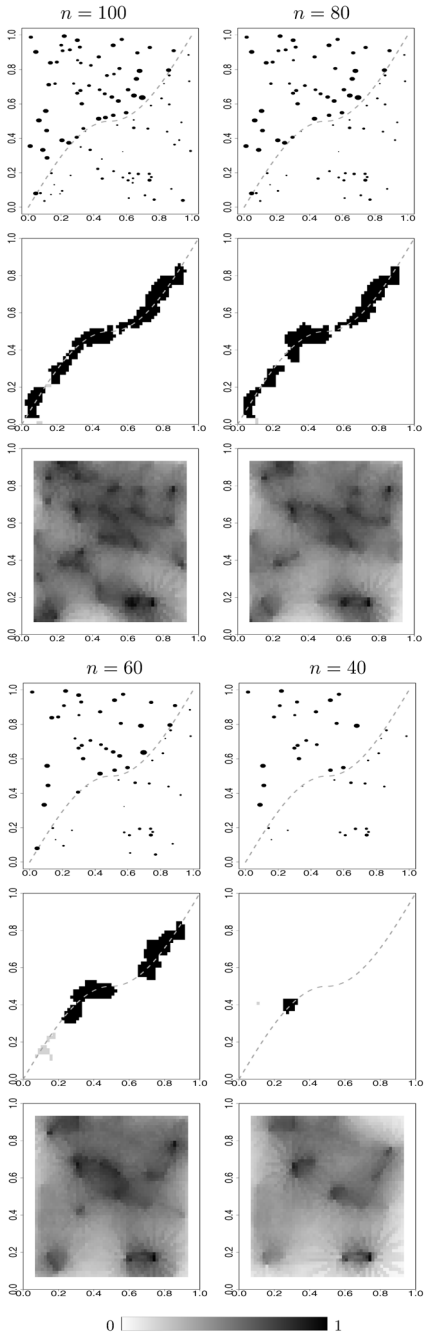
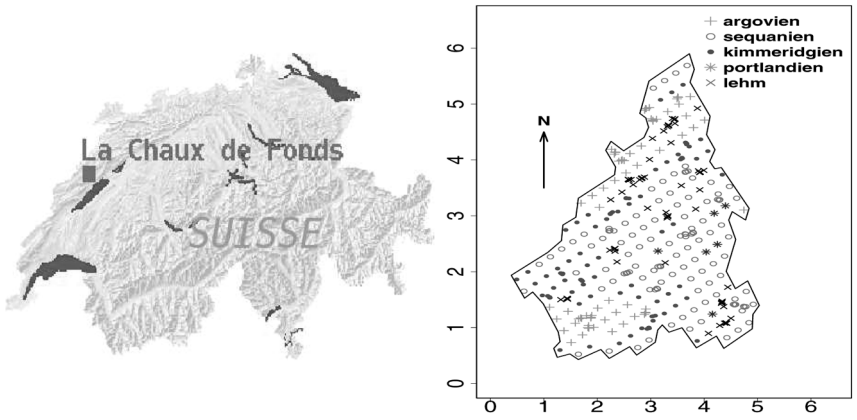


FIG 4. — Première ligne : échantillons de points de différentes densités (voir texte) ; Deuxième ligne : ZCAs (significatives en noir, non significatives en gris) ; Troisième ligne : puissance du test de détection.



(a) *(b)*

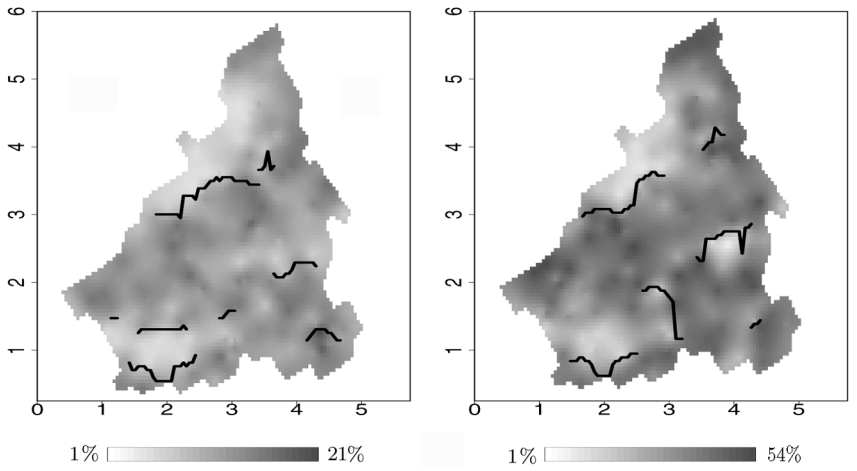
FIG 5. — (*a*) Localisation de la région. (*b*) Carte géologique et schéma d'échantillonnage.

5.2. Détection de ZCAs

Nous avons estimé à chaque étape de la procédure itérative des fonctions de covariance exponentielle : $C_Z(\mathbf{h}) = \hat{\sigma}^2 \exp(-\|\mathbf{h}\|/\hat{b})$, où $\hat{\sigma}^2$ et $\hat{b}$ sont les estimations des paramètres de palier et de portée. D'après l'équation (12), la portée intégrale d'une fonction de covariance exponentielle définie dans $\mathbb{R}^2$ est $A = 2\pi b^2$. Les niveaux locaux α ont été calculés à partir de l'équation (13) : $\alpha_G = 1 - (1-\eta)^{2\pi\hat{b}/|\mathcal{D}|}$, avec η fixé à 5%. La procédure itérative a convergé à la deuxième itération pour le cobalt et à la troisième pour le nickel.

La figure 6 représente les ZCAs détectées sur une grille 100×100, superposées à la carte d'interpolation des variables obtenue par krigeage ordinaire. Nous pouvons remarquer que les concentrations de nickel et de cobalt ont les mêmes variations spatiales. Les concentrations sont faibles dans le nord, le sud-ouest et le centre-est de la région. Les ZCAs sont tracées via le même procédé qu'à la figure 1*c*. Elles sont essentiellement détectées le long des transitions entourant les zones de faible concentration. Nous retrouvons les mêmes ZCAs pour le cobalt et le nickel dans les parties nord, centre-est et sud-ouest de la région. L'observation de la carte géologique de la région (figure 5*b*) montre qu'elles se situent le long des frontières entre l'argovien et le non-argovien.

Nous retrouvons les résultats de Atteia *et al.* (1994) : les concentrations de cobalt et de nickel ont des distributions spatiales similaires à celles de la géologie, ce qui suggère que ces métaux sont principalement issus de la roche de fond et non de l'industrie ou de l'agriculture. L'utilisation de la méthode de détection de ZCAs pourrait donc permettre aux scientifiques d'explorer leur données afin de détecter les traits structuraux les plus importants.



(a) *(b)*

FIG 6. — Carte d'interpolation et ZCAs détectées dans les concentrations de : *a*) cobalt, *b*) nickel.

6. Discussion

Nous avons considéré que les variations du sol pouvaient être modélisées par la somme d'un champ moyen, pouvant présenter des discontinuités, et d'un résidu aléatoire spatialement corrélé. Dans ce contexte, nous avons proposé une méthode permettant de tester l'hypothèse nulle de stationnarité de la variable contre l'hypothèse alternative que les données présentent des ZCAs correspondant à des discontinuités ou des variations fortes de la moyenne. Par conséquent, si une variable présente des variations liées à la fonction aléatoire et possède une moyenne constante ou avec de faibles variations, aucune ZCA ne sera détectée.

Plusieurs hypothèses ont été faites afin d'établir la méthode. Nous avons en particulier supposé la stationnarité d'ordre 2 et la régularité de la fonction de covariance. Les deux hypothèses $\mathcal{H}_1$ et $\mathcal{H}_2$ peuvent être affaiblies. L'hypothèse $\mathcal{H}_1$ peut être remplacée par $\mathcal{H}_1'$: $Z(\cdot)$ satisfait l'hypothèse intrinsèque de stationnarité (section 2.2). Dans ce cas, l'interpolateur de krigeage est le krigeage ordinaire (Equation 4) dans lequel le variogramme remplace la fonction de covariance. L'hypothèse $\mathcal{H}_2$ porte sur la régularité de la fonction de covariance en dehors de 0. Nous avons supposé que la fonction de covariance est indéfiniment différentiable pour tout $\mathbf{h}$ tel que $||\mathbf{h}|| > 0$. En réalité seule la différentiabilité à l'ordre trois est nécessaire.

Nous avons également supposé que les données étaient gaussiennes. Cette hypothèse est nécessaire pour établir les résultats théoriques présentés dans la section 3 qui rendent possible la prise en compte de l'autocorrélation entre les données afin de détecter des ZCAs à partir d'échantillons de points de faible densité en présence d'autocorrélation. Si les données ne sont pas

gaussiennes elles peuvent être transformées, par exemple en utilisant une transformation logarithmique ou une transformation de Cox. Cela n'assure pas la "gaussiannité" du champ aléatoire, mais semble être suffisant pour que la méthode puisse raisonnablement bien fonctionner en pratique.

La généralisation à des données multivariées reste un problème ouvert. Une telle extension requiert d'une part la démonstration du théorème 3 dans un contexte multivariable, et d'autre part la mise en place d'une procédure entièrement automatique d'estimation du variogramme.

7. Annexe

Dans le théorème 3, $\mathbf{\Lambda}_i(\mathbf{x})$ est la matrice de covariance du gradient de $U_i(\mathbf{x})$, champ gaussien intervenant dans la décomposition du champ de $\chi^2(2)$ $T(\mathbf{x})$:

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

avec $W(\mathbf{x}) = \big(W_1(\mathbf{x}), W_2(\mathbf{x})\big)'$, $\sigma_i(\mathbf{x}) = \sqrt{\mathbf{\Sigma}_{[ii]}(\mathbf{x})}, i = 1,2$ et $\rho(\mathbf{x}) = \mathbf{\Sigma}_{[12]}(\mathbf{x})/\sigma_1(\mathbf{x})\sigma_2(\mathbf{x})$.

Nous allons donner les expressions de $\mathbf{\Lambda}_i(\mathbf{x}) = \mathbb{E}\big[\partial U_i(\mathbf{x}) \partial U_i'(\mathbf{x})\big]$. Pour cela, notons $\sigma_i = \sigma_i(\mathbf{x}),\ i = 1,2,\ \rho = \rho(\mathbf{x})$ et $D_i = D_i(\mathbf{x}) = \partial_i C(\mathbf{x}),\ i = 1,2$ où $\partial_i f = \partial f / \partial x^i$.

Pour $k, l = 1, 2$:

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

8. Références

Adler R. (1981) *The Geometry of Random Fields*. New-York : Wiley.

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. *Research Report 2*, Unité de Biométrie, INRA-Avignon. Available at : www.avignon.inra.fr/internet/unites/biometrie/publications_scientifiques/RR.htm

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

Atteia O., Dubois J-P. and Webster R. (1994) Geostatistical analysis of soil contamination in the Swiss Jura. *Environmental Pollution*, **86**, 315-327.

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.

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. Available at :
www.maths.lancs.ac.uk/ gabriel/papers/TheseGabriel.pdf

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* P. Monestiez, D. Allard and R. Froidevaux (eds), Kluwer, The Netherlands, pp. 311-322.

Goovaerts P. 1997. *Geostatistics for Natural Resources Evaluation*. Oxford Univ. Press, New-York.

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.

Matheron G. (1962) Traité de Géostatistique Appliquée. Tome **I**. *Mémoires du Bureau de Recherches Géologiques et Minières*. **No. 24**. Editions Bureau de Recherches Géologiques et Minières, Paris.

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.

YAGLOM A. M. (1986) *Correlation theory of stationary and related random functions*. Springer-Verlag, Berlin.