Le traitement des variables régionalisées en écologie : apports de la géomatique et de la géostatistique

Philippe Aubry

N° d'ordre 13 - 2000 Année 2000

THÈSE

présentée

devant l'UNIVERSITÉ CLAUDE BERNARD - LYON 1

pour l'obtention

du DIPLÔME DE DOCTORAT

Spécialité Biométrie

(arrêté du 30 mars 1992)

par

Philippe AUBRY

LE TRAITEMENT DES VARIABLES RÉGIONALISÉES EN ÉCOLOGIE APPORTS DE LA GÉOMATIQUE ET DE LA GÉOSTATISTIQUE

Soutenue le 6 janvier 2000 devant le jury composé de :

M. Christian GAUTIER	Président
M. Joël CHADOEUF	Rapporteur
M. François HOULLIER	Rapporteur
M. François BOUILLÉ	Examinateur
M. Domitien DEBOUZIE	Examinateur
M. Frederick DELAY	Examinateur

Laboratoire de Biométrie et de Biologie Évolutive - UMR CNRS 5558
Université Claude Bernard - Lyon 1
43, boulevard du 11 novembre 1918
69622 Villeurbanne Cedex

ABSTRACT

Processing ecological regionalised variables without taking their spatial properties into account raises a number of problems. To address these problems, we develop geomatical methods based on computerized procedures, and geostatistical methods based on the theory of random functions.

After introducing fundamentals of geomatics and geostatistics and clarifying the specific nature of spatial autocorrelation, we introduce design-based and model-based inferences. We use random functions to study sampling efficiency, to optimize predictors and compute prediction intervals.

We propose a procedure to optimize distance classes when computing semivariograms. We also examine the use of semivariogram integral and justify semivariogram modelling by using weighted least squares fitting. We discuss semivariogram accuracy under design-based and model-based inferences, and we clarify the meaning of jackknifing.

The optimization of sampling from a finite population in order to estimate the spatial mean or the semivariogram is examined using several combinatorial heuristics and simulations of random functions.

The problem of testing correlation or association between two regionalised variables is studied, using again simulations of random functions. Several methods are reviewed and we recommend tests which explicitly take spatial autocorrelation into account.

In the frame of spatial association between regionalised variables, we propose a hybrid method using quadtrees and an editing distance between recursive rooted trees.

Finally, we study measures of spatial complexity, criticize fractal analysis and propose alternative methods such as a measure of topological complexity for isolines maps.

Key Words: regionalised variables, geomatics, spatial autocorrelation, geostatistics, semivariogram, statistical ecology, design-based and model-based inferences, simulation, optimization, association, complexity.

Title: Regionalised variables processing in Ecology: contribution of Geomatics and Geostatistics.

RÉSUMÉ

Face à la contradiction consistant à traiter les variables régionalisées écologiques sans tenir compte de leurs propriétés spatiales, nous développons des méthodes géomatiques, utilisant des techniques informatiques, et géostatistiques, appliquant la théorie des fonctions aléatoires.

Après avoir introduit des éléments de géomatique et de géostatistique, et avoir précisé la nature spécifique de l'autocorrélation spatiale, nous introduisons les inférences *design-based* et *model-based*. A l'aide de fonctions aléatoires, nous étudions l'efficacité de l'échantillonnage, optimisons des prédicteurs, et calculons des intervalles de prédiction.

Nous proposons une procédure d'optimisation des classes de distances lors du calcul du variogramme. Nous examinons également l'utilisation de l'intégrale du variogramme, et justifions la modélisation du variogramme par ajustement aux moindres carrés pondérés. Nous discutons de la précision du variogramme dans les cadres *design-based* et *model-based*, et au sens du *jackknife*.

L'optimisation de l'échantillonnage d'une population finie en vue de l'estimation de la moyenne spatiale ou du variogramme est examinée à l'aide de plusieurs heuristiques d'optimisation combinatoire et de simulations de fonctions aléatoires.

Le problème du test de la corrélation ou de l'association entre deux variables régionalisées est étudié, à nouveau en utilisant des simulations de fonctions aléatoires. Nous passons en revue plusieurs méthodes et recommandons les tests qui font explicitement référence à l'autocorrélation spatiale des variables régionalisées.

Dans le cadre de la définition de l'association spatiale entre variables régionalisées, nous proposons une méthode hybride utilisant des *quadtrees* et une distance d'édition entre arborescences récursives.

Enfin, nous étudions des mesures de la complexité spatiale, critiquons l'analyse fractale et proposons des méthodes alternatives, notamment une mesure de complexité topologique d'une carte en isolignes.

Mots-clés : variables régionalisées, géomatique, autocorrélation spatiale, géostatistique, variogramme, écologie statistique, inférences *design-based* et *model-based*, simulation, optimisation, association, complexité.

Titre : Le traitement des variables régionalisées en écologie: apports de la géomatique et de la géostatistique.

Remerciements

"Quand on est parti de rien, et qu'on n'est pas arrivé à grand chose, on n'a de merci à dire à personne." (Pierre Dac)

Mes remerciements vont en premier lieu à Domitien Debouzie qui m'a fait con..ance en acceptant d'être mon directeur de recherches en DEA, puis en thèse, et qui m'a permis de développer mon projet avec beaucoup de liberté.

Je remercie François Bouillé (Professeur, Université Pierre & Marie Curie), Joël Chadœuf (Directeur de Recherches, INRA), Frederick Delay (Maître de Conférences, Université Pierre & Marie Curie), Christian Gautier (Professeur, Université Claude Bernard) et François Houillé (Directeur de Recherches, CIRAD), pour avoir accepté de juger ma thèse. Les éventuelles erreurs, omissions ou imprécisions de mon mémoire de thèse restent de ma seule responsabilité.

Merci à Daniel Chessel et à Pierre Chauvet pour avoir répondu à mes questions au début de ma thèse. Toute ma reconnaissance va aux chercheurs qui m'ont con..é le traitement de leurs données spatiales, chronologiquement : Frédéric Lardeux, Claude Dutreix, Xavier Fauvergue, Christian Biémont et Hervé Piegay. Toutes ces collaborations m'ont été extrêmement pro..tables. Toute ma sympathie va également à Mô Dang pour nos échanges concernant la classi..cation des données spatiales. Je tiens aussi à remercier Andrew Solow et Allan Stewart-Oaten, ainsi que les reviewers anonymes dont les commentaires m'ont permis d'améliorer les quelques éléments de cette thèse qui ont fait l'objet des deux articles publiés par la revue Ecology. Merci en..n à Anne-Marie Gonidec pour m'avoir fait découvrir BiblioMacPC et pour nos nombreuses discussions à bâtons rompus.

Cette thèse est l'aboutissement d'une chaîne causale qui prend son origine au début des années 1980, époque à laquelle mes parents ont mis sur mon chemin trois classes d'objets qui ont décidé de toute la suite:

- un ..let à papillons, un guide d'identi..cation de Lépidoptères Rhopalocères et du matériel de collection d'Insectes,
- deux cartes de l'IGN au 1: 25 000 et au 1: 50 000 couvrant les alentours de mon lieu de villégiature, dans le Sud du département de la Creuse,
- un micro-ordinateur ORIC 1 équipé de 48 ko de RAM et d'un microprocesseur 8 bits 6502 de chez Rockwell.

Par la suite, plusieurs personnes ont joué le rôle de catalyseur, par leur exemple, leurs écrits ou leurs cours magistraux. Je tiens tout particulièrement à citer, chronologiquement, Jean-Christophe Richard sans qui j'aurais certainement abandonné la collection des Insectes, Jean-Marie Sibert qui m'a fait découvrir le projet de Cartographie des Invertébrés Européens (CIE), Claude Dutreix dont la thèse de 1986 m'a fait découvrir la biogéographie quantitative des Lépidoptères Rhopalocères, François Bouillé qui m'a enseigné le concept d'algorithme, EXEL (ou ADL), les graphes, HBDS et la géomatique, et en..n Frederick Delay dont les cours de krigeage m'ont permis de prendre pied dans la géostatistique.

Je dédie cette thèse à mes parents et à mes grands parents qui m'ont toujours donné les moyens de "faire des choses" — notamment sous la forme de prêts ..nanciers à taux négatifs.

Table des matières

Liste des figures

Liste des tableaux

Chapitre 1

Introduction

> *"we have to face one fact: the inability of common statistics to take into account the spatial aspect of the phenomenon, which is precisely its most important feature."* (Matheron 1963)
>
> *"Even though spatial pattern is ostensibly being studied, the spatial relationship* [...] *is often ignored."* (Pielou 1969)
>
> *"It is sad to recognize that many distributions and mosaics proposed in books of mathematical ecology* [...] *are quite often irrelevant."* (Margalef 1979)

L'objet de l'écologie est d'étudier de façon globale les interactions entre les organismes et leur environnement. Ces interactions se déroulent à la fois dans les trois dimensions de l'espace et dans le temps. Le cadre fondamentalement spatio-temporel de l'écologie est à l'origine de nombreuses difficultés méthodologiques. Pour simplifier les études, il est classique de découpler l'espace et le temps afin de réaliser, d'une part des études strictement spatiales, et d'autre part, des études strictement temporelles, quitte à procéder ultérieurement à une intégration des résultats de ces deux types d'études et à parler, abusivement, d'étude "spatio-temporelle".

Prendre en considération uniquement la dimension temporelle revient à privilégier la dynamique des phénomènes écologiques aux dépens de leur structuration spatiale. L'approche complémentaire consiste à traiter les phénomènes écologiques d'un point de vue statique, instantané, autrement dit, en ne considérant que l'espace géographique. Dans ce mémoire, nous traitons exclusivement de problèmes méthodologiques introduits par l'espace dans l'étude des phénomènes écologiques : la prise en compte simultanée de l'espace et du temps, bien qu'idéale, est hors de notre propos.

1.1 Structuration spatiale

L'écologie identifie un ensemble de facteurs biotiques et abiotiques constituant l'environnement des organismes. Ces facteurs influencent la survie et la reproduction différentielle des organismes, et par intégration, l'évolution des systèmes intermédiaires entre les organismes et la biosphère tels que les populations locales (ou dèmes), les métapopulations, les peuplements, les écosystèmes, les paysages, voire même, les biomes. En se plaçant à

l'échelle d'observation adéquate, tous les facteurs écologiques présentent de l'*hétérogénéité spatiale* (ou variabilité spatiale), cette hétérogénéité étant caractérisée par un certain degré de *structuration spatiale*.

La structuration spatiale de l'environnement repose sur la variabilité des conditions abiotiques (chaleur, humidité, etc.), la distribution hétérogène des ressources (nutriments, proies, etc.) et l'introduction de contraintes à travers des relations de voisinage, qu'elles soient considérées au niveau des écosystèmes (Margalef 1979) ou des organismes eux-mêmes (Addicott *et al.* 1987). Les effets de cette structuration spatiale se manifestent dans de nombreux processus ou phénomènes écologiques : dynamique des populations (Fahrig 1988, Hanski 1994a, 1994b, Moilanen & Hanski 1998), dispersion (Murray 1988, Hanski *et al.* 1994), compétion (Tilman 1994, Holt 1997, Takenaka *et al.* 1997), diversité spécifique (Tilman 1994, Takenaka *et al.* 1997), etc.

La structuration spatiale de l'environnement détermine en grande partie la distribution spatiale des organismes, et cette distribution spatiale joue un rôle d'autant plus important dans les phénomènes écologiques que les organismes sont peu mobiles. Ainsi, la distribution spatiale est une composante particulièrement importante dans l'interaction parmi les plantes, influençant la compétition, la survie, la fécondité et la dispersion des propagules (Cardina *et al.* 1996). L'importance de la distribution spatiale est confortée par le fait que la prise en compte des aspects spatiaux au sein des modèles de dynamique des populations de plantes change radicalement les conditions de persistence et de coexistence dans les communautés végétales (Czárán & Bartha 1992).

Au-delà des aspects les plus fondamentaux de l'écologie, la prise en compte de la structuration spatiale des populations revêt une importance considérable dans le domaine de la biologie de la conservation (Buckland & Elston 1993, Fahrig & Merriam 1994, Hastings 1994, Hanski *et al.* 1995), dans le maintien de la diversité génétique (Charmet & Balfourier 1995) ou dans le développement des stratégies rationnelles de contrôle des populations de ravageurs (Liebhold *et al.* 1993).

L'importance de la structuration spatiale, quel que soit le niveau d'intégration considéré (population, peuplement, paysage, etc.), se traduit actuellement par une explosion du nombre d'articles traitant de l'espace dans les systèmes écologiques, et en biologie des populations en particulier (Blondel & Lebreton 1996). En conséquence, il n'est pas exagéré d'affirmer que le problème des structures spatiales constitue le problème central de l'écologie (Levin 1992).

1.2 Analyse des structures spatiales

L'écologie est l'une des disciplines scientifiques qui fait le plus appel aux méthodes statistiques (Gaines & Denny 1993), et c'est naturellement le plus souvent sur une base statistique que l'analyse des structures spatiales y est envisagée. Les statistiques spatiales utilisées en écologie ont pour objet l'étude la répartition de points dans le plan (*point pattern analysis*) (Diggle 1983, Ripley 1981, pp. 130-190, Ripley 1988a, pp. 22-73, Upton & Fingleton 1985, pp. 53-104, Cressie 1991, pp. 577-723), ou la variation spatiale d'une *variable régionalisée* (VR), autrement dit, une variable dont les valeurs sont localisées dans l'espace (Ripley 1981, pp. 28-101, Cliff & Ord 1981, Ripley 1988a, pp. 9-21, Upton & Fingleton 1985, pp. 151-212, Cressie 1991, pp. 29-379). En fait, les problèmes posés

en termes de *point pattern analysis* peuvent éventuellement être reformulés en termes de variation spatiale d'une VR, simplement en utilisant la méthode des quadrats. Dans ce qui suit, nous considérons exclusivement l'étude des variables régionalisées et ne traitons pas de la *point pattern analysis*.

L'étude des structures spatiales par des moyens statistiques a une longue histoire en écologie, notamment dans un domaine comme la foresterie (Ripley 1988b). Nous ne cherchons pas à dresser un panorama historique exhaustif des méthodes, dans aucun domaine en particulier, mais nous pouvons mettre en évidence trois types d'approches selon que l'information spatiale est ignorée, prise en compte de façon implicite ou explicite.

1.2.1 Méthodes a-spatiales

De nombreux indices ont été utilisés — ou seulement proposés — afin de classer les distributions spatiales étudiées en structures aléatoires, régulières ou agrégées (exemples dans Taylor *et al.* 1978). Citons, notamment, le paramètre b de la loi de Taylor (Taylor 1961, Taylor 1971), le rapport s^2/m, le paramètre k de la loi binomiale négative, les indices de Lloyd ou de Morisita (Pielou 1969, pp. 90-98, Bliss 1971, Stiteler & Patil 1971, Ripley 1981, pp. 102-106, Upton & Fingleton 1985, pp. 29-31), et plus généralement, les paramètres des distributions discrètes (Cassie 1962, Pielou 1969, pp. 79-89, Kemp 1971, Boswell & Patil 1971, Gurland & Hinz 1971, Ripley 1981, pp. 106-108).

Aucun de ces indices ne tient compte de la localisation spatiale des données, ce qui conduit évidemment à une perte d'information préjudiciable à l'analyse des structures spatiales (Chessel 1978, Larkin *et al.* 1995). En termes statistiques, aucun de ces indices ne satisfait au principe de *suffisance*, au sens où une statistique *suffisante* résume les données sans perte d'information essentielle pour le problème traité (Cassel *et al.* 1977). Dans le cas des indices, il est totalement abusif de parler de *méthodes d'analyse spatiale* (*e.g.*, Taylor 1984) puisque, précisément, les méthodes en question ignorent la localisation spatiale des données (Ripley 1981).

1.2.2 Méthodes spatiales implicites

Une étape importante dans la pratique de l'analyse des structures spatiales en écologie est due à Greig-Smith (1952). Greig-Smith (1952) recommande d'effectuer un échantillonnage par quadrats contigus au lieu d'un échantillonnage par quadrats aléatoires, *i.e.* sans aucun relevé de l'information spatiale (coordonnées des quadrats). Bien que Greig-Smith (1952) continue à s'intéresser à la discrimination classique des distributions spatiales (aléatoires, régulières ou agrégées), l'utilisation de quadrats contigus a pour conséquence :

- de couvrir le domaine d'étude de façon continue,
- d'introduire implicitement l'information spatiale — au moins topologique — à travers les relations de voisinage entre quadrats,
- de permettre le calcul de la variance à différentes échelles spatiales par agrégation successive des quadrats contigus en quadrats plus grands,
- d'identifier approximativement la taille d'un *patch* dans le cas d'une distribution spatiale de type agrégé.

Greig-Smith (1952) propose d'effectuer l'analyse de variance hiérarchique associée au regroupement successif des quadrats contigus, mais cette pratique n'est pas sans poser quelques difficultés statistiques (revues dans Pielou 1969, pp. 104-106, Chessel 1978, Ripley 1981, pp. 108-112). Le test proposé ultérieurement par Mead (1974) ne semble pas non plus exempt de défaut (Upton 1984).

En pratique, la variance est représentée graphiquement en fonction du niveau d'agrégation, la présence d'extrema locaux traduisant des changements de la structure spatiale. Afin de pallier les inconvénients de l'approche de Greig-Smith (1952), d'autres méthodes ont été proposées, notamment par M.O. Hill (1973), Zahl (1974), Goodall (1974) et Ludwig & Goodall (1978) (revues dans Turner *et al.* 1991, pp. 19-26, Ver Hoef *et al.* 1993). Ces propositions ont eu notamment pour conséquence de représenter la variabilité, non plus en fonction du niveau d'agrégation des quadrats, mais en fonction de la distance entre quadrats, conduisant ainsi à des méthodes spatialement explicites.

1.2.3 Méthodes spatiales explicites

L'analyse de l'autocorrélation spatiale a été transférée depuis la géographie (Cliff & Ord 1973), d'abord vers la géologie (Henley 1976), puis vers l'écologie (Jumars *et al.* 1977, Sokal & Oden 1978a, 1978b), soit pratiquement à la même époque que les propositions de méthodes décomposant la variabilité spatiale en fonction de la distance. Ces deux types d'analyses, pratiquement équivalents, conduisent à une description des données beaucoup plus pertinente que ne le permettent les indices, mais ne constituent toutefois qu'une étape intermédiaire vers un traitement exhaustif de l'information disponible.

Dans cette perspective, l'avènement des ordinateurs a eu un impact profond sur l'étude des structures spatiales. En effet, la révolution informatique a contribué au déplacement des questions élémentaires telles que "est-ce que cette distribution spatiale est aléatoire, régulière ou agrégée? " vers des résumés plus complets des données et l'ajustement de modèles stochastiques (Ripley 1988b). Autrement dit, l'informatique permet, non seulement d'analyser de plus en plus finement les structures spatiales, mais également de manipuler des modèles opérationnels concernant ces structures. Dans le cadre de cette révolution méthodologique, l'objectif de notre travail est d'illustrer l'apport de deux domaines directement concernés par le traitement des variables régionalisées : la *géomatique* et la *géostatistique*.

1.2.3.1 La géomatique

Les valeurs d'une VR sont mesurées ou observées sur des domaines bornés nommés *supports*. La définition des supports conditionne en partie les propriétés de la VR correspondante. En conséquence, l'étude du traitement des VR écologiques ne peut pas faire l'économie de considérations concernant le traitement des supports eux-mêmes, ce qui implique de recourir à la géomatique.

La *géomatique* peut être définie comme *"l'ensemble des applications de l'informatique au traitement des données géographiques, et en particulier à la cartographie"* (Mathieu 1990). Par ailleurs, l'informatique peut être vue comme la *"science de toutes les sciences"* puisque toute science peut lui déléguer le traitement de l'information qu'elle

manipule (Le Moigne 1984). Selon ce point de vue, la géomatique est alors la science du traitement de l'information spatiale de "toutes les sciences", et en particulier de l'écologie.

Les réalisations les plus connues de la géomatique sont évidemment les SIG (Systèmes d'Information Géographique) que l'on peut définir comme des systèmes informatiques permettant la saisie, le stockage, la gestion, l'analyse et la représentation graphique de l'information localisée (revues dans Burrough 1986, Tomlin 1990, Laurini & Thompson 1992, Collet 1992). Les SIG sont de plus en plus mentionnés en écologie (revues dans Johnson 1990, Haslett 1990, Caloz & Collet 1997), en entomologie (revue dans Liebhold *et al.* 1993), en agronomie (revue dans Petersen *et al.* 1995), ainsi qu'en épidémiologie (revue dans Croner *et al.* 1996).

Nous ne considérons pas ici l'analyse spatiale telle qu'elle est envisagée dans les SIG, notamment par modélisation multi-critère ou par analyse de proximité (*cf.* Tomlin 1990, Walker 1996), mais plus fondamentalement, nous présentons certains concepts, objets et opérateurs de la géomatique qui permettent de définir de nouveaux modes de traitement des VR écologiques.

1.2.3.2 La géostatistique

La randomisation spatiale dans les dispositifs de quadrats, de points ou de transects utilisés en écologie statistique est suffisante pour assurer l'existence d'estimateurs non biaisés, quelle que soit la distribution spatiale des organismes ou la variabilité spatiale des facteurs abiotiques étudiés (Robson 1982). Cette pratique soulève cependant deux problèmes :

- une véritable randomisation est généralement difficile à assurer sur le terrain,
- les estimations sont peu efficaces parce qu'elles n'exploitent pas l'information concernant la structure spatiale étudiée.

En géologie, il est souvent impossible d'obtenir des échantillons aléatoires : les roches sont collectées où elles se trouvent, où elles sont exposées et accessibles (Watson 1983). En particulier dans le domaine minier, les échantillons ne peuvent pas être sélectionnés librement (Laslett 1997), de sorte que le problème crucial posé par les échantillons non aléatoires a motivé le développement d'une théorie connue sous le nom de *géostatistique* ou *théorie des variables régionalisées* (Matheron 1963, 1965). Le terme de *géostatistique* apparu en 1962 (Chauvet 1994) désigne actuellement, en toute rigueur, essentiellement le recours à la théorie des fonctions aléatoires pour modéliser les VR (Journel 1983b). La géostatistique peut s'appliquer à la géologie, à la pédologie, à l'agronomie, à la foresterie, à l'épidémiologie, et plus généralement, à n'importe quelle discipline manipulant des données localisées dans l'espace et nécessitant des modèles décrivant la dépendance spatiale entre ces données (Cressie 1988a).

En fait, une méthodologie équivalente à la géostatistique s'est élaborée indépendamment dans le domaine de la foresterie grâce à Matérn (1960). Il n'est donc pas surprenant que, dès le début des années 1970, la théorie des variables régionalisées de Matheron (1965) ait fait l'objet d'une attention toute particulière de la part de biométriciens forestiers français (*cf.* les références citées dans Houllier 1992). Au niveau international, le transfert de la géostatistique s'est effectué d'abord vers la pédologie, au début des années

1980 (Webster & Burgess 1980a, 1980b, Burgess & Webster 1980), puis vers l'écologie, à la fin des années 1980 (Robertson 1987). Aujourd'hui, la géostatistique est utilisée dans de nombreux domaines tels que l'écologie des populations (*e.g.*, Webster & Boag 1992a, 1992b, Liebhold *et al.* 1993, Wallace & Hawkins 1994, Rossi *et al.* 1995, Franceshini *et al.* 1997, Cannavacciuolo *et al.* 1998), la génétique des populations (*e.g.*, Le Corre *et al.* 1998), la biogéographie (*e.g.*, Villard & Maurer 1996), l'halieutique (*e.g.*, Guillard *et al.* 1992, Simard *et al.* 1993, Petitgas 1993), la foresterie (*e.g.*, Höck *et al.* 1993, Kohl & Gertner 1997), la phytopathologie (*e.g.*, Chellemi 1988, Lecoustre *et al.* 1989, Nelson *et al.* 1994, 1999), l'agriculture de précision (*e.g.*, Donald 1994) et même l'écologie microbienne (revue dans Brockman & Murray 1997). Dans toutes ces disciplines, le terme *géostatistique* est utilisé essentiellement pour désigner un ensemble d'outils descriptifs (*geostatistics*) plutôt que l'application de la théorie des fonctions aléatoires.

Notre objectif est d'examiner l'apport réel de la géostatistique (au sens large) dans le traitement des VR écologiques, ce qui nécessite de replacer la géostatistique dans un contexte statistique général, et d'aborder un ensemble de problématiques essentielles en écologie statistique.

1.3 Plan de la thèse

Les trois premiers chapitres sont destinés à fournir les éléments fondamentaux nécessaires à la compréhension des chapitres suivants. Le Chapitre 2 introduit les objets et les opérateurs géomatiques qui sont mentionnés tout au long de la thèse. Ce chapitre traite spécifiquement des supports des valeurs des VR, de leurs propriétés géométriques, de leurs relations topologiques et métriques, et de leur modélisation numérique. Le Chapitre 3 introduit le concept d'autocorrélation spatiale, propriété qui émerge de la mise en relation des valeurs d'une VR avec la localisation de ses supports. Les principales définitions opératoires et les principaux tests de l'autocorrélation spatiale sont détaillés. Les notions de corrélation et de décorrélation spatiales sont introduites afin de mieux faire comprendre la nature intime de l'autocorrélation spatiale. Le Chapitre 4 expose les éléments essentiels de la géostatistique, notamment la modélisation des VR par des fonctions aléatoires (FA), la simulation numérique des FA, le concept de régularisation et la notion de variance d'erreur d'estimation.

A partir du Chapitre 5, les données sont vues comme résultant de l'échantillonnage spatial d'une population. Le recours aux FA est replacé dans le contexte statistique général des modèles de superpopulations. La distinction entre l'inférence *design-based* et l'inférence *model-based* est introduite, et le concept de représentativité est discuté. Incidemment, les notions de dépendance spatiale et de dépendance statistique sont explicitées. Enfin, le thème de l'efficacité de l'échantillonnage spatial est examiné.

La première moitié du Chapitre 6 est entièrement consacrée à l'estimation spatiale, locale ou globale, et particulièrement à la méthode du krigeage. La seconde moitié de ce chapitre aborde le problème de la précision des estimations spatiales, essentiellement des estimations globales. Le Chapitre 7 traite de l'estimation, de la modélisation et de la précision du variogramme. En s'appuyant sur les Chapitres 4 à 7, le Chapitre 8 aborde les aspects techniques de l'optimisation combinatoire de l'échantillonnage en vue de l'estimation spatiale globale, locale, et de l'estimation du variogramme.

Les Chapitres 9 et 10 explorent le problème de la mesure et du test de l'association entre deux VR. Le Chapitre 9 se place dans un cadre a-spatial classique et étudie le problème statistique posé par le test de la corrélation entre deux VR autocorrélées. Le Chapitre 10 introduit la notion de corégionalisation et aborde le problème de la mesure de l'association spatiale entre cartes binaires, puis entre cartes quantitatives.

Enfin, le Chapitre 11 est entièrement consacré à la mesure de la complexité spatiale des cartes choroplèthes, des cartes isoplèthes et des images.

1.4 Terminologie

Bien que les concepts et les notations soient définis dans le corps du texte, au fur et à mesure de leur introduction, et bien que les nombreuses abréviations employées soient regroupées dans l'Annexe A, il convient d'apporter certaines précisions quant à la terminologie adoptée dans ce mémoire, afin d'éviter tout malentendu.

Le terme *phénomène régionalisé* désigne tout phénomène se déroulant dans un espace, en l'occurrence, dans l'espace géographique. Le terme *variable régionalisée* désigne au sens strict une fonction prenant des valeurs définies en chaque point de l'espace (Matheron 1963). Dans ce mémoire, nous définissons une variable régionalisée comme une variable, de structure algébrique quelconque (quantitative, ordinale, nominale), prenant ses valeurs sur un ensemble de supports, fini ou infini.

Par abus de langage, le terme *autocorrélation spatiale* est utilisé comme un synonyme de *dépendance spatiale*. Les termes *hétérogénéité spatiale*, *variabilité spatiale* et *complexité spatiale* sont considérés comme pratiquement synonymes, et nous utilisons l'un ou l'autre selon le contexte.

Le terme *distribution* pouvant prêter à confusion, nous précisons systématiquement s'il s'agit de la *distribution spatiale* ou de la *distribution statistique* d'un ensemble (objets, événements, valeurs).

Le terme *population* est utilisé au sens d'ensemble (de supports, de valeurs) soumis à l'analyse et à la modélisation et ne doit pas être confondu avec le concept de *population biologique* (*cf.* Legay & Debouzie 1985).

Bien qu'il faille en toute rigueur considérer plusieurs types d'erreurs au cours du traitement des données spatiales (*cf.* Matérn 1960, pp. 55-56, Arbia 1993, Caloz & Collet 1997), nous considérons uniquement l'erreur entre la valeur réelle d'une grandeur (*e.g.*, la moyenne d'une population) et sa valeur estimée ou prédite (*e.g.*, la moyenne d'un échantillon), et pas les erreurs de mesure des valeurs, ni les erreurs de localisation des supports.

Le terme *échelle* est utilisé au sens que lui donne l'écologie du paysage (*cf.* Turner et Gardner 1991, p. 7) et non pas au sens d'*échelle cartographique* (*cf.* Mathieu 1990, p. 42). Ainsi, une étude écologique à petite échelle correspond à une grande échelle cartographique (*e.g.*, 1: 25 000) et inversement, une étude à grande échelle correspond à une petite échelle cartographique (*e.g.*, 1: 200 000). L'échelle est définie à la fois par la taille du domaine d'étude et la *résolution spatiale* des données, autrement dit, la taille des supports des valeurs. La résolution spatiale augmente lorsque la taille des supports diminue.

Le terme *patch* désigne une région relativement homogène vis-à-vis d'une certaine variable, assez nettement délimitée dans l'espace, *e.g.* une zone de valeurs élevées entourée de valeurs faibles.

Le terme *pattern* désigne à l'origine la structure spatiale qui résulte de la distribution des organismes qui interagissent entre eux et avec leur environnement (Hutchinson 1953). Hutchinson (1953) distingue cinq sortes de *patterns* :

- vectoriels, lorsque la distribution des organismes est déterminée par des forces externes, gradients de température, de lumière, d'humidité, courants, vents, etc.,
- reproducteurs, lorsque la distribution est déterminée par le fait que la progéniture reste proche des parents,
- sociaux, lorsque des signaux entraînent l'agrégation ou l'évitement,
- coactifs, résultant de l'interaction entre espèces en compétition pour une ou plusieurs ressources,
- stochastiques, en invoquant des forces aléatoires.

Le dernier type de *pattern* constitue une catégorie fourre-tout, qui sert à classer tous les phénomènes dont l'origine nous échappe — du moins à un moment donné — et qui sont justifiables d'une modélisation stochastique plutôt que déterministe[1]. Comme nous traitons de la distribution spatiale des organismes et de la variabilité spatiale des facteurs environnementaux dans un même cadre méthodologique, d'un point de vue strictement descriptif et statique, nous ne parlons pas de *pattern* au sens d'Hutchinson (1953), mais plus généralement de *structures spatiales*, en évacuant d'emblée la question de leur genèse.

[1] Comme le fait remarquer Matheron (1978), il s'agit là essentiellement d'une décision méthodologique, car il n'y a pas de probabilités dans la nature, il n'y a que des modèles probabilistes.

Chapitre 2

Modèles géomatiques

> *"Although these investigations are not concerned with the stochastic processes sampled, they are intimately connected with the geometric questions that seem to emerge in almost all discussions of planar and spatial processes"* (Matérn 1960)
>
> *"Further study of the analysis of graph theory* [...] *might lead to fruitful applications of graph theoretic techniques in the analysis of geographic regions"* (Gabriel & Sokal 1969)

En ne tenant pas compte de la dimension temporelle des phénomènes écologiques, l'espace dans lequel sont localisées les valeurs d'une VR est l'espace géographique, *i.e.* la surface de la Terre ou *géoïde*[1]. Les valeurs d'une VR sont repérées sur une surface de référence qui représente, localement ou globalement, une bonne approximation du géoïde, *e.g.* l'ellipsoïde de Clark pour le territoire français métropolitain. En pratique, une étude écologique est nécessairement restreinte à un domaine du géoïde que nous notons D. Lorsque D est suffisamment petit, il est possible de l'assimiler directement à un domaine borné du plan, soit $D \subset \mathbb{R}^2$. Si D est trop vaste pour que l'approximation précédente soit licite, et afin de simplifier toutes les opérations géométriques, il est nécessaire de recourir à une projection de l'ellipsoïde de référence sur le plan, cette opération n'étant évidemment pas sans conséquences sur la validité des résultats des traitements ultérieurs. Pour simplifier, nous considérons donc désormais que la VR est définie sur $D \subset \mathbb{R}^2$.

Quelle que soit la VR étudiée, un ensemble de valeurs $\{z_i \mid i = 1, \ldots, n\}$ est associé à un ensemble de supports $\{v_i \mid \ i = 1, \ldots, n\}$ inclus dans D. Par conséquent, le traitement des VR implique nécessairement le traitement des supports, *i.e.* le recours à la *géomatique*. La géomatique désigne un domaine cohérent qui regroupe toutes les techniques informatiques de modélisation numérique ou symbolique des objets spatiaux, ainsi que tous les traitements informatiques qui opèrent sur ces modèles. La géomatique ne constitue donc pas une discipline autonome mais une spécialisation de l'informatique. La géomatique fait appel à des domaines aussi divers que la géométrie algorithmique (Lee & Preparata 1984, Preparata & Shamos 1985), la géométrie discrète (Chassery & Montanvert 1991), la théorie des graphes et des hypergraphes (Berge 1970), les techniques de conception et de gestion des bases de données (Bouillé 1977, David 1991, Laurini & Milleret-Raffort 1993), les techniques d'indexation spatiale (Ooi 1990, Samet 1990, Li 1992), la sémiologie

[1] Le *géoïde* est défini comme une surface de niveau de la pesanteur terrestre (Mathieu 1990).

(Bertin 1977, Vieu 1991), les méthodes de l'Intelligence Artificielle (Titeux 1989, Chun 1990), etc. Dans ce qui suit, nous nous contentons de définir :

- les objets fondamentaux manipulés par la géomatique, *i.e.* dans notre contexte, les différents types de supports d'une VR,
- quelques opérateurs fondamentaux qui sont mentionnés au cours des chapitres suivants,
- les modèles de cartographie.

2.1 Objets fondamentaux

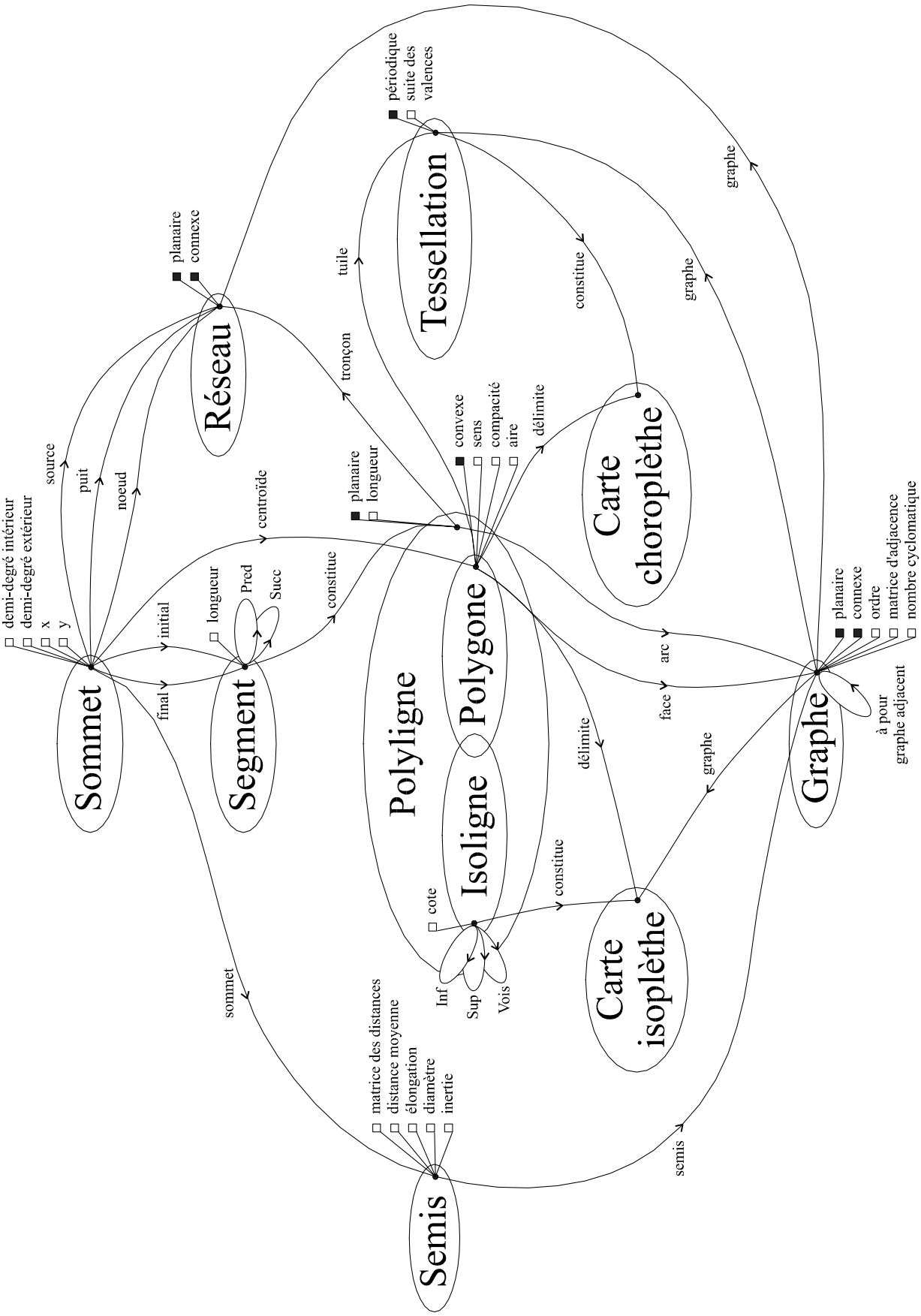
Les objets fondamentaux de la géomatique ne sont pas autre chose que des objets mathématiques tels que les points, les segments de droite, les polygones ou les graphes, mais qui sont considérés selon un point de vue opératoire, *i.e.* informatique. Ces objets peuvent tous être décomposés sous la forme de vecteurs, ce qui conduit à parler d'*objets vectoriels* et de géomatique en *mode vecteur*.

Les objets vectoriels peuvent être répartis en groupes nommés *types abstraits de données* ou *classes* selon la terminologie informatique utilisée, l'idée sous-jacente étant que ces groupes sont homogènes quant aux propriétés des objets qui les composent. Ces classes ne sont pas indépendantes mais entretiennent des liens les unes avec les autres. Afin de représenter de façon synthétique — si ce n'est exhaustive — les classes et leurs relations mutuelles, il convient d'utiliser un modèle graphique de structuration des connaissances. Parmi les modèles de structuration des connaissances utilisés en géomatique (Laurini & Milleret-Raffort 1989, pp. 17-24, Laurini & Thompson 1992, pp. 444-476) le modèle HBDS (*Hypergraph Based Data Structure*) (Bouillé 1977) se révèle à la fois :

- parfaitement adapté, puisqu'il est utilisé couramment à l'IGN (Institut Géographique National) (Pressensé & Salgé 1987),
- conceptuellement très riche, puisqu'il fait référence à la théorie des hypergraphes de Berge (Bouillé 1977) et à la théorie des catégories de Eilenberg & Mac Lane (Cousin 1988).

Le modèle orienté-objet HBDS est constitué de classes d'objets munies d'*attributs* — *i.e.* des types de propriétés partagées par tous les objets d'une même classe — et qui entretiennent entre elles des *liens*. L'inclusion d'une classe dans une autre correspond à une spécialisation, la sous-classe héritant des attributs de la classe supérieure : il s'agit d'*héritage simple*. Par ailleurs, une classe C qui résulte de l'intersection de deux classes A et B hérite à la fois des attributs de A et de ceux de B : il s'agit alors d'*héritage multiple*. Les concepts informatiques de classe, d'objet, d'attribut, d'héritages simple et multiple sont exposés notamment dans Masini *et al.* (1990), Delobel *et al.* (1991), Benzaken & Doucet (1993), Bouché (1994) et Larvet (1994), et ne sont pas davantage développés dans ce mémoire.

En complément de la structure HBDS que nous proposons (Fig. 2.1), nous définissons succinctement les classes d'objets vectoriels, en distinguant des *objets élémentaires* et des *objets composés*. Bien entendu, d'autres présentations peuvent être envisagées (*e.g.*, David 1991, Laurini & Milleret-Raffort 1993, pp. 143-175).



Figure˜2.1: Modèle HBDS (*Hypergraph Based Data Structure*) des classes des objets fondamentaux de la géomatique en mode vecteur. Tous les liens entre classes admettent des liens réciproques, non figurés. Un carré noir correspond à un attribut booléen.

2.1.1 Objets élémentaires

Les objets élémentaires peuvent être classés en fonction de leur dimension topologique, ce qui conduit à distinguer des objets :

- ponctuels, de dimension 0,
- linéaires, de dimension 1, *i.e.* séparés en deux parties par suppression d'un point,
- surfaciques, de dimension 2, *i.e.* séparés en deux parties par suppression d'une ligne.

2.1.1.1 Objets ponctuels

Les objets ponctuels sont modélisés par des points. Dans le plan, un *point* ou *sommet* s est généralement repéré par ses coordonnées cartésiennes $(x(s), y(s))$ bien que d'autres systèmes puissent être employés (*e.g.*, coordonnées polaires). Un sommet isolé ne possède aucun attribut morphologique, métrique, ou topologique.

2.1.1.2 Objets linéaires

Les objets linéaires sont modélisés par des segments et des polylignes. Un *segment* $s_1 \leftrightarrow s_2$ est défini par ses deux sommets extrêmes s_1 et s_2. Un segment peut être muni d'une orientation, ce qui conduit à distinguer le segment $s_1 \rightarrow s_2$ du segment $s_2 \rightarrow s_1$ bien que la géométrie soit la même. L'attribut essentiel d'un segment est sa longueur. Une *polyligne* $s_1 \leftrightsquigarrow s_n$ d'extrémités s_1 et s_n est décrite par une suite de sommets $(s_1, s_2, \ldots, s_n)$ tels que $s_i \leftrightarrow s_{i+1}$ avec $i = 1, \ldots, n-1$. Une polyligne est orientée si tous les segments qui la compose sont identiquement orientés, soit $s_1 \rightsquigarrow s_n$ si $s_i \rightarrow s_{i+1}$ pour $i = 1, \ldots, n-1$, avec s_1 le *sommet initial* et s_n le *sommet final*. Une polyligne possède une longueur et un attribut topologique de *planarité*[2]. Une polyligne est dite *planaire* ou *simple* si elle n'est pas en intersection avec elle-même en dehors de ses extrémités. En écologie, la trajectoire d'un animal constitue un exemple de polyligne, éventuellement planaire (*e.g.*, Perry 1995, Fig. 5e), mais plus généralement non planaire (*e.g.*, Blanché *et al.* 1996, Fig. 10 & 11, Anderson *et al.* 1997, Fig. 3).

2.1.1.3 Objets surfaciques

Les objets surfaciques sont modélisés par des polygones. Un *polygone* $P = s_1 \circlearrowleft$ est décrit par une polyligne $s_1 \leftrightsquigarrow s_n$ fermée sur elle-même, *i.e.* telle que $s_1 \equiv s_n$ avec s_1 le premier sommet de la description. Pour une même suite de sommets $(s_1, s_2, \ldots, s_{n-1})$ d'un polygone P, il est possible de distinguer $n-1$ descriptions différentes $s_i \circlearrowleft$ avec $i = 1, \ldots, n-1$, bien que la géométrie soit la même. De même que pour une polyligne, un polygone est orienté si tous les segments qui le composent sont identiquement orientés, ce qui conduit à définir *un sens de rotation*. Un polygone hérite des attributs d'une polyligne mais la longueur de la polyligne qui le décrit se nomme *périmètre*. Un polygone peut être *convexe* ou *concave*.

[2] A la différence de David (1991, p. 53), nous n'utilisons pas le terme *planéité* qui désigne en fait le caractère d'une surface plane.

En vertu du théorème de Jordan, un polygone simple P sépare le plan en deux composantes connexes disjointes, *i.e.* l'*intérieur* et l'*extérieur* du polygone (Chassery & Montanvert 1991, p. 35, Voĭtsekhovskiĭ 1995). Notons $\overset{\circ}{P}$ l'intérieur de P, $\overline{P}$ son adhérence et $\complement P$ son complémentaire dans le plan ou extérieur[3] de P. La frontière de P peut se noter $\partial P = \overline{P} \cap \complement P$ ou encore $\partial P = \overline{P} - \overset{\circ}{P}$ (Choquet 1992, p. 16). L'*aire* de P est un attribut quantifiant son intérieur $\overset{\circ}{P}$ ou ce qui revient au même, son adhérence $\overline{P}$.

Il est possible d'associer un *centroïde* à un polygone P, *i.e.* un sommet $s_c \in \overset{\circ}{P}$ le plus éloigné possible de ∂P. Si P est convexe, s_c est défini comme le barycentre des sommets décrivant ∂P, en revanche, si P est concave, son barycentre n'appartient pas nécessairement à son intérieur de sorte que plusieurs définitions opératoires du centroïde s_c peuvent être envisagées (*e.g.*, Laurini & Thompson 1992, pp. 269-270).

A notre connaissance, la fonction densité de probabilité des distances entre points aléatoires situés à l'intérieur du polygone P est un attribut qui n'est jamais considéré en géomatique. L'importance de cet attribut lorsque P est le support d'une VR apparaîtra dans la Section 4.6.3.1.

2.1.2 Objets composés

Les objets composés intègrent plusieurs objets élémentaires, cette intégration conduisant à l'émergence de nouveaux attributs. Nous identifions trois classes fondamentales d'objets composés obtenus directement à partir des objets ponctuels, linéaires et surfaciques :

- les semis,
- les réseaux,
- les tessellations.

Chacune de ces trois classes peut être vue comme un cas particulier de la classe des graphes, le semis constituant toutefois un cas limite.

2.1.2.1 Semis

Un *semis* est un ensemble de n sommets distincts $S = \{s_i \mid i \in I\}$, avec $I = \{1, \ldots, n\}$. A partir des n sommets il est possible de calculer $\binom{n}{2} = n(n-1)/2$ distances $d_{ij} = d(s_i, s_j)$, avec $i < j \in I$. Un semis S peut être caractérisé, notamment, par sa matrice de distances $\mathbf{D} = ((d_{ij}))$. $\mathbf{D}$ peut être synthétisée par la distance moyenne :

$$\bar{d} = \frac{2}{n(n-1)} \sum_{i<j} d_{ij} \tag{2.1}$$

la distance minimale :

$$d_{\min} = \min_{i,j \in I} d_{ij} \tag{2.2}$$

[3]Les définitions de ces notions de topologie générale peuvent se trouver notamment dans Dixmier (1981, pp. 17-20) et Choquet (1992, pp. 13-16).

et la distance maximale ou *diamètre* :

$$d_{\max} = \max_{i,j \in I} d_{ij} \tag{2.3}$$

Des attributs de type morphologique peuvent être définis, tels que l'inertie du semis, son élongation, etc.

2.1.2.2 Réseaux

Soit trois polylignes ayant une extrémité commune $a \leftrightsquigarrow d$, $b \leftrightsquigarrow d$ et $c \leftrightsquigarrow d$; nous dirons que a, b et c sont des *sommets pendants*, que d est un *noeud*, et que les trois isolignes sont trois *tronçons* d'un *réseau*. Un réseau est donc constitué d'un ensemble de polylignes en contact par leurs extrémités. En parcourant le réseau, il est évidemment possible de définir d'autres polylignes que celles qui constituent les tronçons. Lorsque les tronçons sont orientés, alors il est possible de distinguer les sommets pendants qui correspondent à des sommets initiaux ou *sources*, et ceux qui correspondent à des sommets finaux ou *puits*. En généralisant l'attribut de planarité des polylignes, un réseau est qualifié de *planaire* s'il n'est pas en intersection avec lui-même.

2.1.2.3 Tessellations

Une *tessellation* $T = \{P_i \mid i \in I\}$ d'un domaine $D \subset \mathbb{R}^2$ est une partition de D en n polygones P_i nommés *tuiles* ou *tesselles*. Formellement, T est une famille finie $(P_i)_{i \in I}$ de parties de D qui vérifie (Arnold & Guessarian 1993) :

$$\overset{\circ}{P}_i \neq \emptyset \;\; \forall i \in I, \;\; \overset{\circ}{P}_i \cap \overset{\circ}{P}_j = \emptyset \;\; \forall i \neq j \in I, \;\; D = \bigcup_{i \in I} P_i \tag{2.4}$$

Autrement dit, une tessellation de D est constituée d'un ensemble de polygones non chevauchants qui couvrent entièrement D. On désigne par *tuile type* P une tuile dont plusieurs occurrences participent à T. Une tessellation peut comporter $m = 0$, $m = 1$ ou $m > 1$ tuiles types P.

Le cas $m = 0$ correspond aux tessellations irrégulières caractéristiques des phénomènes naturels organisés selon une *mosaïque* (*cf.* Pielou 1969, Chiarello 1994).

Le cas $m = 1$ correspond à un *pavage* au sens strict, et la tessellation T peut être définie par la géométrie de P ainsi qu'un groupe d'isométries[4] de P, *i.e.* une façon de reproduire P pour tesseller D. Imposons que P soit convexe. Soit r le nombre de sommets d'une tuile qui appartiennent à au moins 3 tuiles distinctes ; on montre que $r \geq 3$ et ne dépend par de P (Berger 1979). En tournant le long de la frontière ∂P, les sommets correspondants sont $s_1, \ldots, s_r$ et α_i est le nombre de tuiles contenant le sommet s_i. On montre que la *suite des valences* d'une tessellation périodique $(\alpha_i \mid i = 1, \ldots, r)$ ne dépend pas — à un renversement et à une permutation circulaire près — de la tuile P, et qu'elle satisfait à la relation (Berger 1979, pp. 53, Chassery & Montanvert 1991, pp. 29) :

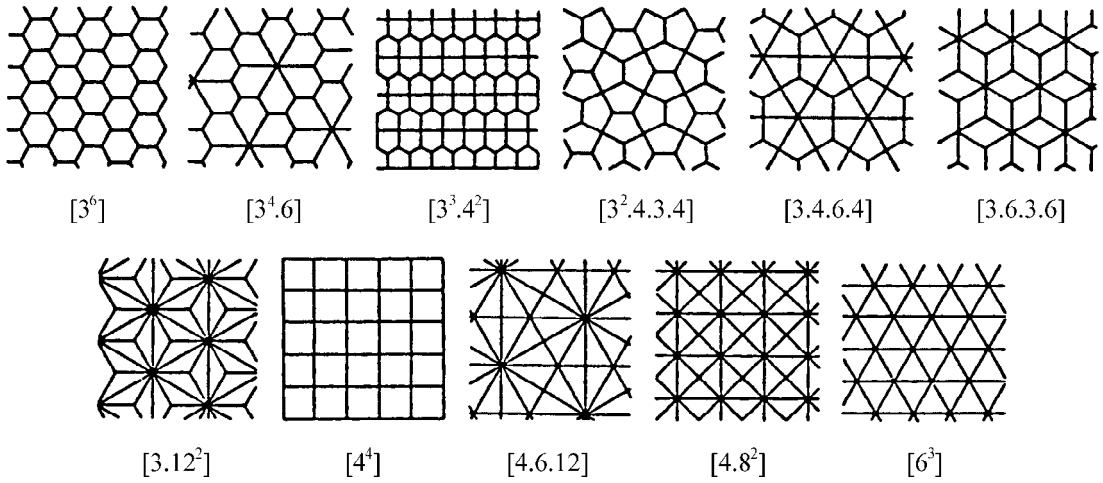
$$\frac{r}{2} - 1 = \sum_{i=1}^{r} \frac{1}{\alpha_i} \tag{2.5}$$

[4]Les *groupes des paveurs* sont décrits notamment dans Berger (1979, pp. 33-54), Sénéchal (1979, pp. 98-115) et Lord & Wilson (1986, pp. 152-155).

Il n'existe que 21 suites de valences[5] qui satisfont à (2.5) (Chassery et Montanvert 1991). Seules 11 suites de valences donnent lieu à des tessellations par des polygones convexes identiques, nommées *tessellations semi-régulières* (Pasquier 1987, Chassery & Montanvert 1991) ou *tessellations de Shubnikov-Laves* (Ivanov 1995). Notons la suite (a, a, a, b, b, c) sous la forme abrégée $[a^3.b^2.c]$; les tessellations semi-régulières sont définies par les suites de valences $[3^6]$, $[3^4.6]$, $[3^3.4^2]$, $[3^2.4.3.4]$, $[3.4.6.4]$, $[3.6.3.6]$, $[3.12^2]$, $[4^4]$, $[4.6.12]$, $[4.8^2]$ et $[6^3]$ (Fig. 2.2). Une restriction supplémentaire consiste à imposer que P soit un polygone régulier, ce qui implique $\alpha_i = \alpha$, $\forall i \in \{1, \ldots, r\}$. On montre que α et r sont liés par la relation (Ore 1970, pp. 116-118) :

$$(\alpha - 2)(r - 2) = 4 \tag{2.6}$$

cette relation étant bien conforme à (2.5) pour $\alpha_i = \alpha$, $\forall i \in \{1, \ldots, r\}$. Les seuls couples d'entiers qui satisfont (2.6) sont $(\alpha = 3, r = 6)$, $(\alpha = 4, r = 4)$ et $(\alpha = 6, r = 3)$, d'où l'on déduit que P est un hexagone, un carré ou un triangle équilatéral. Les tessellations qui correspondent aux suites de valences $[\alpha^r]$, *i.e.* $[3^6]$, $[4^4]$ et $[6^3]$ sont dites *régulières* (Pasquier 1987, Chassery & Montanvert 1991).



Figure~2.2: Les onze tessellations semi-régulières possibles et leurs suites de valences (d'après Pasquier 1987).

[5] D'après Berger (1979) il existerait 23 suites de valences possibles.

2.1.2.4 Graphes

Un *graphe* $G = (S, A)$ est un couple constitué par un ensemble de *sommets* $S = \{s_i \mid i \in I\}$ et par une famille $A = (a_i \mid i = 1, \ldots, m)$ d'éléments du produit cartésien $S \times S = \{(s_i, s_j) \mid i, j \in I\}$. L'élément (s_i, s_j) peut apparaître plusieurs fois dans la famille A : s'il apparaît au plus p fois, alors G est un *p-graphe* (Berge 1970). Le nombre de sommets du graphe G est appelé l'*ordre* de G. Un élément a_i de la famille A est une *arête* du graphe G.

Si le couple de sommets de toute arête a_i est orienté, alors il est possible de distinguer un *sommet initial* $\mathrm{ini}(a_i)$ et un *sommet final* $\mathrm{fin}(a_i)$. Dans ce cas, a_i est un *arc* de G, et G est qualifié de *graphe orienté*. Une arête peut être vue comme un arc dont on ignore l'orientation de sorte que tout graphe peut être considéré comme fondamentalement orienté[6] (Berge 1970). Chaque sommet s de G est caractérisé par ses demi-degrés extérieur $d_G^+(s)$ et intérieur $d_G^-(x)$, et par son degré $d_G(x) = d_G^+(x) + d_G^-(x)$ (Berge 1970). Le *demi-degré extérieur* $d_G^+(x)$ représente le nombre d'arcs incidents à s vers l'extérieur (*i.e.* ayant s comme sommet initial). On définit de même le *demi-degré intérieur* $d_G^-(x)$ pour les arcs incidents à s vers l'intérieur (*i.e.* ayant s comme sommet final).

Une *chaîne* de longueur q est une séquence d'arcs $(a_i \mid i = 1, \ldots, q)$ telle que chaque arc ait une extrémité en commun avec l'arc précédent, et l'autre extrémité en commun avec l'arc suivant. Un *cycle* est une chaîne telle que le même arc ne figure pas deux fois dans la séquence, et dont les sommets aux deux extrémités coïncident. Un *chemin* est une chaîne particulière pour laquelle $\mathrm{fin}(a_i) \equiv \mathrm{ini}(a_{i+1})$ avec $i = 1, \ldots, q-1$. Dans le cas d'un 1-graphe, un chemin est entièrement déterminé par la succession des sommets $\mathrm{ini}(a_1), \mathrm{fin}(a_1), \mathrm{fin}(a_2), \ldots, \mathrm{fin}(a_q)$. L'*écart* $d(s_i, s_j)$ entre deux sommets s_i et s_j est la longueur du plus court chemin entre s_i et s_j. L'*écartement* d'un sommet s_i est alors défini par (Berge 1970, p. 58) :

$$\mathrm{e}(s_i) = \max_{\substack{s_j \in S \\ s_j \neq s_i}} d(s_i, s_j) \tag{2.7}$$

Les graphes possèdent de nombreux attributs, dont deux attributs topologiques particulièrement importants de *connexité* et de *planarité*. Un graphe est dit *connexe* si, pour toute paire de sommets (s_i, s_j) avec $i \neq j$, il existe une chaîne reliant s_i et s_j. Un graphe G est dit *planaire* s'il est possible de le représenter sur un plan de sorte que ses sommets soient des points distincts, les arêtes des courbes simples, et que deux arêtes ne se rencontrent pas en dehors de leurs extrémités (Berge 1970, p. 16, Alekseev 1995).

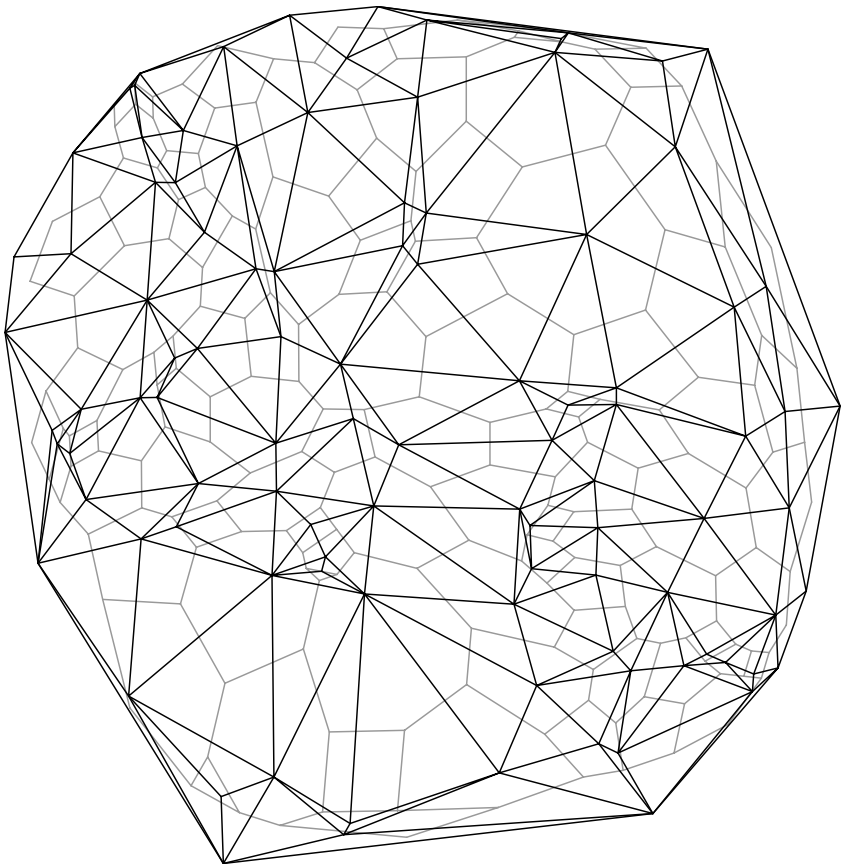
Formellement, les graphes planaires incluent les semis, les réseaux et les tessellations. En effet, un semis représente le cas limite d'un graphe dépourvu d'arête ($m = 0$), *i.e.* dont tous les sommets sont isolés les uns des autres. Un réseau est évidemment modélisé par un graphe G dont les sommets sont les extrémités des tronçons et dont les arêtes sont décrites par les tronçons. Dans le cas d'une tessellation comportant f tuiles, les sommets de G sont ceux des tuiles, les arêtes de G sont les côtés des tuiles, les intérieurs $\overset{\circ}{P}_i$ définissent les *faces* de G de *contours* ∂P_i. La frontière ∂D est le contour de G et le complémentaire $\complement D$ est la *face infinie* de G.

[6] En théorie des graphes, il s'agit d'une convention adoptée par l'Ecole française, mais l'Ecole américaine adopte pour sa part la convention inverse (Berge 1970).

Dans un *graphe planaire connexe*, les nombres de sommets n, d'arêtes m et de faces f sont reliés par la formule d'Euler :

$$n - m + f = 2 \tag{2.8}$$

Soit G un graphe planaire connexe ; associons un sommet $s^\bullet$ à chaque face f de G et à toute arête a de G faisons correspondre une arête $a^\bullet$ reliant les sommets $s^\bullet$ associés aux faces situées de part et d'autre de a dans G. Le graphe $G^\bullet$ défini pas les sommets $s^\bullet$ et les arêtes $a^\bullet$ est le graphe dual de G, tel que $(G^\bullet)^\bullet = G$ (Berge 1970, p. 20). Si aucun sommet n'est associé à la face infinie de G, il convient de ne pas parler de graphe dual mais plutôt de *graphe d'adjacence* G^* pour lequel $(G^*)^* \neq G$ (Fig. 2.3).



Figure˜2.3: Exemple de graphe G (figuré en noir) accompagné de son graphe d'adjacence G^* (figuré en gris).

Soit $A_{i\to j} = \{a \mid s_i = \mathrm{ini}\,(a)\,, s_j = \mathrm{fin}\,(a)\}$ l'ensemble des arcs ayant s_i comme sommet initial et s_j comme sommet final ; notons $m^+_G\,(s_i, s_j) = \mathrm{Card}\,(A_{i\to j})$ la *multiplicité* du couple de sommets (s_i, s_j). A tout graphe G il est possible d'associer une *matrice d'adjacence* $\mathbf{M} = ((m_{ij}))$, avec la simplification d'écriture $m_{ij} = m^+_G\,(s_i, s_j)$.

En pratique, il est courant d'associer une fonction de valuation aux sommets de S et/ou aux arêtes de A. C'est évidemment le cas lorsque les sommets sont les supports d'une VR et que les arêtes sont valuées par la distance euclidienne entre leurs extrémités. C'est également le cas lorsqu'un graphe est utilisé pour représenter les relations spatiales

entre sous-populations au sein d'une métapopulation ou entre les éléments du paysage formant une mosaïque (*i.e.*, une tessellation irrégulière). Dans ce contexte, les valuations portées par les arêtes des graphes peuvent notamment quantifier la facilité avec laquelle les individus peuvent passer d'une sous-population à une autre ou coloniser de nouveaux habitats (Buckland & Elston 1993).

Les graphes présentent l'intérêt d'introduire l'information topologique au sein des modèles de métapopulation qui ne tiennent compte généralement que des distances entre patchs d'habitat (*e.g.*, Hanski *et al.* 1994, Wahlberg *et al.* 1996). L'introduction des relations topologiques est importante parce que la connectivité entres patchs a un effet sur l'extinction des populations locales (Fahrig & Merriam 1985, Merriam 1988). L'hypothèse d'une connectivité égale entre les sous-populations étant irréaliste pour la plupart des métapopulations (Fahrig & Merriam 1985, Hanski 1994a), les graphes peuvent être utilisés afin de quantifier la connectivité entre patchs. D'un point de vue plus général, lorsqu'un paysage est modélisé par un graphe, il est important de mesurer la connectivité entre les éléments du paysage parce qu'une connectivité élevée implique davantage de possibilités de dispersion des animaux et des propagules des végétaux, et davantage de possibilités de transfert de matière et d'énergie entre éléments (Cantwell & Forman 1993). A cet effet, Gabriel & Sokal (1969) proposent de mesurer la *connectivité* d'un graphe[7] connexe par :

$$c(G) = \frac{\nu(G)}{\max \nu(G)} \tag{2.9}$$

pour $n \geq 3$, avec au numérateur le nombre cyclomatique $\nu(G) = m - n + 1$ d'un graphe connexe (Berge 1970, p. 15) et au dénominateur la valeur maximale que peut prendre $\nu(G)$. Dans le cas d'un graphe complet comportant n sommets ou *n-clique* (Berge 1970), le nombre d'arêtes qu'il est possible de former est $m = n(n-1)/2$, d'où $\max \nu(G) = (n-1)(n-2)/2$. Dans le cas général, la connectivité $c(G)$ vaut donc au maximum 1 dans le cas d'une n-clique, et au minimum 0 dans le cas d'un graphe sans cycle (ou *arbre*) pour lequel $m = n - 1$. Si G est un graphe planaire, alors $\max \nu(G) = 2n - 5$ (Gabriel & Sokal 1969). Dans ce cas, $c(G)$ atteint sa valeur maximale si G est une *triangulation* de S (Matula & Sokal 1980).

2.2 Opérateurs fondamentaux

Les procédures qui peuvent opérer sur les objets des classes du modèle HBDS que nous proposons (Fig. 2.1) sont trop nombreuses pour que nous puissions toutes les faire figurer, et *a fortiori* les décrire. Par conséquent, le but de cette section est uniquement de présenter quelques opérateurs particulièrement importants qui sont mentionnés dans les chapitres suivants, en distinguant des *opérateurs internes* et des *opérateurs externes*.

Les algorithmes correspondant aux opérateurs cités relèvent essentiellement de la géométrie analytique, du traitement des graphes et de la géométrie algorithmique. Dans ce qui suit, nous ne détaillons aucun algorithme, certains algorithmes géométriques étant très compliqués. Néanmoins, dans le cas des problèmes très bien documentés, nous faisons largement référence à la littérature.

[7] Cette définition diffère de celle de la théorie des graphes (*cf.* Berge 1970, Sapozhenko 1995).

2.2.1 Opérateurs internes

Au sein de chaque classe d'objets il est possible de définir des *opérateurs internes*. Ces opérateurs internes constituent des attributs de classes nommés *méthodes*. Il est possible de distinguer des opérateurs internes de type :

- morphologique,
- métrique,
- topologique,
- ensembliste.

Les opérateurs morphologiques calculent uniquement des attributs d'objets. En revanche, les opérateurs métriques et topologiques peuvent calculer des attributs d'objets ou renvoyer un résultat concernant la relation spatiale pour un couple d'objets. Enfin, les opérateurs ensemblistes ne calculent pas d'attributs d'objets mais peuvent générer un nouvel objet de la classe à partir d'un couple d'objets originels.

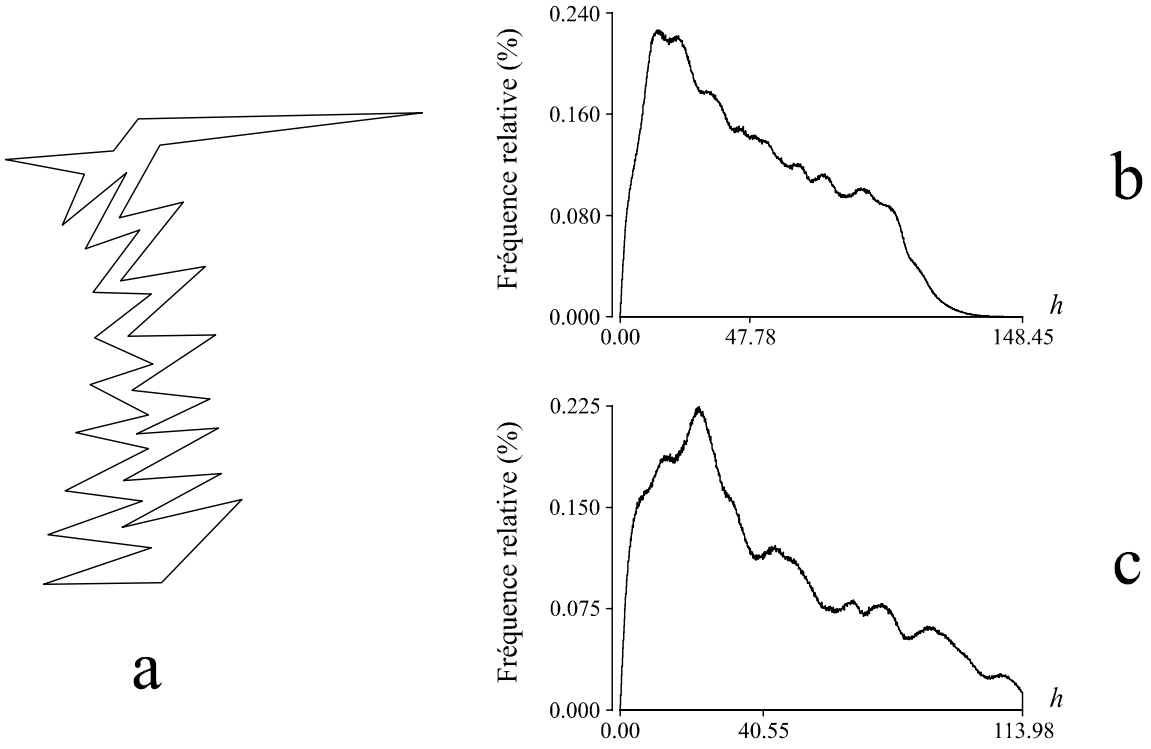
2.2.1.1 Opérateurs morphologiques

Par définition, les résultats des opérateurs morphologiques ont pour objectif de caractériser la forme de l'objet. Par exemple, la fonction densité de probabilité des distances entre points situés aléatoirement à l'intérieur d'un polygone P constitue un attribut qui est lié à la forme de P. Lorsque P est concave, il est possible de distinguer deux distributions de probabilités pour la norme des vecteurs inter-points selon que lesdits vecteurs sont autorisés ou non à traverser ∂P (*e.g.*, Fig. 2.4). Ce type d'attribut est cependant difficile à calculer par voie analytique dans le cas d'un convexe quelconque[8], et pratiquement impossible dans le cas d'un concave, ce qui nécessite de recourir à une méthode de Monte-Carlo. La définition opératoire de cette méthode de Monte-Carlo fait elle-même appel à des opérateurs géomatiques, éventuellement assez compliqués dans le cas du traitement des polygones concaves (Section 8.5.1, p. 259).

De nombreux indices morphologiques sont définis en écologie du paysage (Blackburn & Milton 1996, Gustafson 1998). Les attributs morphologiques sont également utiles dans le domaine de la biologie de la conservation où la définition optimale de la forme des réserves naturelles constitue un sujet qui reste largement débattu (Kunin 1997). Les attributs morphologiques les plus simples sont l'*élongation* et la *compacité*.

Elongation L'*élongation* d'un objet peut être définie sur la base du rapport $e = \lambda_1/\lambda_2$, avec $\lambda_1 \geq \lambda_2$ les valeurs propres de la matrice d'inertie calculée à partir des coordonnées des sommets de l'objet (Harvey 1981, Tough & Miles 1984, Postaire 1987, pp. 171-173, Tough 1988). Cette approche revient donc à effectuer une ACP (Analyse en Composantes Principales) sur les sommets de l'objet afin de déterminer les deux axes principaux du semis, puis à calculer le rapport de la part d'inertie associée à chacun d'eux. Par définition, l'élongation est infinie pour un segment puisque dans ce cas $\lambda_2 = 0$.

[8] La fonction densité de probabilité des distances entre points aléatoires est établie par Ghosh (1951), dans le cas d'un rectangle dont les côtés sont parallèles aux axes des coordonnées.



Figure˜2.4: Distributions de probabilités des distances entre points localisés au hasard à l'intérieur d'un polygone concave selon que la frontière du polygone ∂P constitue un obstacle ou pas. (a) Exemple de polygone P présentant de nombreuses concavités. (b) Distribution obtenue lorsque les vecteurs inter-points sont autorisés à traverser ∂P. (c) Distribution obtenue lorsque les vecteurs inter-points ne peuvent pas traverser ∂P.

Compacité La *compacité* d'un polygone P se définit classiquement à partir de son aire $[P]$ et de son périmètre L comme $c_1 = 4\pi\,[P]\,/L^2$ (Postaire 1987, pp. 170-171), $c_2 = L^2/4\pi\,[P]$ (Serra 1982, p. 336) ou encore $c_3 = L/\sqrt{4\pi\,[P]}$ (Rodríguez & Lewis 1997). La compacité prend sa valeur maximale $c_1 = 1$ pour un cercle parfait, et sa valeur minimale $c_1 = 0$ pour un polygone aplati, *i.e.* un segment de droite. Serra (1982, p. 336) et Lagro (1991) font remarquer que ce type d'indice ne permet pas de discriminer entre des formes très différentes. Par exemple, dans le cas d'un polygone dont le nombre de côtés tend vers l'infini — *i.e.* dont la frontière est très sinueuse — alors $L \to \infty$ et par conséquent $c_1 \to 0$ ($c_2, c_3 \to \infty$), même si P tend vers un cercle (Serra 1982, Fig. X.10, p. 336). Le recours aux outils développés par la morphologie mathématique permet de résoudre ce type de problème (Serra 1982, pp. 336-338).

2.2.1.2 Opérateurs métriques

Le résultat d'un opérateur métrique est en rapport avec la géométrie ou la position des objets. Les procédures qui calculent la longueur d'une polyligne, le périmètre ou l'aire d'un polygone sont des exemples très simples d'opérateurs déterminant les valeurs d'attributs métriques.

Les opérateurs métriques peuvent également mesurer la proximité spatiale de deux objets d'une même classe, dans un sens qui reste à définir. Le vocabulaire n'étant pas encore fixé, il est possible de parler d'éloignement ou de distance au sens large. Dans ce contexte, l'exemple d'opérateur le plus simple est évidemment le calcul de la distance euclidienne entre deux points s_1 et s_2, *i.e.* la distance en ligne droite.

D'autres mesures doivent être définies dès qu'interviennent des objets qui imposent des contraintes d'accessibilité à l'espace. Un opérateur qui mesure la distance en présence de contraintes spatiales est d'un intérêt évident en écologie parce qu'il permet de calculer des distances biologiquement réalistes. Ainsi, il est possible d'imposer que la distance entre deux points situés sur une polyligne L soit mesurée le long de L, comme ce serait le cas pour des rats musqués repérés le long d'une rivière (Le Boulengé *et al.* 1996) ou bien strictement à l'intérieur d'un domaine concave D, comme ce serait le cas pour des poissons localisés dans un estuaire (Little *et al.* 1997).

La présence d'obstacles modélisés sous la forme d'un ensemble de n polygones $P = \{P_i \mid i \in I\}$ peut également imposer de calculer la distance entre deux points s_1 et s_2 en contournant au plus près les polygones de P qui sont situés entre s_1 et s_2. De tels obstacles peuvent être des régions où la *fitness* d'un organisme est nulle ou relativement faible (Addicott *et al.* 1987), par exemple des étendues d'eau pour un insecte terrestre n'ayant pas la possibilité de voler ou pour un gastéropode terrestre. Dans le cadre de la théorie des métapopulations, ce type de considération est particulièrement important dans la modélisation de la colonisation de nouveaux habitats (Buckland & Elston 1993). D'une façon plus générale, la présence d'obstacles doit être prise en compte explicitement dans toutes les méthodes d'écologie statistique qui font référence à la distribution des distances inter-points (Switzer 1983). La distance doit alors être définie comme la longueur de la *géodésique* qui sépare les deux points s_1 et s_2. Une *géodésique* est une trajectoire (donc une polyligne) qui minimise la distance parcourue entre ses deux extrémités et qui peut se décrire selon la proposition suivante (Tournassoud 1988, p. 64) :

Proposition 1 *Dans l'espace libre, une géodésique se décompose en tronçons de trajectoires qui alternent nécessairement entre les types 1 et 2 ci-dessous :*

1. *déplacement en ligne droite sans contact avec les obstacles en dehors des deux extrémités, avec tangence aux obstacles aux deux points de contact,*

2. *glissement au contact d'une partie convexe de la frontière d'un obstacle.*

Le calcul de géodésiques ne constitue pas un problème facile dans un espace continu mais peut se simplifier par discrétisation de cet espace en un semis de points. En effet, il suffit alors de rechercher le plus court chemin au sein d'un graphe G dont aucun sommet n'appartient à la région R définie comme :

$$R = \bigcup_{i \in I} P_i \tag{2.10}$$

et dont aucune arête ne traverse un polygone P_i, $\forall i \in I$. La construction du graphe G nécessite évidemment des prédicats topologiques spécifiques. Cette solution est approximative, mais elle est d'autant plus précise que la discrétisation est fine, et la recherche d'une solution exacte est certainement de peu d'intérêt en pratique. Cependant, en affinant la discrétisation, le temps de calcul du plus court chemin dans le graphe augmente, ce qui peut nécessiter de recourir à des algorithmes approximatifs du type A^* (*e.g.*, Tournassoud 1988, pp. 153-156, Pearl 1990, pp. 67-68).

Un opérateur métrique également intéressant en écologie est celui qui calcule la distance entre deux polygones P_1 et P_2 comme la distance entre les points les plus proches appartenant aux frontières ∂P_1 et ∂P_2 :

$$d(P_1, P_2) = \min \{ d(\alpha, \beta) \mid \alpha \in \partial P_1, \beta \in \partial P_2 \} \tag{2.11}$$

Par exemple, les modèles de dynamique des métapopulations spatialement explicites ne mesurent pas la distance entre les patchs mais entre leurs centroïdes, parce que c'est évidemment plus facile (Hanski *et al.* 1994), mais cela peut fausser les résultats des modèles si la taille des patchs n'est pas négligeable par rapport à la distance inter-patchs. Dans ce cas, le recours à un tel opérateur élimine une source d'erreurs potentielles, sans pour autant modifier le modèle proposé.

2.2.1.3 Opérateurs topopologiques

Un opérateur topologique retourne un résultat en rapport avec la topologie ou la position des objets. Les prédicats qui testent la planarité d'une polyligne, d'un polygone, ou plus généralement d'un graphe, sont des exemples d'opérateurs déterminant la valeur d'un attribut topologique. Il y a cependant deux façons de concevoir la planarité. La première correspond à la définition donnée par la théorie des graphes (Berge 1970, p. 16, Alekseev 1995). Dans ce contexte, un algorithme testant si un graphe admet une représentation planaire est proposé par Hopcroft & Tarjan (1974). Selon une acception de la planarité plus restrictive, il ne s'agit plus de tester si un graphe G admet **une** représentation dans le plan, sans intersection des arêtes en dehors de leurs extrémités, mais si **la** représentation

de G dans le plan est effectivement planaire, les sommets et les arêtes étant fixés, aucune déformation topologique du graphe n'étant admise.

Les opérateurs topologiques peuvent également caractériser la position relative de deux objets d'une même classe, par exemple de deux polygones P_1 et P_2, ce qui peut permettre de répondre à des questions telles que :

- P_1 et P_2 se chevauchent-ils ?
- P_1 et P_2 se touchent-ils ?
- P_1 et P_2 ont-ils un côté en commun ?
- un des polygones est-il inclus dans l'autre ?

2.2.1.4 Opérateurs ensemblistes

Les opérateurs ensemblistes sont destinés à produire de nouveaux objets de la classe, par union, intersection ou différence de deux objets initiaux, par exemple deux polygones[9] P_1 et P_2, ce qui peut permettre, notamment, de calculer l'aire de l'intersection $P_1 \cap P_2$ (*e.g.*, chevauchement de deux domaines vitaux).

2.2.2 Opérateurs externes

En croisant les différentes classes d'objets, il est possible de définir de nombreux *opérateurs externes*. Nous nous contentons d'évoquer les opérateurs qui sont mentionnés dans les chapitres suivants.

2.2.2.1 Test du point dans le polygone

L'opérateur topologique le plus fondamental est un prédicat $\mathcal{P}$ testant si un point appartient à l'intérieur d'un polygone P, est situé sur sa frontière ∂P ou bien à l'extérieur de P. Ce prédicat permet notamment de réaliser une analyse "*point-in-polygon*", par exemple l'association d'un site de mesure à la classe géologique correspondante (Söderström & Eriksson 1996). Le problème du test du point dans un polygone est abordé par Shimrat (1962), Hacker (1962), Anderson (1976), Salomon (1978), Davis & David (1980), Preparata & Shamos (1985, pp. 41-45) et Huang & Shih (1997).

2.2.2.2 Segment en intersection avec un polygone

Le découpage résultant d'une situation d'intersection entre un segment $s_1 \leftrightarrow s_2$ et la frontière ∂P d'un polygone P quelconque constitue une opération essentielle. Le résultat de cette procédure de découpage peut être constitué d'un ensemble de segments et/ou d'un ensemble de polygones. Il est également possible d'associer une information topologique à chaque segment généré par le découpage afin de connaître sa position relative par rapport à ∂P. L'algorithmique d'un tel opérateur est partiellement traitée par Schweizer (1987, pp. 456-467).

[9] Le problème des opérations ensemblistes entre polygones est abordé par Schweizer (1987, pp. 469-493) et van Oosterom (1994).

2.2.2.3 Association d'un polygone à un semis

La façon la plus simple d'associer un polygone à un semis S consiste à définir son *rectangle minimum englobant*, *i.e.* le rectangle dont les côtés sont parallèles aux axes des coordonnées (O, x) et (O, y), et qui est défini par les coordonnées extrêmes des points $s \in S$.

Une autre association très classique consiste à calculer le plus petit polygone contenant tous les points du semis, *i.e.* l'*enveloppe convexe* de S également nommée *polygone convexe minimum*. En écologie, l'enveloppe convexe est utilisée, notamment, pour délimiter "objectivement" le domaine vital d'un animal d'après un ensemble de positions obtenues par radio-pistage (*cf.* Gallerani-Lawson & Rodgers 1997) ou le domaine d'échantillonnage d'après un ensemble de sites (*e.g.*, Minns *et al.* 1996). Le problème du calcul de l'enveloppe convexe d'un semis de points est abordé par Graham (1972), Jarvis (1973), Eddy (1977), Anderson (1978), Bykat (1978), Akl & Toussaint (1978), Andrew (1979), Fournier (1979), McCallum & Avis (1979), Akl (1979), Overmars & van Leeuwen (1980), Devroye & Toussaint (1981), Maus (1984), Demazure (1988, pp. 430-433), Larkin (1991) et Sedgewick (1991, pp. 377-390).

2.2.2.4 Association d'une tessellation à un semis

Classiquement, une tessellation est associée à un semis de points $S \subset D$ en délimitant la *zone d'influence* de chaque sommet s_i, *i.e.* l'ensemble des points $x \subset D$ plus proches de s_i que de n'importe quel autre sommet du semis. La tessellation obtenue est formée de *polygones de Voronoï*[10] P_i associés à chaque sommet de S (Toussaint 1980, Gordon & Finden 1985) :

$$P_i(S) = \{x \mid d(x, s_i) < d(x, s_j),\ \forall i \neq j \in I\} \tag{2.12}$$

La *tessellation de Voronoï* est largement utilisée en écologie (*e.g.*, Mead 1971, Vincent *et al.* 1976, Czárán & Bartha 1992, Perry 1995, Minns *et al.* 1996, Mercier 1997). Les propriétés de la tessellation de Voronoï sont revues dans Preparata & Shamos (1985, pp. 205-211), et le problème de sa construction est traité notamment par Rhynsburger (1973), Green & Sibson (1977), Brassel & Reif (1979), Preparata & Shamos (1985, pp. 211-220), Elbaz & Spehner (1990), Tipper (1991) et Tsai (1993). D'autres tessellations peuvent être dérivées de la tessellation de Voronoï au sens strict en pondérant les points du semis (*cf.* Chadœuf & Monestiez 1989, Mercier 1997, pp. 25-26).

2.2.2.5 Association d'un graphe à un semis

L'association d'un graphe à un semis S comportant n points peut s'effectuer de très nombreuses façons, même en se limitant à des 1-graphes connexes. Le problème sous-jacent peut parfois indiquer le graphe le plus approprié, mais en pratique, il est généralement difficile de présenter des arguments en faveur d'un graphe particulier (Gordon & Finden 1985).

[10] Les *polygones de Voronoï* sont également connus sous le nom de *régions de Dirichlet*, *polygones de Thiessen*, ou encore *cellules de Wigner-Seitz* (Preparata & Shamos 1985, p. 204).

Un graphe associé à un semis S traduit souvent une certaine conception du voisinage de chaque point $s \in S$. Dans ce contexte, le graphe est généralement défini selon des critères géométriques, les définitions les plus connues étant celles :

- de la triangulation de Delaunay (DT),
- du graphe de Gabriel (GG),
- du graphe de voisinage relatif (RNG),
- de l'arbre de poids minimum euclidien (EMST).

Une propriété remarquable des graphes cités est qu'ils sont tous des sous-graphes les uns des autres (Preparata & Shamos 1985, p. 263) :

$$EMST \subseteq RNG \subseteq GG \subseteq DT \tag{2.13}$$

Cette relation d'inclusion est illustrée à partir d'un semis de $n = 1096$ positions d'arbres[11] (Fig. 2.5). Il existe d'autres graphes de voisinage définis selon des critères géométriques — par exemple le k-RNG qui généralise le RNG (*cf.* Su & Chang 1991) — la plupart étant unifiés en un graphe à deux paramètres nommé *graphe de γ-voisinage* (Veltkamp 1992).

Triangulation de Delaunay Dans le plan, la triangulation de Delaunay[12] ou DT (*Delaunay Triangulation*) est définie comme le dual du diagramme de Voronoï (Preparata & Shamos 1985, Gordon & Finden 1985, Berstel *et al.* 1991) :

$$DT(S) = \{(s_i, s_j) \mid P_i \text{ et } P_j \text{ ont un segment frontière en commun}\} \tag{2.14}$$

Dans un domaine D borné, la DT peut être déduite de la tessellation de Voronoï représentée sous la forme d'un graphe G, au moyen d'un opérateur calculant son graphe d'adjacence G^*. Il est également possible de construire directement la DT (Maus 1984, Pasquier 1987, pp. 102-142, Berstel *et al.* 1991, pp. 93-109, Devillers *et al.* 1992, Joe 1993, Tsai 1993).

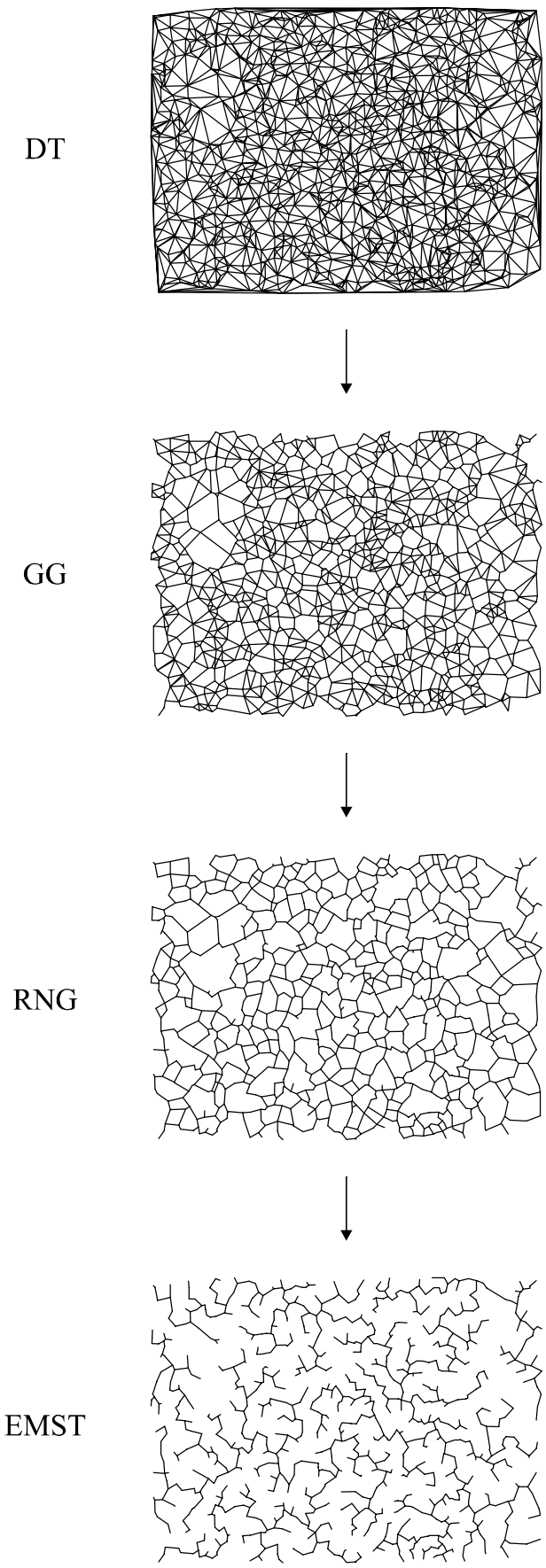
Graphe de Gabriel Le graphe de Gabriel ou GG (*Gabriel Graph*) a été introduit par Gabriel & Sokal (1969) dans l'étude de la variation géographique en biologie. Le graphe de Gabriel peut être défini comme (Matula & Sokal 1980, Gordon & Finden 1985) :

$$GG(S) = \left\{(s_i, s_j) \mid d_{ij}^2 < d_{ik}^2 + d_{jk}^2 \ , \ \forall k \neq i, j \in I\right\} \tag{2.15}$$

avec la simplification d'écriture $d_{ij} = d(s_i, s_j)$. En étudiant les propriétés des GG, Matula & Sokal (1980) montrent en particulier que $GG \subseteq DT$, ce qui permet d'envisager une construction efficace à partir de la DT.

[11] Données communiquées par le Dr. Marie-Agnès Moravie.

[12] Parmi toutes les triangulations possibles d'un semis de points donné, la triangulation de Delaunay maximise l'angle minimum, mais il est possible de faire référence à d'autres critères géométriques à optimiser (*e.g.*, Edelsbrunner *et al.* 1992, Schumaker 1993).



Figure˜2.5: Relation d'inclusion des quatre principaux graphes de voisinage pour un semis de $n = 1096$ positions d'arbres. DT : triangulation de Delaunay. GG : graphe de Gabriel. RNG : graphe de voisinage relatif. EMST : arbre de poids minimum euclidien.

Graphe de voisinage relatif Le graphe de voisinage relatif ou RNG (*Relative Neighbourhood Graph*) a été introduit par Toussaint (1980) dans le domaine de la reconnaissance des formes. Le graphe de voisinage relatif peut être défini comme (Preparata & Shamos 1985, p. 263) :

$$RNG\,(S) = \{(s_i, s_j) \mid d_{ij} \leq \min \max\,(d_{ik}, d_{jk})\,,\ \forall k \neq i, j \in I\} \tag{2.16}$$

Toussaint (1980) montre que le RNG se situe quelque part entre la DT et l'EMST, soit $EMST \subseteq RNG \subseteq DT$, et discute du problème de sa construction. O'Rourke (1982), Urquhart (1982a, 1982b) et Supowit (1983) étudient la complexité algorithmique du calcul du RNG. La construction directe du RNG est connue pour être considérablement plus coûteuse que celle du GG (Preparata & Shamos 1985, p. 263), mais la complexité algorithmique décroît évidemment fortement si le RNG est construit d'après le GG ou la DT.

Arbre de poids minimum euclidien L'arbre de poids minimum euclidien ou EMST (*Euclidean Minimum Spanning Tree*) associé au semis S est un arbre dont les sommets sont ceux de S et dont la longueur est minimale. Au sein de la n-clique induite par les n sommets de S, il existe n^{n-2} arbres distincts (Berge 1970, p. 41, Preparata & Shamos 1985, p. 189). Parmi tous ces arbres, le problème consiste à identifier l'EMST. En fait, il peut exister plusieurs EMST lorsque le semis présente des symétries internes[13].

La construction de l'EMST peut se poser en termes de construction du MST (*Minimum Spanning Tree*) de la n-clique dont les arêtes sont valuées par la distance euclidienne. Le problème général de la construction du MST dans un graphe a été résolu indépendamment par Kruskal (1956), Prim (1957) et Dijkstra (1959) (Preparata & Shamos 1985, p. 189). Les algorithmes de Kruskal et de Prim sont exposés notamment dans Sedgewick (1991, pp. 473-488), Froidevaux *et al.* (1993, pp. 487-502), Aho & Ullman (1993, pp. 521-527) et Prins (1994, pp. 248-262). L'utilisation de l'algorithme de Dijkstra revient à formuler le problème en termes de plus court chemin absolu dans un graphe incluant l'EMST. L'algorithme de Dijkstra est expliqué par Froidevaux *et al.* (1993, pp. 465-472), Aho & Ullman (1993, pp. 547-559) et Prins (1994, pp. 147-154). Des algorithmes efficaces sont proposés ou comparés par Yao (1975), Cheriton & Tarjan (1976), Whitaker (1977), Hung & Divoky (1988), Ahuja *et al.* (1990). Deux algorithmes efficaces pour construire le MST d'un graphe selon qu'il est orienté ou pas, sont proposés par Gabow *et al.* (1986). Une autre stratégie de construction de l'EMST est également exposée par Preparata & Shamos (1985, pp. 226-230).

Quel que soit l'algorithme considéré, il est évident que l'efficacité de la construction de l'EMST augmente de plus en plus si l'on considère la DT plutôt que de la n-clique, le GG plutôt que la DT, et enfin le RNG plutôt que le GG.

2.3 Modèles de cartographie

En traitant des phénomènes régionalisés (PR) du point de vue géomatique, nous faisons la distinction entre une *représentation cartographique* des données, une *cartographie* et une

[13] Dans ce cas, Matula & Sokal (1980) montrent que le GG inclut tous les EMST possibles.

carte. Une *représentation cartographique* consiste simplement à dessiner chaque support et à lui associer un chiffre, une couleur, un poncif ou un symbole, afin de représenter graphiquement la valeur associée au support. Au sens large, une *cartographie* est une procédure qui, à partir des données, renseigne sur le PR **en tout point de l'espace**. La cartographie constitue par conséquent un type de modélisation d'un PR, le modèle produit étant nommé simplement une *carte*. Il convient alors de distinguer deux situations selon que le PR (Gabriel & Sokal 1969) :

- présente des changements abrupts — au moins à l'échelle d'observation employée — ce qui nécessite de définir plusieurs régions dans lesquelles le PR est considéré comme homogène, l'ensemble des régions formant une *carte choroplèthe* (*e.g.*, une carte pédologique[14]),
- varie de façon continue, cette variation pouvant alors être représentée sous la forme d'une carte en isolignes ou *carte isoplèthe* (*e.g.*, une carte topographique).

L'objet de cette section est de préciser la notion de cartographie, puis de présenter les modèles vectoriels des cartes choroplèthe et isoplèthe. Enfin, nous considérons l'approximation discrète des cartes vectorielles sous la forme d'images.

2.3.1 Notion de cartographie

En géomatique, la *cartographie* d'un PR sur un domaine $D \subset \mathbb{R}^2$ muni d'une mesure d'aire μ peut être définie comme une fonction $\varphi(\cdot)$ qui à tout point $x \in D$ associe une *classe cartographique*, soit (Pasquier 1987, David 1991) :

$$\varphi : D \rightarrow E \tag{2.17}$$

avec E un ensemble de classes cartographiques muni d'un élément particulier u nommé *classe indéterminée*. La fonction $\varphi(\cdot)$ est telle que (Pasquier 1987) :

$$\mu\left[\varphi^{-1}(u)\right] = 0 \tag{2.18}$$

autrement dit, la classe cartographique de tout support inclus dans D est déterminée. Deux types d'opérations peuvent être définis sur les cartographies afin d'obtenir une *cartographie dérivée* et une *cartographie produit*.

2.3.1.1 Cartographie dérivée

Soit F un ensemble de classes cartographiques différent de E et une fonction $f : E \rightarrow F$, la fonction composée $f \circ \varphi$ est appelée *cartographie dérivée* de φ (Pasquier 1987). Considérons un exemple très simple dans lequel φ est la cartographie géologique de D, avec E comportant de nombreux types de roches. Si la fonction f associe à chaque type de roche sa nature géochimique, soit carbonatée, soit siliceuse, alors la composition $f \circ \varphi$ définit la cartographie des roches carbonatées et siliceuses.

[14] L'adéquation du modèle de carte choroplèthe à la représentation de la pédologie est discutée par Burrough *et al.* (1997), Zhu *et al.* (1996) et Zhu (1997), dans le contexte de la logique floue.

2.3.1.2 Cartographie produit

Soient $\varphi : D \rightarrow E$ et $\psi : D \rightarrow F$ deux cartographies de D, le produit $\varphi \times \psi : D \rightarrow E \times F$ qui a chaque point $x \in D$ fait correspondre le couple $(\varphi(x), \psi(x))$ est appelée *cartographie produit* (Pasquier 1987). Cette notion s'étend au produit d'un nombre quelconque de cartographies et correspond à la production des cartes polythématiques par superposition de cartes élémentaires (*map-overlay*). Considérons par exemple que φ est la cartographie des sols et que ψ est celle de la couverture végétale, alors le produit $\varphi \times \psi$ est la cartographie qui associe à la fois le type de sol et le type de végétation.

2.3.1.3 Cartographie et variable régionalisée

De même qu'une variable régionalisée (VR), une cartographie est une fonction de l'espace géographique. Il convient cependant de faire la distinction entre les deux notions. Une VR correspond à une formulation mathématique de **la réalité** du PR, et ses valeurs ne peuvent être connues que par des mesures ou des observations. Une cartographie est une procédure de modélisation de la VR, et une carte, comme tout modèle, ne constitue qu'une approximation de la réalité. La carte résultant d'une cartographie est strictement équivalente à une VR uniquement lorsque la cartographie correspond à une campagne exhaustive de mesure ou d'observation.

Considérons par exemple le diamètre des arbres : si la cartographie sur D consiste à inventorier tous les arbres présents dans D et à mesurer leur diamètre, alors la carte obtenue est équivalente à la VR. En revanche, si la cartographie comporte une procédure d'estimation du diamètre des arbres, alors la carte produite diffère de la VR.

La distinction entre cartographie et VR disparaît dès lors que l'on ne fait plus explicitement la différence entre les vraies valeurs de la VR et les estimations de ces valeurs.

2.3.2 Carte choroplèthe

Une carte choroplèthe est modélisée par une tessellation limitée par la frontière de D. Selon le type de traitement envisagé, la tessellation peut être vue, soit comme une ensemble de polygones contigus mais indépendants, soit comme un graphe planaire[15]. La modélisation par un graphe planaire est particulièrement adaptée lorsque la topologie doit être exploitée, par exemple pour calculer le graphe d'adjacence des polygones de la carte.

2.3.3 Carte isoplèthe

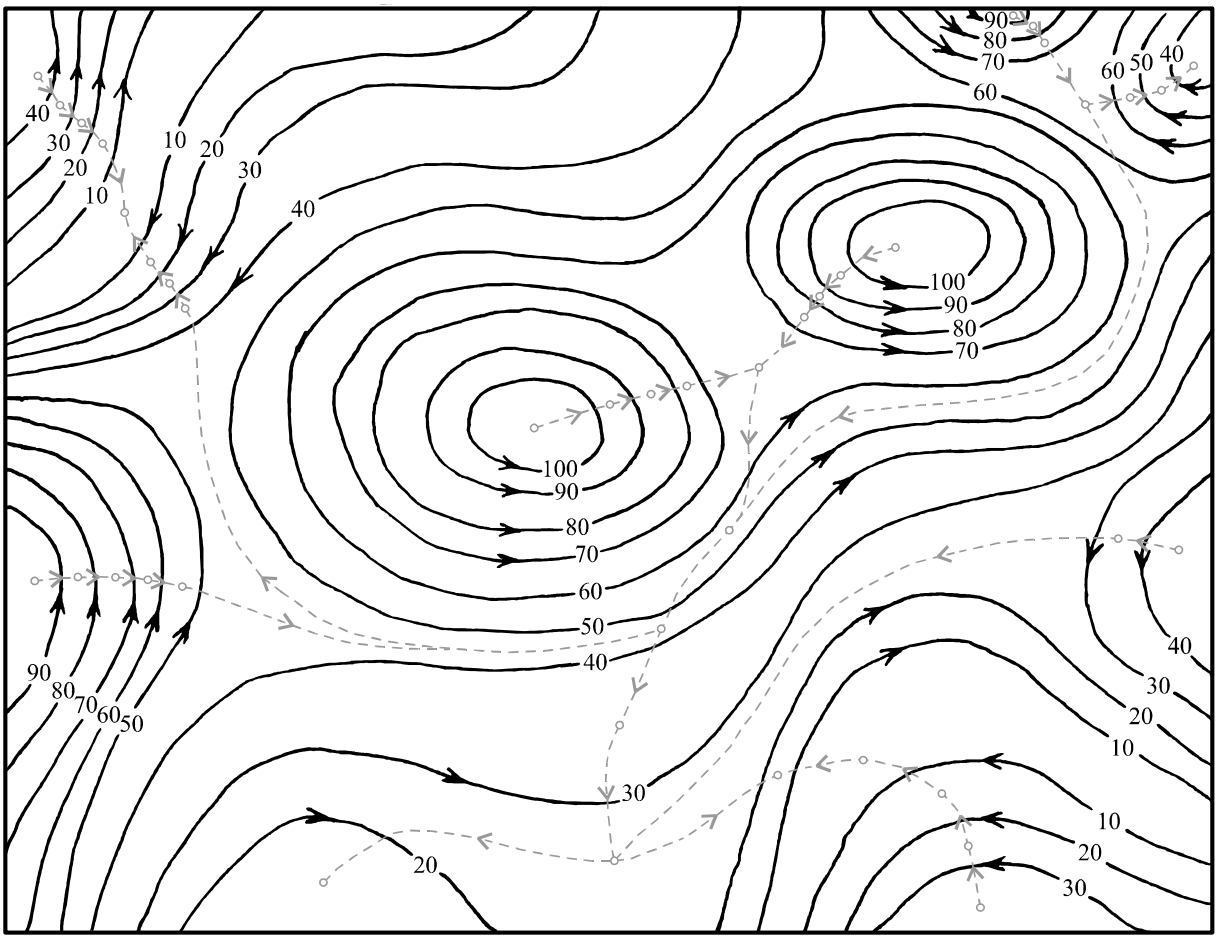
Une carte isoplèthe est composée d'un ensemble d'isolignes[16], *i.e.* de polylignes valuées par des cotes. Une carte isoplèthe peut éventuellement être vue comme un cas particulier de carte choroplèthe dans laquelle les polygones définissent des domaines correspondant

[15] Il existe également un modèle de *carte combinatoire* (*cf.* David 1991, pp. 51-67), mais son utilisation semble encore marginale.

[16] Le terme *isoligne* est parfois distingué de *isoplèthe* ou de *isarithme* dans un contexte de cartographie manuelle (*e.g.*, Mathieu 1990).

à des intervalles de cotes $[\alpha, \beta]$. Cependant, la plupart des traitements portant sur une carte isoplèthe nécessitent un modèle exploitant ses propriétés spécifiques, notamment topologiques.

Un modèle adéquat de carte isoplèthe est constitué d'un polygone définissant le domaine D et d'un graphe planaire G dont les arêtes sont les isolignes. Les arêtes de G sont naturellement valuées par les cotes des isolignes. Le graphe peut être orienté par la *convention du dahu*, *i.e.* en considérant que toute isoligne sépare un domaine supérieur situé à sa gauche d'un domaine inférieur situé à sa droite (Bouillé 1975). La modélisation par un graphe et l'introduction de la convention d'orientation des arêtes permettent de représenter la topologie associée à une carte isoplèthe par le graphe d'adjacence G^* (Bouillé 1975). La Figure 2.6 donne un exemple de carte isoplèthe modélisée par un graphe orienté.



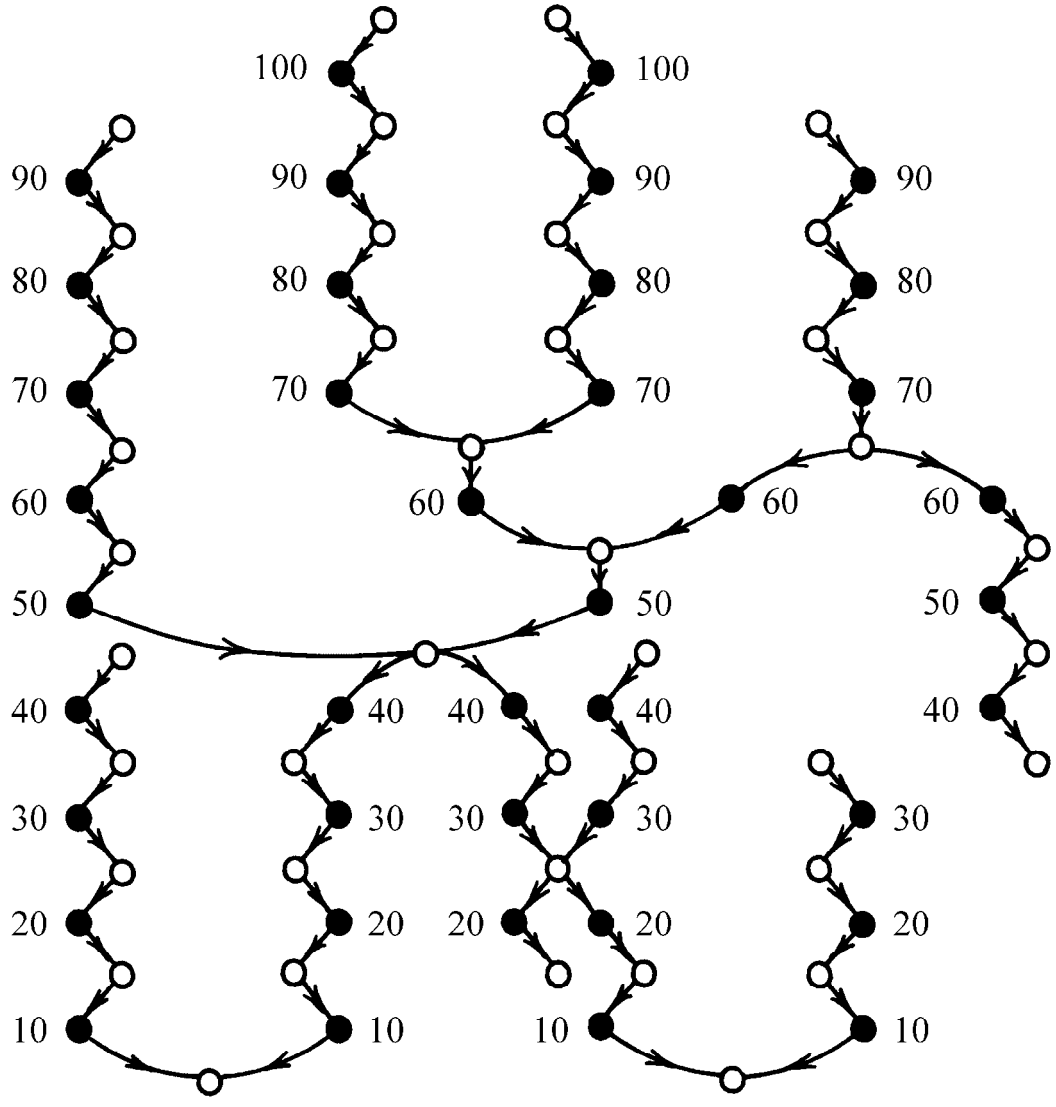
Figure˜2.6: Carte isoplèthe modélisée par un graphe G orienté par la convention du dahu (figuré en noir) et son graphe d'adjacence G^* (figuré en gris) (d'après Bouillé 1975).

Un sommet de G^* est associé à chaque face interne de G (Fig. 2.6 & 2.7). Comme G est orienté, son graphe d'adjacence G^* l'est également. Par une série de transformations opérant sur G^*, on obtient un graphe T (Fig. 2.8) décrivant pour chaque isoligne quelles sont les isolignes inférieures, supérieures ou de même cote que l'on peut joindre sans croiser d'isolignes (Bouillé 1975).

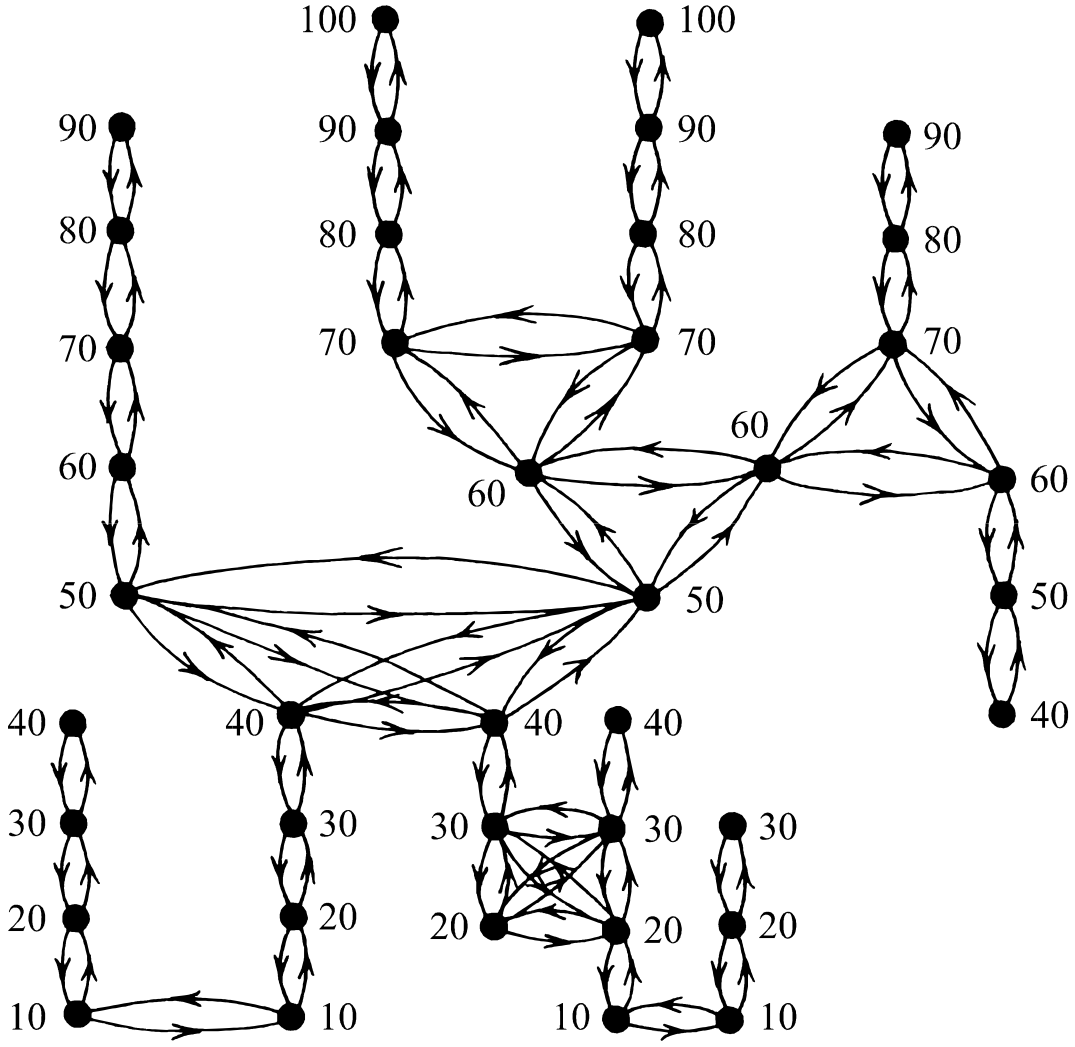
Le graphe T comporte donc trois types d'arcs exprimant à la fois :

- les relations d'ordre dans l'ensemble des cotes ($\prec$, $=$, $\succ$),
- une relation topologique de voisinage a Jonc b telle que l'on peut passer de l'isoligne a à l'isoligne b sans croiser une autre isoligne.

Notons Inf, Sup et Vois les relations de voisinage d'une isoligne avec les isolignes de cote respectivement inférieure, supérieure, et égale. En exploitant les propriétés des relations Inf, Sup et Vois (Annexe C) il est possible de concevoir des algorithmes calculant T directement à partir des polylignes et de ∂D, *i.e.* en mode vecteur et sans aucune information topologique explicite.



Figure˜2.7: Graphe d'adjacence G^*. Les points blancs représentent les domaines compris entre les isolignes (sommets du graphe d'adjacence) et les points noirs représentent les isolignes traversées par le graphe d'adjacence (d'après Bouillé 1975).



Figure˜2.8: Graphe orienté T décrivant les relations de voisinage topologique entre les isolignes. Les flèches horizontales correspondent à la relation Vois, et les flèches verticales correspondent aux relations Inf et Sup (d'après Bouillé 1975).

2.3.4 Images

On considère en général que le mode vecteur se prête difficilement au traitement statistique de l'information spatiale contenue dans une carte (Arbia 1993). Une solution consiste à discrétiser la carte en appliquant sur D une tessellation régulière de maille carrée. Ce type de discrétisation ou *rasterisation* produit une *image* constituée d'une grille de *pixels*[17] Ω. A la différence du mode vecteur, le mode image ou *raster* offre de nombreuses possibilités analytiques (*e.g.*, Collet 1992).

Le mode raster est généralement préféré au mode vecteur pour la modélisation écologique (Chiarello 1994, Heil & van Deursen 1996). En particulier, les images se prêtent bien à l'étude d'un phénomène à différentes échelles parce qu'elles peuvent être traitées à différentes résolutions. La question des échelles spatiales étant fondamentale dans toutes les investigations écologiques (Wiens 1989, Milne 1992) et en particulier en écologie du paysage (Turner *et al.* 1989, Chiarello 1993, 1994, 1996, Qi & Wu 1996, Gardner 1998), il est important de pouvoir manipuler les images sous la forme de structures hiérarchiques.

Néanmoins, le mode raster présente quelques inconvénients dans la mesure où la rasterisation a pour principales conséquences :

- d'approximer la carte en mode vecteur,
- de supprimer l'identité de chaque objet spatial initial en le décomposant en un ensemble de pixels noyés dans la masse des autres pixels de l'image.

L'utilisation du mode raster peut par conséquent entraîner de sérieuses erreurs lorsque le problème à traiter concerne la géométrie ou la topologie d'objets spatiaux. Le mode raster nécessite en particulier de redéfinir une géométrie et une topologie adaptées à un espace discret (Chassery & Montanvert 1991). Il convient notamment de préciser la topologie de l'image.

2.3.4.1 Structures hiérarchiques

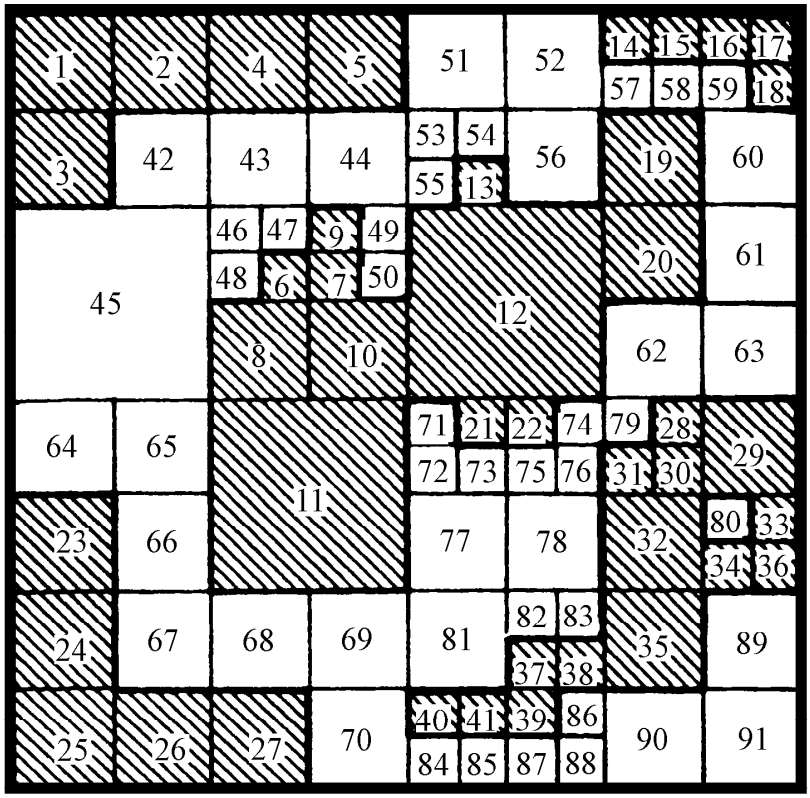
Afin de manipuler plusieurs échelles spatiales, il faut représenter l'image au moyen d'une structure hiérarchique telle qu'un *quadtree* ou une *pyramide*. Nous n'envisageons pas les structures hiérarchiques comme un moyen parmi d'autres de représenter une image, mais plutôt pour leur utilité en termes de description et de traitement des formes (Laurini & Milleret-Raffort 1989, Chassery & Montanvert 1991).

Quadtree La structure d'une image carrée $\Omega = 2^r \times 2^r$ peut être représentée sous forme hiérarchique grâce à un *quadtree* (Besançon 1988, pp. 317-356, Samet 1990, Chassery & Montanvert 1991, pp. 102-127). Un *quadtree* est un arbre qui correspond à la décomposition de l'image en quatre quadrants, cette décomposition se répétant de façon récursive sur chaque quadrant comportant des pixels de valeurs différentes, jusqu'à ce que tous les quadrants soient homogènes (Fig. 2.9 & 2.10). Les différents niveaux du *quadtree* correspondent à différentes résolutions de l'image : la racine de l'arbre représente l'image originelle (quadrant unique), les noeuds correspondent aux quadrants successifs constitués

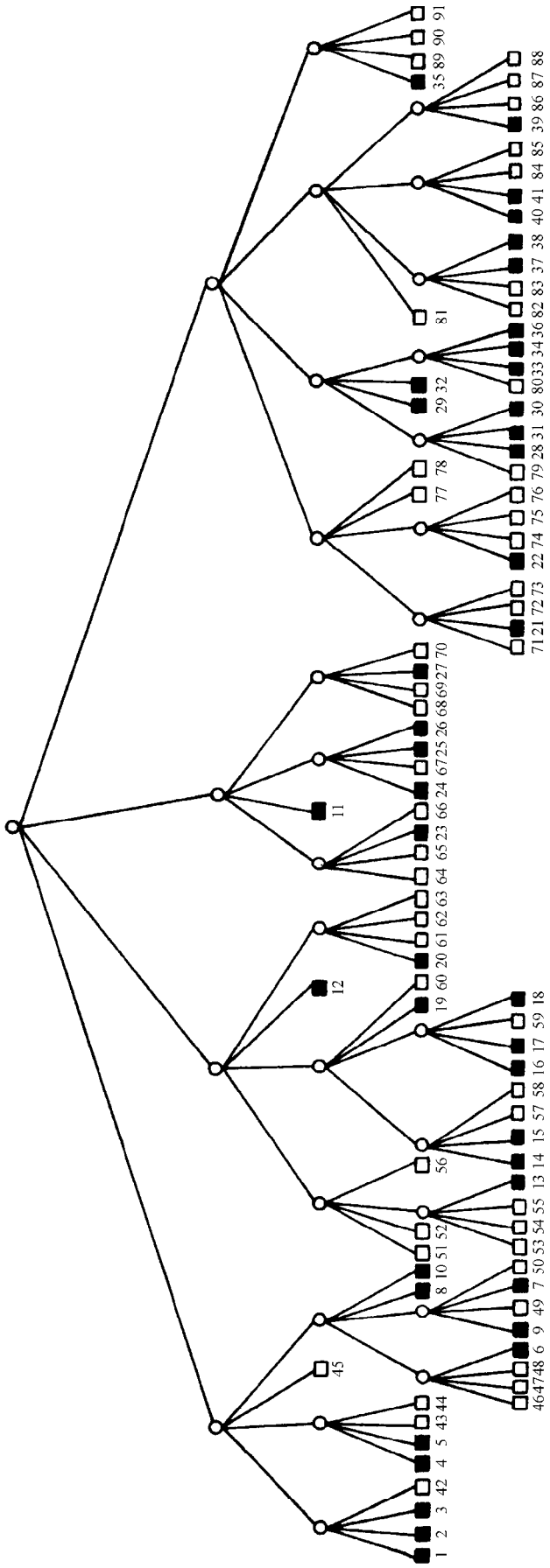
[17] Le néologisme *pixel* a été formé par contraction de *picture element* (David 1991).

lors de la décomposition récursive, et les feuilles de l'arbre correspondent à des quadrants homogènes, qui peuvent être, le cas échéant, des pixels de l'image.

La racine, les noeuds et les feuilles du *quadtree* sont étiquetés par la valeur moyenne du quadrant correspondant s'il s'agit d'une image en niveaux de gris. Dans le cas d'une image binaire, les quadrants homogènes sont noirs ou blancs, et les quadrants inhomogènes sont gris.



Figure˜2.9: Blocs de pixels homogènes d'une image binaire 16×16 (d'après Samet 1981).



Figure˜2.10: *Quadtree* de l'image binaire 16×16 (d'après Samet 1981, corrigé).

Pyramide Une *pyramide* est un mode de représentation multi-résolution qui peut être associé à une image carrée $\Omega = 2^r \times 2^r$. La construction de la pyramide consiste à effectuer la décomposition récursive des quadrants jusqu'au niveau des pixels, indépendamment d'un critère d'homogénéité (Rosenfeld 1983). Par exemple, une image de résolution 128×128 sera également mémorisée à la résolution 64×64 par agrégation des pixels au sein de chaque quadrant de 2×2 pixels, puis aux résolutions 32×32, 16×16, et finalement 8×8 (*e.g.*, Laurini & Milleret-Raffort 1989, p. 30). En associant un arbre de décomposition hiérarchique à la pyramide, on obtient un *quadtree* dont tous les noeuds sont développés jusqu'aux feuilles qui correspondent aux pixels de l'image de résolution la plus élevée (*e.g.*, Csillag & Kabos 1996).

2.3.4.2 Topologie d'une image

Deux pixels $a, b \in \Omega$ sont dits voisins si $\partial a \cap \partial b \neq \emptyset$. Dans une image, un pixel entretient avec ses huit voisins deux types de voisinages (Postaire 1987, Pasquier 1987, Chassery & Montanvert 1991) :

- le voisinage direct[18] dans lequel deux pixels voisins ont un côté commun,
- le voisinage indirect lorsque deux pixels n'ont qu'un sommet en commun.

Parmi les huit voisins d'un pixel, il y a donc quatre pixels en voisinage direct et quatre pixels en voisinage indirect. En se restreignant au voisinage direct, on définit le *4-voisinage*, tandis qu'en considérant à la fois le voisinage direct et le voisinage indirect, on définit le *8-voisinage*. Le choix d'un type de voisinage définit *ipso facto* un type de connexité.

Connexité dans une image Une chaîne connexe reliant deux pixels $a, b \in \Omega$ est un ensemble de n pixels $a \leftrightarrow p_1 \leftrightarrow p_2 \leftrightarrow \cdots \leftrightarrow p_n \leftrightarrow b$ tel que a est voisin de p_1, b est voisin de p_n et p_i est voisin de p_{i+1} pour $i = 1, \ldots, n-1$ (Postaire 1987, Pasquier 1987, Chassery & Montanvert 1991). Un ensemble de pixels $A \subset \Omega$ est connexe si tout pixel $p \in A$ peut être relié à tout autre pixel de A par une chaîne connexe dont les éléments appartiennent à A. Selon le type de voisinage retenu, deux types de connexités sont donc définis :

- le 4-voisinage définit la *4-connexité* ou *4-adjacence*,
- le 8-voisinage définit la *8-connexité* ou *8-adjacence*.

Choix de la connexité On montre qu'il n'est pas possible de construire une topologie cohérente pour une image sur la base de la 8-connexité puisque, paradoxalement, deux chaînes 8-connexes peuvent se croiser sans se couper (Pasquier 1987, p. 58, Chassery & Montanvert 1991, p. 44). Par la suite, nous faisons exclusivement référence à la 4-connexité (4-adjacence).

[18] Dans la littérature, on utilise souvent les mouvements des pièces d'un jeu d'échec afin de caractériser le voisinage direct (mouvement de la Tour) et le voisinage indirect (mouvement du Fou) (*e.g.*, Cliff & Ord 1981, Upton & Fingleton 1985).

Chapitre 3

Autocorrélation spatiale

> *"Peut-on, dans un paysage de phénomènes, reconnaître un objet ou une chose si l'on n'en a pas préalablement le concept ? C'est aussi simple que cela. Si l'on n'a pas le concept d'un objet, on ne le reconnaîtra pas."* (Thom 1991)
>
> *"The outcome of a spatial autocorrelation test will be sensitive to* [...] *operational definitions."* (Upton & Fingleton 1985)

Une variable régionalisée (VR) présente généralement une structure spatiale : si tel n'est pas le cas, le caractère régionalisé est sans intérêt pour l'analyse et la modélisation de la variable[1]. Le traitement d'une VR suppose donc *a priori* une structure spatiale dont l'absence constitue une hypothèse nulle qu'il est possible d'évaluer statistiquement.

Quantifier la structuration spatiale d'une VR nécessite de définir une statistique. Généralement, cette statistique traduit de façon opératoire une propriété essentielle d'une VR spatialement structurée appelée *autocorrélation spatiale*. En termes intuitifs, l'autocorrélation spatiale peut être vue comme la ressemblance des valeurs d'une VR en fonction de la position de leurs supports. Une VR présente de l'autocorrélation spatiale positive si les valeurs mesurées sur des supports voisins se ressemblent davantage qu'elles ne ressemblent aux autres valeurs. On parle d'autocorrélation négative dans le cas d'une dissemblance. Toute statistique mesurant l'autocorrélation spatiale repose donc sur la mise en relation de deux informations :

- le voisinage des supports,
- la ressemblance entre valeurs.

Gatrell (1979) distingue l'*espace absolu*, dans lequel les supports sont localisés, de l'*espace relatif*, défini par les relations entre supports. Définir différentes relations revient alors à définir différents espaces relatifs au sein desquels il est possible de mesurer la distance de nombreuses façons. Une telle distinction met l'accent uniquement sur les supports, alors que la mesure de l'autocorrélation spatiale nécessite de définir des proximités tant dans l'espace des supports que dans celui des valeurs. Evidemment, dans chacun de ces espaces, il est possible de définir différents types de proximités. En outre, il existe plusieurs approches pour mettre en relation les proximités entre les supports et celles entre les valeurs, ce qui conduit à de nombreuses mesures de l'autocorrélation spatiale.

[1] De façon abusive, certains auteurs considèrent qu'une variable localisée dans l'espace, mais qui ne présente pas de structure spatiale, n'est pas une variable régionalisée (*e.g.*, Lecoustre *et al.* 1989).

Nous proposons de classer les principales mesures d'autocorrélation spatiale en croisant deux approches statistiques et deux modes de calcul :

- approche statistique :
  - corrélation entre deux matrices de proximités,
  - rapport de deux variances (ou covariances),
- mode de calcul :
  - global, *i.e.* pour tous les supports à la fois,
  - fractionné, *i.e.* pour certaines classes de supports.

3.1 Corrélation entre deux matrices de proximités

Soit une VR $z(\cdot)$ dont on connaît n valeurs $\{z_i \mid i = 1, \ldots, n\}$ mesurées ou observées sur un ensemble de supports $\{x_i \mid i = 1, \ldots, n\}$, avec $z_i = z(x_i)$ pour $i = 1, \ldots, n$. Notons X_{ij} la proximité entre deux supports x_i et x_j et Y_{ij} la similarité entre les valeurs z_i et z_j. Dans le cas le plus général, les mesures X et Y sont symétriques, ce qui conduit à considérer uniquement les $m = \binom{n}{2} = n(n-1)/2$ entrées (i, j) de demi-matrices inférieures $n \times n$. Notons Z une mesure de la corrélation entre les valeurs X_{ij} et Y_{ij}. Tout triplet (X, Y, Z) définit de façon opératoire une forme particulière d'autocorrélation spatiale. Tester l'autocorrélation spatiale sur la base d'un triplet (X, Y, Z) revient en fait à tester l'adéquation entre les données et un modèle de la variation spatiale proposé *a priori*. Dans une telle approche, il faut donc choisir judicieusement :

- une proximité entre supports (X),
- une similarité entre valeurs (Y),
- une statistique Z pour mesurer la corrélation entre les valeurs des deux matrices $\mathbf{X} = ((X_{ij}))$ et $\mathbf{Y} = ((Y_{ij}))$.

Une fois le choix d'un triplet (X, Y, Z) effectué, il convient ensuite d'examiner le test de la statistique obtenue.

3.1.1 Choix de la proximité

La proximité entre supports peut être définie de nombreuses façons pour un même jeu de données $\{z(x_i) \mid i = 1, \ldots, n\}$ (Cliff & Ord 1981). Il est possible de distinguer :

- les matrices d'adjacence de graphes, pas nécessairement planaires, et éventuellement pondérés,
- les matrices de proximités basées sur une métrique.

3.1.2 Choix de la similarité

Il n'existe pas de règle précise pour choisir la similarité entre valeurs. Cette similarité (ou dissimilarité) peut être basée sur une semi-métrique, une métrique, plus rarement une ultra-métrique, et choisie en fonction :

- de la structure algébrique de la variable,
- de l'idée que l'on se fait de la ressemblance entre valeurs,
- de la robustesse à l'asymétrie de la distribution statistique des valeurs,
- de la robustesse aux valeurs extrêmes.

3.1.3 Choix de la corrélation

Indépendamment du choix des définitions des proximités X et Y, différentes familles de tests peuvent être construites selon le choix de la corrélation Z :

- corrélation paramétrique de Pearson,
- corrélation non paramétrique de Spearman (*i.e.*, corrélation de Pearson calculée sur les rangs des valeurs),
- corrélation non paramétrique de Kendall, calculée sur les rangs des valeurs.

Pour des définitions de X et Y fixées, il est possible de se demander quelle corrélation conduit au test de l'autocorrélation spatiale le plus puissant (Dietz 1983). Dans ce contexte, et en absence d'une évaluation fiable et générale de la puissance comparée de ces trois corrélations, le coefficient de Pearson constitue un choix par défaut.

3.1.4 Test de la corrélation entre matrices

Considérons le coefficient de corrélation de Pearson entre les X_{ij} et les Y_{ij} (Manly 1991, 1993) :

$$r = \frac{\sum\limits_{i<j} X_{ij}Y_{ij} - \sum\limits_{i<j} X_{ij} \sum\limits_{i<j} Y_{ij}/m}{\left[\left\{\sum\limits_{i<j} X_{ij}^2 - \left(\sum\limits_{i<j} X_{ij}\right)^2 /m\right\}\left\{\sum\limits_{i<j} Y_{ij}^2 - \left(\sum\limits_{i<j} Y_{ij}\right)^2 /m\right\}\right]^{1/2}} \tag{3.1}$$

L'hypothèse nulle H_0 est l'absence de corrélation entre les X_{ij} et les Y_{ij}. Cette hypothèse ne peut pas être testée par la procédure paramétrique habituelle puisque ni les X_{ij}, ni les Y_{ij} ne sont des réalisations de variables aléatoires mutuellement indépendantes, et puisque leur distribution conjointe n'est certainement jamais normale.

Néanmoins, il est possible de recourir à une approche permutationnelle. Dans ce cadre, Mantel (1967) considère la statistique :

$$Z = \sum_{i<j} X_{ij}Y_{ij} \tag{3.2}$$

Cette statistique est équivalente au r de Pearson puisque les termes autres que Z qui figurent dans (3.1) sont invariants par permutation des indices (Manly 1991). Soit Π l'ensemble des $n!$ permutations des indices $i = 1, 2, \ldots, n$. La distribution de Z sous H_0 est obtenue en assignant une probabilité égale $(1/n!)$ aux valeurs de Z calculées pour les $n!$ permutations des lignes et colonnes d'une des deux matrices (Dietz 1983, Manly 1991, 1993). Bien entendu, la distribution sous H_0 est conditionnelle aux matrices $\mathbf{X}$ et $\mathbf{Y}$, et la conclusion du test concerne uniquement les données analysées. A partir de la distribution exacte de Z sous H_0, la *p-value*[2] est calculée comme la proportion de valeurs supérieures ou égales à la valeur observée Z_{obs} (ou inférieures ou égales, selon l'hypothèse alternative considérée) :

$$p = \frac{\text{Card}\left(\{Z \mid Z \geq Z_{obs}\}\right)}{n!} \tag{3.3}$$

Le *test permutationnel* peut être envisagé pour $n \leq 10$. Dans la plupart des cas ($n > 10$), l'énumération de toutes les permutations est impraticable et le test permutationnel est approximé par un *test de randomisation* (Edgington 1986, 1987) en construisant un ensemble de valeurs obtenu par échantillonnage aléatoire et avec remise de Π (Dietz 1983). La *p-value* obtenue dans le test de randomisation est une estimation sans biais de la *p-value* exacte (3.3) :

$$\widehat{p} = \frac{\text{Card}\left(\{Z \mid Z \in \Omega,\ Z \geq Z_{obs}\}\right)}{\text{Card}\left(\Omega\right)} \tag{3.4}$$

avec la valeur observée Z_{obs} figurant dans l'ensemble Ω des valeurs de Z pour les permutations considérées. La distribution d'échantillonnage de cette *p-value* est asymptotiquement normale, de moyenne p et de variance $p\left(1-p\right)/m$. Il est conseillé de générer un nombre élevé de permutations aléatoires afin de réduire la fluctuation de la convergence et atteindre une estimation précise de la *p-value* (Jackson & Somers 1989). A cette fin, un nombre de 10^4 permutations semble constituer un minimum, et nous utilisons couramment 5×10^4 ou 10^5 permutations aléatoires[3] pour $n = 100$.

Comme alternative au test de randomisation, Mantel (1967) et Mantel & Valand (1970) suggèrent que la distribution de Z sous H_0 est approximativement normale. En effet, Mantel (1967) indique que Z a l'apparence classique d'une *U-statistique* dont une des propriétés est d'être asymptotiquement distribuée selon la loi normale (Hoeffding 1948), du moins dans des conditions typiques (*cf.* Serfling 1988). Mielke (1978) note justement que la statistique de Mantel (3.2) ne satisfait pas à un théorème nécessaire pour qu'une U-statistique suive la loi normale. En conséquence, Mielke (1978) propose une distribution de Pearson de type III (*cf.* Ord 1985). Il semble donc que la distribution du Z de Mantel soit non normale de sorte que les *p-values* obtenues à partir de l'approximation normale sont trop faibles (Mielke 1978, Luo & Fox 1996). En cas de doute concernant l'hypothèse de normalité, Mantel lui-même recommande de tester la corrélation entre matrices grâce à un test de randomisation (Mantel 1967).

[2] Concernant l'interprétation des *p-values* et l'usage d'un seuil (*e.g.*, $\alpha = 0.05$), on peut consulter Gibbons (1986), Casella & Berger (1987a), Berger & Sellke (1987), Yoccoz (1991) et Stewart-Oaten (1995). Nous considérons de façon classique qu'une *p-value* mesure la force de l'évidence contre l'hypothèse nulle, conditionnellement aux données (Edgington 1987, Casella & Berger 1987b). En pratique, l'interprétation d'une *p-value* doit s'effectuer en tenant compte de la quantité d'information disponible (Hinkey 1987).

[3] L'algorithme utilisé par Dietz (1983), et Manly (1991, 1993), n'échantillonne pas très bien la distribution exacte. Il est préférable de permuter deux valeurs choisies au moyen d'un couple de nombres pseudo-aléatoires différent d'une permutation à l'autre.

Enfin, en considérant un jeu de données réelles, Faust & Romney (1985) étudient la sensibilité du test de Mantel à l'asymétrie des distributions des valeurs des matrices $\mathbf{X}$ et $\mathbf{Y}$. Bien que Faust & Romney (1985) calculent les *p-values* à partir de 500 permutations aléatoires seulement, les écarts observés entre les résultats obtenus à partir des valeurs brutes, des valeurs log-transformées, et de leurs rangs, montrent clairement que le test de Mantel est sensible à l'asymétrie des distributions des X_{ij} et Y_{ij}. Ce résultat n'est pas surprenant puisque le test de Mantel n'est pas autre chose qu'un test du coefficient de corrélation de Pearson entre les valeurs des deux matrices ; or, le test du coefficient de corrélation de Pearson est connu pour être sensible à l'asymétrie de la distribution des données (Kowalski 1972).

3.2 Indices d'autocorrélation spatiale

Une façon de mesurer l'autocorrélation spatiale est de constituer le rapport d'une variance (ou covariance) utilisant l'information spatiale — *i.e.* la localisation des supports des valeurs — à une variance (ou covariance) n'utilisant pas cette information. L'écart entre la valeur observée de ce rapport et son espérance sous H_0 constitue une mesure de l'intensité de l'autocorrélation spatiale qui peut être testée. Deux indices illustrent ce paradigme : le c de Geary et le I de Moran.

3.2.1 c de Geary

Geary introduit en géographie un indice c destiné à tester si les statistiques données pour chaque comté d'un pays sont distribuées au hasard ou bien au contraire, si elles présentent une structure spatiale (Geary 1954) :

$$c = \frac{(n-1)}{2K} \frac{\sum_{i \neq j} (z_i - z_j)^2}{\sum_i (z_i - \bar{z})^2} \tag{3.5}$$

avec n le nombre de comtés, z_i la mesure pour le comté i ($i = 1, \ldots, n$) et $K = \sum k_i$ où k_i est le nombre de connexions entretenues par le comté i avec ses voisins. Sous cette forme, le c de Geary concerne donc des supports organisés selon une carte choroplèthe (Section 2.3.2), et K est la somme des arêtes du graphe d'adjacence associé à cette carte (Section 2.1.2, p. 17).

Le c de Geary constitue en fait une généralisation à l'espace bidimensionnel du rapport η de von Neumann défini pour les séries temporelles (von Neumann 1941) :

$$\eta = \frac{\delta^2}{s^2} \tag{3.6}$$

$$\delta^2 = \frac{1}{n-1} \sum_{i=1}^{n-1} (z_{i+1} - z_i)^2 \tag{3.7}$$

avec $s^2 = n^{-1} \sum (z_i - \bar{z})^2$ l'estimateur classique de la variance. La statistique δ^2 est une estimation de la variance destinée à minimiser l'effet de l'autocorrélation temporelle sur la

dispersion statistique. Cette estimation repose sur les différences entre valeurs successives de la série temporelle. En revanche, s^2 estime la variance indépendamment de l'ordre des observations et n'inclut donc pas l'effet d'une éventuelle autocorrélation. L'autocorrélation temporelle peut être testée en comparant δ^2 et s^2 (von Neumann *et al.* 1941, von Neumann 1941). A cet effet, les deux premiers moments du rapport η sont obtenus par Williams (1941), et la distribution de η est étudiée par von Neumann (1941). Par ailleurs, la statistique (3.6) est algébriquement identique à la statistique d introduite par Durbin & Watson (1950) pour tester l'autocorrélation sérielle du terme d'erreur dans un modèle linéaire, en utilisant les résidus calculés à partir d'un ajustement aux moindres carrés (Pettitt 1982). Ainsi, le c de Geary est également lié au d de Durbin-Watson (Sokal & Oden 1978a). Enfin, Okabe (1976) propose une interprétation du c de Geary basée sur l'analyse de variance, et Chessel (1981) mentionne les relations entre le c de Geary, l'ANOVA hiérarchique proposée par Greig-Smith (1952), l'indice non paramétrique de dispersion de Chessel & Croze (1978), et la statistique de Walter (1974).

3.2.1.1 Généralisation du c de Geary

La forme originelle du c de Geary peut être généralisée en considérant une matrice $\mathbf{W}$ de poids w_{ij} affectés à tous les couples de supports (i, j)(Cliff & Ord 1981) :

$$c = \frac{(n-1)}{2W} \frac{\sum\limits_{i,j} w_{ij} \left(z_i - z_j\right)^2}{\sum\limits_{i} \left(z_i - \bar{z}\right)^2} \tag{3.8}$$

avec $W = \sum w_{ij}$. La forme (3.5) correspond ainsi au cas particulier où $\mathbf{W}$ est la matrice associée au graphe d'adjacence d'une carte choroplèthe.

3.2.1.2 Variogramme et c de Geary

En considérant en particulier les $N(\mathbf{h})$ couples de supports (i, j) séparés par un vecteur $\mathbf{h}_{ij} = \mathbf{h}$, le c de Geary s'écrit comme la fonction :

$$c(\mathbf{h}) = \frac{\widehat{\gamma}(\mathbf{h})}{s^2_{n-1}} \tag{3.9}$$

avec $s^2_{n-1} = (n-1)^{-1} \sum \left(z_i - \bar{z}\right)^2$ et $\widehat{\gamma}(\mathbf{h})$ le variogramme expérimental[4] défini par Matheron (1965), qui peut s'écrire (Isaaks & Srivastava 1989) :

$$\widehat{\gamma}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{(i,j)|\mathsf{h}_{ij}=\mathsf{h}} \left(z_i - z_j\right)^2 \tag{3.10}$$

Le variogramme expérimental standardisé (3.9) apparaît donc comme un cas particulier de la forme générale (3.8).

[4] Contrairement à l'usage de plus en plus répandu, la terminologie correcte est *demi-variogramme*, le variogramme étant en fait 2γ (Isaaks & Srivastava 1989, Chauvet 1994).

3.2.1.3 Interprétation

Le c de Geary n'admet pas de borne supérieure mais admet une borne inférieure évidente $\min(c) = 0$ qui correspond à l'autocorrélation spatiale maximale, *i.e.* lorsque les valeurs comparées sont identiques. Par construction, le c de Geary est un rapport d'estimations de la variance respectivement spatiale et a-spatiale. Par conséquent, la valeur attendue du c de Geary en absence d'autocorrélation spatiale est $\mathrm{E}[c] = 1$ (Geary 1954, Cliff & Ord 1981). Les valeurs comprises dans l'intervalle $[0, 1[$ témoignent donc d'une autocorrélation spatiale positive tandis que les valeurs dans $]1, +\infty[$ traduisent la présence d'autocorrélation négative.

La matrice de pondération $\mathbf{W}$ joue exactement le même rôle que la matrice $\mathbf{X}$ dans les tests utilisant la corrélation entre deux matrices (Section 3.1). Ainsi, choisir la forme de $\mathbf{W}$ revient à proposer *a priori* un modèle de variation spatiale dont il est possible de tester l'adéquation aux données (Jumars *et al.* 1977, Cliff & Ord 1981, p. 168, 174).

3.2.2 I de Moran

Moran considère les phénomènes distribués dans un espace bidimensionnel discrétisé par n supports répartis selon une grille d'indices $i = 1, \ldots, p$ et $j = 1, \ldots, q$, et propose de mesurer la corrélation entre plus proches voisins comme (Moran 1950)[5] :

$$r_{11} = \frac{pq}{2pq - p - q} I \tag{3.11}$$

$$I = \frac{A + B}{\sum_{i,j} (z_{ij} - \overline{z})^2} \tag{3.12}$$

avec

$$A = \sum_{i=1}^{p} \sum_{j=1}^{q-1} (z_{ij} - \bar{z})(z_{i,j+1} - \bar{z}) \tag{3.13}$$

$$B = \sum_{i=1}^{p-1} \sum_{j=1}^{q} (z_{ij} - \bar{z})(z_{i+1,j} - \bar{z}) \tag{3.14}$$

Dans I (3.12) figurent $p(q-1) + (p-1)q = 2pq - p - q$ termes au numérateur et pq termes au dénominateur de sorte que r_{11} (3.11) consiste simplement à diviser le numérateur et le dénominateur de I par le nombre de termes correspondants. Notons que le terme A concerne la relation de voisinage entre colonnes (pour les lignes $i = 1, \ldots, p$) tandis que le terme B concerne la relation entre lignes (pour les colonnes $j = 1, \ldots, q$) : implicitement, le graphe de voisinage considéré est donc 4-connexe. Considérons à présent le cas général où n supports sont répartis de façon quelconque.

[5] Chessel & Sabatier (1993), Hansen (1994) et Chessel *et al.* (1997) citent à tort l'article de Moran (1948) qui définit en fait une autre statistique, pour des données binaires.

3.2.2.1 Généralisation du I de Moran

De la même façon que pour le c de Geary, la statistique (3.11) se généralise en considérant une matrice $\mathbf{W}$ de poids w_{ij} affectés à tous les couples de supports (i, j)(Cliff & Ord 1981) :

$$I = \frac{n}{W} \frac{\sum_{i,j} w_{ij} (z_i - \bar{z}) (z_j - \bar{z})}{\sum_i (z_i - \bar{z})^2} \tag{3.15}$$

avec $W = \sum w_{ij}$ et n le nombre total de valeurs. Bien que la statistique (3.15) soit classiquement désignée comme le I *de Moran* elle ne correspond pas à l'expression originale de I (3.12) mais plutôt à r_{11} (3.11). Dans sa forme générale (3.15), le I de Moran[6] se présente donc comme un coefficient d'autocorrélation faisant intervenir au numérateur une autocovariance spatiale et au dénominateur l'autocovariance a-spatiale, *i.e.* la variance des valeurs.

3.2.2.2 Fonction de covariance et I de Moran

En considérant en particulier les $N(\mathbf{h})$ couples de supports (i, j) séparés par un vecteur $\mathbf{h}_{ij} = \mathbf{h}$, le I de Moran s'écrit :

$$I(\mathbf{h}) = \frac{\widehat{C}(\mathbf{h})}{s_n^2} \tag{3.16}$$

avec $s_n^2 = n^{-1} \sum (z_i - \bar{z})^2$ et $\widehat{C}(\mathbf{h})$ la fonction de covariance :

$$\widehat{C}(\mathbf{h}) = \frac{1}{N(\mathbf{h})} \sum_{(i,j)|\mathsf{h}_{ij}=\mathsf{h}} (z_i - \bar{z}) (z_j - \bar{z}) \tag{3.17}$$

3.2.2.3 Interprétation

Le I de Moran ne s'interprète pas exactement comme un coefficient de corrélation classique en ce qu'il n'est pas restreint à l'intervalle $[-1, +1]$ et ne vaut pas exactement 0 en absence d'autocorrélation spatiale. En général, la borne supérieure $|I|$ est inférieure à 1, bien qu'elle puisse dépasser 1 dans certains cas particuliers (Cliff & Ord 1981, p. 21). La valeur attendue du I de Moran en absence d'autocorrélation spatiale est $\mathrm{E}[I] = -(n-1)^{-1}$ (Moran 1950, Cliff & Ord 1981). Une valeur inférieure à $\mathrm{E}[I]$ traduit de l'autocorrélation négative tandis qu'une valeur supérieure témoigne d'une autocorrélation positive.

Les commentaires au sujet de la matrice de pondération $\mathbf{W}$ de la Section 3.2.1.3 s'appliquent évidemment au I de Moran.

3.2.3 Test des indices d'autocorrélation

Le c de Geary et le I de Moran sont tous les deux asymptotiquement normalement distribués sous l'hypothèse nulle H_0 d'absence d'autocorrélation. Afin de tester les valeurs observées en faisant référence à la loi normale, les moments des deux statistiques sont

[6] Nous désignons désormais par I *de Moran* l'expression générale donnée par Cliff & Ord (1981).

donnés par Cliff & Ord (1981). En ce qui concerne le c de Geary, la validité de l'approximation normale est discutée dans Cliff & Ord (1981). La normalité de la distribution du I de Moran sous H_0 a donné lieu à une littérature plus abondante que dans le cas du c de Geary, citons notamment Cliff & Ord (1971), Sen (1976), Cliff & Ord (1977), Sen (1977), Cliff & Ord (1981), et plus récemment, Terui & Kikuchi (1994) et Waldhör (1996).

Un test permutationnel s'avère particulièrement intéressant dans le cas où le nombre de supports est faible ($n \leq 10$). Pour des valeurs de n couramment rencontrées ($n > 10$) le test permutationnel est approximé par un test de randomisation. Le principe est le même que dans la Section 3.1.4 sauf que les permutations concernent directement les valeurs z_i. Un nombre élevé de permutations aléatoires — 10^3, 10^4 ou 10^5, selon la valeur de n — permet d'obtenir une estimation fiable de la *p-value*, calculée comme en (3.4).

Dans le cas d'une matrice $\mathbf{W}$ correspondant à une classe de vecteurs $\mathbf{h}$, le c de Geary et le I de Moran sont équivalents, respectivement :

- au variogramme standardisé par s^2_{n-1} (3.9),
- à la fonction de covariance standardisée par s^2_n (3.16).

Ainsi, dans le cadre d'un test permutationnel ou d'un test de randomisation, le variogramme (3.10) et la fonction de covariance (3.17) sont des statistiques équivalentes, respectivement au c de Geary et au I de Moran.

L'assertion selon laquelle il n'est pas possible de tester le variogramme (Legendre & Troussellier 1988, Legendre & Fortin 1989, Rossi 1996) est erronée[7] : toute statistique mesurant l'autocorrélation spatiale peut être testée au moyen d'un test de randomisation.

3.2.4 Comparaison du c de Geary et du I de Moran

Le c de Geary et le I de Moran peuvent être appliqués aussi bien à des données quantitatives qu'à des données binaires (Geary 1954, Cliff & Ord 1981), bien que des statistiques spécifiques aient été développées dans le cas des données binaires (*e.g.*, Moran 1948). Le variogramme (3.10) étant équivalent au c de Geary — à un facteur de standardisation près — l'assertion selon laquelle il ne peut être utilisé qu'avec des données quantitatives (Legendre & Fortin 1989, Rossi *et al.* 1995) est erronée, le variogramme étant largement utilisé dans le cas des indicatrices (Carr *et al.* 1985, Rossi *et al.* 1992, Liebhold *et al.* 1993, Hohn *et al.* 1993, Halvorson *et al.* 1995, 1996, Sharov *et al.* 1995, Goovaerts *et al.* 1997, Juang & Lee 1998a, Western *et al.* 1998). Il en est évidemment de même en ce qui concerne la fonction de covariance (3.17). L'applicabilité du c de Geary et du I de Moran étant identique du point de vue de la structure algébrique de la variable, la comparaison entre les deux statistiques peut porter sur :

- leur définition opératoire,
- leur sensibilité à la forme de la distribution statistique des valeurs,
- leur efficacité et leur puissance.

[7] Cette erreur est corrigée par Legendre (1993) qui indique que le variogramme peut être testé *via* le test développé pour le c de Geary, puisque le c de Geary lui est étroitement lié.

3.2.4.1 Relation entre le c de Geary et le I de Moran

Considérons les définitions générales (3.8) et (3.15). Soit T la quantité (Sokal 1979a) :

$$T = \frac{(n-1)}{2W} \frac{\sum_i (w_{i.} + w_{.i}) (z_i - \bar{z})^2}{(z_i - \bar{z})^2} \tag{3.18}$$

avec $w_{i.}$ la somme des poids pour la ligne i et $w_{.i}$ la somme des poids pour la colonne i. Le c de Geary et le I de Moran sont liés par la relation $c = T - I(n-1)/n$. En définissant $d = 1 - c$ et $J = I + (n-1)^{-1}$ on obtient deux statistiques qui varient dans le même sens et dont l'espérance est nulle.

On peut d'autre part écrire $1 - T = 1 - c - I(n-1)/n$ et $d - J = 1 - c - I - (n-1)^{-1}$. Si n est grand, alors $(n-1)^{-1} \simeq 0$ et $(n-1)/n \simeq 1$, ce qui conduit à $d - J \simeq 1 - T$. Ainsi, étudier la différence entre d et J, et par extension la différence entre c et I, revient à étudier la quantité T (Sokal 1979a). Sous l'hypothèse de normalité et sans autocorrélation spatiale, l'espérance de T est $\mathrm{E}[T] = (n-1)/n$. En considérant n grand, alors $\mathrm{E}[T] \simeq 1$.

Interpréter les différences de comportement entre c et I nécessite d'interpréter l'écart de T à son espérance sous l'hypothèse nulle. Le numérateur de T est la somme du produit de deux termes :

- $(w_{i.} + w_{.i})$, d'autant plus élevé que le support i est spatialement important, *i.e.* fortement connecté aux autres supports,
- $(z_i - \bar{z})^2$, d'autant plus élevé que la valeur z_i est une valeur extrême par rapport à la moyenne.

T est par conséquent d'autant plus élevé ($T > 1$) que des sites fortement connectés avec les autres sites présentent des valeurs extrêmes, *i.e.* lorsque les valeurs extrêmes présentent une structure spatiale (Sokal 1979a). Si l'autocorrélation spatiale est positive alors $d > 0$ et $J > 0$, et si $T > 1$ alors $d < J$. Si l'autocorrélation spatiale est négative alors $d < 0$ et $J < 0$, et si $T > 1$ alors $|d| < |J|$. Autrement dit, si le c de Geary montre une autocorrélation spatiale positive moins forte que le I de Moran, ou une autocorrélation spatiale négative plus forte que le I de Moran, alors cela peut indiquer que les valeurs extrêmes sont spatialement structurées, au sens de la matrice de pondération $\mathbf{W}$ utilisée (Sokal 1979a).

3.2.4.2 Sensibilité à la forme de la distribution

Des résultats donnés dans Cliff & Ord (1969, *op. cit.* Cliff & Ord 1981) suggèrent que la variance du I de Moran est moins affectée par la distribution statistique des données que le c de Geary. L'expérience nous a montré que le c de Geary et le I de Moran conduisent à des résultats similaires dans le cas d'une distribution symétrique, mais qu'une forte asymétrie — telle que celle d'une distribution exponentielle négative (*e.g.*, le diamètre des arbres) — peut entraîner un désaccord entre les deux statistiques. Ce désaccord est d'autant plus marqué que l'autocorrélation spatiale est ténue, et le c de Geary s'avère davantage affecté par l'asymétrie de la distribution que le I de Moran.

Remarque 1 *L'influence du choix de la statistique sur la conclusion du test de l'autocorrélation spatiale est un phénomène inévitable mais indésirable. En conséquence, nous recommandons d'utiliser les deux statistiques conjointement :*

- *une conclusion similaire constitue l'assurance que le résultat n'est pas artefactuel, et est bien indépendant de la statistique utilisée,*
- *une conclusion contradictoire doit nécessairement conduire l'écologiste à approfondir son analyse.*

3.2.4.3 Efficacité et puissance

Qu'il s'agisse de l'efficacité asymptotique ou de la puissance, le I de Moran apparaît globalement supérieur au c de Geary, mais l'avantage de l'un sur l'autre reste assez faible (Cliff & Ord 1981 p. 170, 176). Les résultats obtenus sont valides uniquement dans le cas d'une distribution normale, mais Cliff & Ord (1981, p. 178) indiquent que la forme de la matrice $\mathbf{W}$ semble plus importante que le type de données. Cependant nous avons précisé dans la section précédente que l'asymétrie de la distribution des valeurs affecte davantage le c de Geary que le I de Moran, la situation de référence étant une distribution symétrique telle que la loi normale. Une étude de la puissance comparée du c de Geary et du I de Moran dans le cas d'une distribution nettement asymétrique conclurait vraisemblablement en faveur du I de Moran.

3.3 Fonctions d'autocorrélation spatiale

Pour un même jeu de données, la corrélation entre matrices de proximités (*e.g.*, la statistique de Mantel) ou les indices d'autocorrélation (*e.g.*, le I de Moran) peuvent être calculés en faisant varier la définition du voisinage des supports. Il existe deux situations selon que le voisinage est défini en faisant intervenir :

- un graphe de voisinage des supports (éventuellement pondéré),
- les vecteurs formés par les couples de supports (le plus souvent dans un espace muni de la métrique euclidienne).

Si les distances euclidiennes n'ont pas beaucoup de sens, par exemple lorsque la densité relative des supports varie fortement et/ou que les erreurs de localisation ne sont pas négligeables (*e.g.*, McFadden & Aydin 1996), il convient d'utiliser à la place un graphe de voisinage représentant la topologie des connexions entre supports (Legendre & Fortin 1989). Le problème est alors de savoir si la question écologique posée permet de définir ce graphe. Dans le cas de supports ponctuels repérés le long d'une rivière ou d'une ligne de côte, le graphe est naturellement réduit à une simple chaîne (Royaltey *et al.* 1975, Chessel & Sabatier 1993, Thioulouse *et al.* 1995). En revanche, dans le cas de supports ponctuels répartis dans le plan (*e.g.*, les arbres d'une parcelle), la définition du graphe de voisinage est rarement déduite de la question écologique mais est généralement de nature strictement géométrique.

La définition géométrique du graphe de voisinage la plus courante est celle de la triangulation de Delaunay[8] (DT), *i.e.* le graphe d'adjacence du diagramme de Voronoï associé au semis des supports (Section 2.2.2.5, p. 24). Il est préférable de supprimer les arêtes extérieures qui connectent généralement des supports éloignés, ce qui peut se faire au niveau du diagramme de Voronoï lui-même, en supprimant les polygones extérieurs ou marginaux (*e.g.*, Mercier 1997, pp. 65-66). Le graphe de Gabriel est parfois recommandé (Gabriel & Sokal 1969, Royaltey *et al.* 1975,) notamment parce qu'il ne contient pas de longues arêtes telles que celles dues aux triangles très aplatis de la DT (Matula & Sokal 1980). A mesure que l'on considère des sous-graphes de la DT de moins en moins denses — graphe de Gabriel, graphe de voisinage relatif, puis arbre de poids minimum euclidien — il est clair que les résultats de l'analyse d'autocorrélation s'en trouvent de plus en plus modifiés.

Le problème sous-jacent peut parfois indiquer le graphe le plus approprié, mais en pratique il est généralement difficile de présenter des arguments en faveur d'un graphe particulier (Gordon & Finden 1985). L'inconvénient d'une définition strictement géométrique des graphes de voisinage est d'ignorer les caractéristiques des organismes, ainsi que l'existence de rivières, lacs, océans, montagnes, déserts, etc. (Gabriel & Sokal 1969). En conséquence, il peut s'avérer nécessaire de modifer un graphe de voisinage géométrique en tenant compte d'un modèle biologique tel que la direction d'un flux génique ou des barrières écologiques (Sokal & Oden 1978a). Autrement dit, il peut y avoir des raisons d'éditer le graphe de voisinage pour supprimer des arêtes ou en ajouter d'autres (*e.g.*, Sokal *et al.* 1989a, Fig. 2). Dans ce contexte, il faudrait considérer des unions ou des différences entre graphes (Gordon & Finden 1985), ce qui nécessite de recourir à des opérateurs ensemblistes, internes à la classe des graphes.

3.3.1 Voisinage dans un graphe

Dans le cas d'une mise en relation topologique par un graphe de voisinage, il est possible de considérer chaque support et ses voisins à l'ordre 1, *i.e.* les supports que l'on peut joindre en empruntant une seule arête. C'est notamment la définition du voisinage qui est utilisée par Geary (1954) pour une carte choroplèthe, et par Moran (1950) pour une grille de points munie de la relation de 4-connexité.

La notion de voisinage peut être généralisée à l'ordre k, où k peut prendre toutes les valeurs jusqu'à une limite, par exemple $\max(k) \leq n^{-1} \sum \mathrm{e}\,(x_i)$, avec $\mathrm{e}\,(x_i)$ l'écartement du sommet x_i dans le graphe de voisinage (Section 2.1.2.4, p. 16). A l'ordre $k = 1$ on compare en moyenne la similarité entre chaque support et ses voisins directs, puis à l'ordre $k = 2$ entre chaque support et ses voisins directs plus les voisins directs de ceux-ci, etc. En faisant varier k entre 1 et $\max(k)$, l'autocorrélation spatiale est décrite comme une fonction de la taille du voisinage dans le graphe. Le cas d'un graphe dont les arêtes sont pondérées se traite de la même façon, seul le calcul de la statistique étant affecté par la pondération. Il est également possible de considérer une relation de voisinage non symétrique, ce qui revient à utiliser des graphes orientés, et par conséquent des matrices asymétriques de voisinage ou de pondération.

[8] Par exemple, Kenkel *et al.* (1997) utilisent la triangulation de Delaunay afin de tester l'autocorrélation spatiale du diamètre (DBH, *Diameter at Breast Height*) du pin *Pinus banksiana,* dans le sud-est du Manitoba.

3.3.2 Vecteurs inter-supports

Lorsque le voisinage des supports est défini par leurs vecteurs de séparation, il est possible de calculer l'autocorrélation spatiale comme une fonction des normes de ces vecteurs, autrement dit, comme une fonction des distances entre supports. Ces distances peuvent être des valeurs exactes ou des intervalles, selon la répartition des supports et les choix opératoires qui sont faits. En effet, si les données sont réparties de façon parfaitement régulière (*e.g.*, sur une grille à maille carrée), il faut choisir entre :

- des distances multiples de la maille,
- des classes de distances définies par les multiples de la maille et une tolérance.

Lorsque la répartition des supports est irrégulière, il est pratiquement impossible d'utiliser des distances exactes sous peine de ne pas trouver un nombre suffisant de couples de supports qui leur correspondent : dans ce cas, le recours à des classes de distances est indispensable.

Selon les mesures d'autocorrélation utilisées, les fonctions obtenues sont des corrélogrammes (statistique de Mantel standardisée et I de Moran), des covariogrammes (covariance), ou des variogrammes (variogramme et c de Geary).

Dans tous les cas, les fonctions sont dites *omnidirectionnelles* puisqu'elles font intervenir uniquement la norme des vecteurs séparant les supports. De la même façon qu'il est possible d'utiliser des classes de distances, des classes d'angles peuvent être définies afin d'obtenir des fonctions *directionnelles*, le résultat pouvant s'exprimer lui-même sous la forme d'une carte (Oden & Sokal 1986, Isaaks & Srivastava 1989, pp. 140-183, Rossi *et al.* 1992). Cependant, cette pratique nécessite de disposer de nombreux supports parce qu'il faut obtenir un nombre de couples suffisant pour chaque classe (Oden & Sokal 1986, Rossi *et al.* 1992). L'interprétation des cartes montrant l'autocorrélation dans toutes les directions est relativement difficile et nécessite généralement de se reporter à une représentation cartographique des données elles-mêmes (Oden & Sokal 1986, Rossi *et al.* 1992).

3.3.3 Fonctions non ergodiques

Dans le cadre du voisinage défini par les vecteurs inter-supports, deux nouvelles fonctions dites *non ergodiques*[9] ont été introduites par Isaaks & Srivastava (1988) : la covariance non ergodique $\widehat{C}_D$ et la corrélation non ergodique $\widehat{\rho}_D$.

Au lieu de la formule classique (3.17), la covariance non ergodique $\widehat{C}_D$ s'écrit (Isaaks & Srivastava 1989) :

$$\widehat{C}_D(\mathbf{h}) = \frac{1}{N(\mathbf{h})} \sum_{(i,j)|\mathsf{h}_{ij}=\mathsf{h}} (z_i - m_{-\mathsf{h}})(z_j - m_{+\mathsf{h}}) \tag{3.19}$$

avec $N(\mathbf{h})$ le nombre de couples de supports séparés par le vecteur $\mathbf{h}$, $m_{-\mathsf{h}}$ la moyenne de toutes les valeurs dont les supports sont situés à $-\mathbf{h}$ des autres :

$$m_{-\mathsf{h}} = \frac{1}{N(\mathbf{h})} \sum_{i|\mathsf{h}_{ij}=\mathsf{h}} z_i \tag{3.20}$$

[9] L'origine de cette terminologie est donnée dans la Section 4.2.1.2, et dans le Chapitre 7.

et $m_{+\mathsf{h}}$ la moyenne de toutes les valeurs dont les supports sont situés à $+\mathbf{h}$ des autres :

$$m_{+\mathsf{h}} = \frac{1}{N(\mathbf{h})} \sum_{j|\mathsf{h}_{ij}=\mathsf{h}} z_j \tag{3.21}$$

Les valeurs de $m_{-\mathsf{h}}$ et de $m_{+\mathsf{h}}$ sont généralement différentes (Isaaks & Srivastava 1989, p. 59). La corrélation non ergodique $\widehat{\rho}_D$ est la covariance non ergodique standardisée par des écarts-types appropriés :

$$\widehat{\rho}_D(\mathbf{h}) = \frac{\widehat{C}_D(\mathbf{h})}{\sigma_{-\mathsf{h}}\sigma_{+\mathsf{h}}} \tag{3.22}$$

avec $\sigma_{-\mathsf{h}}$ l'écart-type de toutes les valeurs dont les supports sont situés à $-\mathbf{h}$ des autres :

$$\sigma^2_{-\mathsf{h}} = \frac{1}{N(\mathbf{h})} \sum_{i|\mathsf{h}_{ij}=\mathsf{h}} (z_i - m_{-\mathsf{h}})^2 \tag{3.23}$$

et $\sigma_{+\mathsf{h}}$ l'écart-type de toutes les valeurs dont les supports sont situés à $+\mathbf{h}$ des autres :

$$\sigma^2_{+\mathsf{h}} = \frac{1}{N(\mathbf{h})} \sum_{j|\mathsf{h}_{ij}=\mathsf{h}} (z_j - m_{+\mathsf{h}})^2 \tag{3.24}$$

De même que pour les moyennes, les valeurs de $\sigma_{-\mathsf{h}}$ et de $\sigma_{+\mathsf{h}}$ sont généralement différentes (Isaaks & Srivastava 1989, p. 60). Ainsi, la nouveauté introduite par les fonctions non ergodiques est la prise en compte d'éventuelles différences :

- entre moyennes locales grâce à la distinction entre $m_{-\mathsf{h}}$ et $m_{+\mathsf{h}}$,
- entre variances locales grâce à la distinction entre $\sigma^2_{-\mathsf{h}}$ et de $\sigma^2_{+\mathsf{h}}$.

Pour comparer ces deux fonctions entre elles et avec le variogramme standardisé $c(\mathbf{h})$ (3.9) il faut utiliser les transformations suivantes (Rossi *et al.* 1992, Liebhold *et al.* 1993) :

- $c^{(1)}(\mathbf{h}) = \left\{ s^2 - \widehat{C}_D(\mathbf{h}) \right\} / s^2$ pour la covariance non ergodique,
- $c^{(2)}(\mathbf{h}) = 1 - \widehat{\rho}_D(\mathbf{h})$ pour la corrélation non ergodique.

Après transformation, les fonctions varient dans le même sens que le variogramme standardisé et ont même espérance sous H_0 ($\mathrm{E}[c(\mathbf{h})] = 1$). Selon le couple de fonctions considéré, une superposition imparfaite des représentations graphiques montre donc des différences de moyennes locales ($c(\mathbf{h})$ et $c^{(1)}(\mathbf{h})$), de variances locales ($c^{(1}(\mathbf{h})$ et $c^{(2)}(\mathbf{h})$), ou éventuellement des deux à la fois ($c(\mathbf{h})$ et $c^{(2)}(\mathbf{h})$).

3.3.4 Test des fonctions d'autocorrélation

Les fonctions d'autocorrélation spatiale peuvent faire l'objet de trois types de tests, selon que l'on considère :

- les valeurs d'autocorrélation calculées pour chaque voisinage ou chaque classe,
- l'égalité entre deux fonctions d'autocorrélation,
- la fonction d'autocorrélation elle-même, d'un point de vue global.

3.3.4.1 p-gramme

Quelle que soit la statistique utilisée, chaque valeur calculée dans la fonction d'autocorrélation peut être testée individuellement au moyen d'un test de randomisation (Sections 3.1.4 & 3.2.3). En ce qui concerne les fonctions de vecteurs inter-supports, le test de randomisation consiste à réaliser un grand nombre de permutations aléatoires de deux valeurs z_i et z_j, puis à recalculer à chaque fois la fonction entière, *i.e.* pour toutes les classes. Cet algorithme naïf est rapidement inutilisable en pratique et nous avons conçu un algorithme optimisé qui recalcule uniquement les quantités affectées par chaque permutation, et procède ensuite à une mise à jour de la fonction, au cas par cas. Après 10^3, 10^4 ou 10^5 permutations aléatoires — selon la taille du jeu de données — les *p-values* sont calculées pour chaque classe. La démarche classique consiste ensuite à fixer un seuil arbitraire (*e.g.*, $\alpha = 0.05$) pour différencier les valeurs d'autocorrélation qui sont significatives à ce seuil. La *p-value* étant vue comme la force de l'évidence contre l'hypothèse nulle, il est préférable de visualiser toutes les *p-values* afin d'interpréter avec finesse la fonction d'autocorrélation spatiale associée. A cet effet, nous proposons — indépendamment de Walker *et al.* (1997) — de représenter les *p-values* comme des fonctions, conjointement aux fonctions d'autocorrélation spatiale. Walker et *al.* (1997) nomment la fonction des *p-values* un *p-gramme* et nous utilisons par la suite cette terminologie.

Pour faciliter la comparaison entre une fonction d'autocorrélation et le *p*-gramme qui lui est associé, il est souhaitable de les faire varier dans le même sens. Par exemple, dans le cas d'une fonction représentée sous la forme d'un variogramme, il faut associer une faible *p-value* à l'autocorrélation positive et une forte *p-value* à l'autocorrélation négative, et la *p-value* associée à une valeur observée $f(\mathbf{h})_{obs}$ est estimée par :

$$\widehat{p}(\mathbf{h}) = \frac{\mathrm{Card}\left(\{f(\mathbf{h}) \mid f(\mathbf{h}) \in \Omega,\ f(\mathbf{h}) \leq f(\mathbf{h})_{obs}\}\right)}{\mathrm{Card}\left(\Omega\right)} \tag{3.25}$$

avec la valeur observé $f(\mathbf{h})_{obs}$ figurant dans l'ensemble Ω des valeurs de $f(\mathbf{h})$ pour les permutations considérées.

3.3.4.2 Comparaison des fonctions

A notre connaissance, il n'existe pas de statistique permettant de tester l'égalité entre deux fonctions du même type. Pour comparer deux fonctions $f_1(\mathbf{h})$ et $f_2(\mathbf{h})$ comportant k classes, on se contente généralement d'apprécier leur ressemblance (Legendre & Fortin 1989). Cette ressemblance peut s'exprimer indifféremment sous la forme d'une similarité globale ou d'une dissimilarité globale, par exemple en calculant la distance moyenne entre les valeurs en utilisant la métrique L_1 ou *distance de Manhattan* (Sokal 1986, Bocquet-Appel & Sokal 1989, Fortin *et al.* 1989, Sokal & Jacquez 1991) :

$$d_{L_1}(f_1, f_2) = \frac{1}{k}\sum_{i=1}^{k} |f_{1i} - f_{2i}| \tag{3.26}$$

En comparant un ensemble de N fonctions, cette approche produit une matrice symétrique de $N \times N$ distances qui peut être soumise à un algorithme de *classification ascendante hiérarchique* (CAH) afin de produire un *dendrogramme* (revues dans Diday *et al.* 1982, Benzécri 1984, Legendre & Legendre 1984b, Roux 1985).

Une autre approche consiste à résumer l'ensemble des N fonctions au moyen d'une ACP dans laquelle les "individus" sont les fonctions et les "variables" les classes, chaque fonction étant vue commme un hyperpoint dans un espace à k dimensions (Diaz *et al.* 1997). Dans le cas de fonctions exhibant une *portée*, *i.e.* une distance a à partir de laquelle l'autocorrélation est constamment non significative, il faut s'attendre à ce que les "variables" soient assez fortement redondantes au-delà de a.

Pour pallier l'absence de test d'égalité entre deux fonctions $f_1(\mathbf{h})$ et $f_2(\mathbf{h})$, nous proposons de résumer la structure de chaque fonction par une matrice de similarités croisant toutes les valeurs entre elles, puis de tester l'égalité de ces matrices au moyen d'un test de Mantel (Section 3.1.4). Le test peut porter sur la fonction d'autocorrélation spatiale ou sur son p-gramme. La seconde option est préférable dans la mesure où le p-gramme a l'avantage de donner une image accentuée de la variation de l'autocorrélation tout en tenant compte intrinsèquement du nombre de couples dans chaque classe. Des tests effectués sur un ensemble de N fonctions conduisent à une matrice symétrique $N \times N$ des *p-values* qui peut être soumise à un algorithme de CAH.

A la différence de l'approche utilisant la distance de Manhattan, le recours au test que nous proposons a l'avantage de tenir compte de la structure globale de chaque fonction et ne se contente pas de résumer les ressemblances mesurées indépendamment, classe par classe.

3.3.4.3 Signification globale

Au lieu d'étudier individuellement les valeurs d'une fonction d'autocorrélation, il est envisageable de s'intéresser à la fonction d'un point de vue global. Il s'agit alors de savoir si la structure d'autocorrélation spatiale est significative et vaut par conséquent la peine d'être interprétée (Cliff & Ord 1981).

Oden (1984) envisage plusieurs méthodes pour tester un corrélogramme de façon globale, notamment une modification du test *portemanteau* Q utilisé pour les séries temporelles (Hosking 1986), et les corrections de Šidák ou de Bonferroni (Alt 1982, Sokal & Rohlf 1995). En considérant la procédure de Bonferroni, un corrélogramme comportant k coefficients est jugé significatif au seuil α si au moins un coefficient est significatif au seuil $\alpha' = \alpha/k$. Oden (1984) conclut que la méthode de Bonferroni (ou celle de Šidák) est simple et préférable au test Q modifié lorsqu'il y a peu de classes de distances et une faible structure spatiale. L'utilisation de la correction de Bonferroni telle qu'elle est suggérée par Oden (1984) suppose que tous les coefficients sont d'importance égale puisque l'allocation d'erreur est constante (α/k). En fait, cette égalité n'est pas nécessaire puisqu'il faut simplement respecter (Alt 1982) :

$$\sum_{i=1}^{k} \alpha_i = \alpha \tag{3.27}$$

avec α_i le risque affecté au coefficient de la classe i dans le corrélogramme.

La principale objection à l'utilisation de la correction de Bonferroni proposée par Oden (1984) réside dans la nécessité de fixer un seuil α, souvent impossible à justifier, et susceptible de conduire à une importante perte de puissance (*cf.* Hinkley 1987). En outre, le résultat du test dépend fortement du nombre de coefficients k et du degré de structuration spatiale des données. Considérons par exemple un corrélogramme dont la portée a

est légèrement supérieure à la distance moyenne entre les données, ce qui apparaît comme la traduction d'une faible structure spatiale, telle celle mentionnée par Oden (1984). Dans ces conditions, le corrélogramme présentera peu de valeurs en deçà de la portée, et les coefficients ne seront certainement pas très éloignés de l'espérance sous H_0. Si k est petit et α modeste (*e.g.*, $\alpha = 0.05$), alors il est possible que parmi les coefficients des classes en deçà de la portée, il s'en trouve au moins un significatif au seuil $\alpha' = \alpha/k$, auquel cas on concluera que le corrélogramme est significatif. Mais, même dans cette situation "favorable", il suffirait d'augmenter le nombre de coefficients pour qu'aucun des coefficients des classes en deçà de la portée ne soit plus significatif au seuil α', d'où une conclusion opposée à la précédente.

Théoriquement, sous l'hypothèse nulle d'absence d'autocorrélation à toutes les classes, les fonctions sont parfaitement horizontales et s'écrivent :

- $c(\mathbf{h})_{H_0} = 1$ pour le variogramme standardisé et toutes les fonctions représentées sous cette forme,
- $I(\mathbf{h})_{H_0} = -(n-1)^{-1}$ pour le corrélogramme du I de Moran.

Tester si une fonction montre globalement de l'autocorrélation spatiale reviendrait ainsi à tester la fonction observée $f(\mathbf{h})_{obs}$ par rapport à la fonction attendue sous H_0, $f(\mathbf{h})_{H_0}$. En acceptant le test d'égalité entre deux fonctions tel qu'il est proposé dans la section précédente, il suffirait alors de tester l'égalité entre $f(\mathbf{h})_{obs}$ et $f(\mathbf{h})_{H_0}$.

En pratique, pour des jeux de données de taille modeste (*e.g.*, $n \leq 100$), une fonction qui témoigne d'une absence de structure spatiale, présente plutôt des oscillations autour de l'espérance sous H_0, et les *p-values* du *p*-gramme associé oscillent autour de leur valeur moyenne. Pour tester l'absence de structure spatiale significative, il semble alors judicieux de tester le caractère aléatoire de ces oscillations. Dans cette optique, Hossaert-McKey *et al.* (1996) proposent d'utiliser un *run test* (Sokal & Rohlf 1995) par rapport à la médiane afin de tester l'hypothèse nulle selon laquelle les valeurs positives ou négatives du I de Moran sont distribuées aléatoirement le long du corrélogramme. Indépendamment de Hossaert-McKey *et al.* (1996), nous proposons de juger du caractère significatif d'une fonction $f(\mathbf{h})_{obs}$ en testant la corrélation entre valeurs successives, de préférence en considérant le *p*-gramme associé parce que les *p-values* sont directement comparables entre elles. La statistique de corrélation sérielle que nous proposons s'écrit simplement :

$$C = \sum_{i=1}^{k-1} (p_i - \bar{p})(p_{i+1} - \bar{p}) \tag{3.28}$$

avec k le nombre de classes, p_i la *p-value* associée à la valeur de la classe i et $\bar{p}$ la moyenne des *p-values* du *p*-gramme. La statistique (3.28) peut être testée au moyen d'un test de permutation lorsque k est faible ($k \leq 10$), ce qui est généralement le cas. La *p-value* est alors calculée comme la proportion de valeurs supérieures ou égales à la valeur observée :

$$p = \frac{\text{Card}(\{C \mid C \geq C_{obs}\})}{k!} \tag{3.29}$$

Pour $k > 10$, il suffit de substituer un test de randomisation au test de permutation, et la *p-value* est estimée par :

$$\widehat{p} = \frac{\text{Card}(\{C \mid C \in \Omega,\ C \geq C_{obs}\})}{\text{Card}(\Omega)} \tag{3.30}$$

avec la valeur observée C_{obs} figurant dans l'ensemble Ω des valeurs de C pour les permutations considérées.

Une faible *p-value* traduit la présence d'une corrélation entre les valeurs successives du p-gramme. Autrement dit, il existe une structure au sein du p-gramme, et par conséquent, au sein de la fonction d'autocorrélation associée. Au contraire, une *p-value* élevée (*e.g.*, $p = 0.30$, $p = 0.50$) témoigne d'une absence de structure globale. Dans ce cas, même si certaines *p-values* peuvent s'avérer significatives individuellement, elles n'apparaissent pas structurées au sein du p-gramme.

3.4 Corrélation et décorrélation spatiales

L'autocorrélation spatiale traduit la relation entre la distribution spatiale des supports et la distribution statistique des valeurs. Pour bien comprendre la notion d'autocorrélation spatiale, il est donc fondamental d'examiner comment les distributions spatiale et statistique des valeurs conditionnent la forme de l'autocorrélation. Pour ce faire, il convient d'étudier comment l'autocorrélation peut être générée, modifiée ou supprimée. Il faut d'abord distinguer trois modes d'action selon que l'on modifie :

- l'autocorrélation spatiale elle-même, au sens d'une certaine définition opératoire,
- la distribution spatiale des valeurs,
- la distribution statistique des valeurs.

Considérons une grille carrée 30×30 de maille $\Delta = 1$ comportant au total $N = 900$ noeuds auxquels sont affectées au hasard des valeurs $z_i^{(0)} \sim \mathcal{N}(0, 1)$ avec $i = 1, \ldots, N$ (Fig. 3.1.0a & 3.1.0b). Les valeurs ont été affectées aux noeuds dans un ordre aléatoire de sorte que la VR $z^{(0)}(\cdot)$ correspondante n'est pas spatialement autocorrélée (Fig. 3.1.0c).

Imposons à $z^{(0)}(\cdot)$ une structure d'autocorrélation spatiale définie par une matrice de covariance spatiale $\mathbf{C}$, symétrique et définie positive (Annexe D), dont les éléments sont $c_{ij} = C(h_{ij})$, avec h_{ij} la distance entre deux noeuds (i, j). Une technique directe pour corréler spatialement les données selon le modèle représenté par $\mathbf{C}$ nécessite deux étapes. La première étape consiste à factoriser la matrice de covariance $\mathbf{C}$ en un produit de deux matrices, par exemple en utilisant la décomposition de Cholesky $\mathbf{C} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$ (Golub & van Loan 1983). La seconde étape consiste alors à calculer (Ripley 1981) :

$$\mathbf{z}^{(1)} = \mathbf{L}\mathbf{z}^{(0)} \tag{3.31}$$

avec $\mathbf{z}^{(0)} = \left(z_i^{(0)} \mid i = 1, \ldots, N\right)^{\mathrm{T}}$ et $\mathbf{z}^{(1)} = \left(z_i^{(1)} \mid i = 1, \ldots, N\right)^{\mathrm{T}}$. En utilisant cette procédure et une matrice $\mathbf{C}$ définie positive arbitraire, la VR obtenue $z^{(1)}(\cdot)$ présente une distribution spatiale régulière (Fig. 3.1.1a). Cette VR est à présent nettement spatialement autocorrélée comme en témoigne son variogramme standardisé (Fig. 3.1.1c). Imposer la structure d'autocorrélation spatiale définie par $\mathbf{C}$ s'est également traduit par une modification de la distribution statistique, qui ne correspond plus à une loi normale (Fig. 3.1.1b).

Il est possible de décorréler les valeurs de $z^{(1)}(\cdot)$ en effectuant la transformation réciproque :

$$\mathbf{z}^{(0)} = \mathbf{L}^{-1}\mathbf{z}^{(1)} \tag{3.32}$$

Ce type de décorrélation permet de retrouver exactement[10] la distribution statistique de $z^{(0)}(\cdot)$, soit, dans le cas présent, une distribution normale. D'une façon générale, toute transformation du type (3.31) modifie l'autocorrélation spatiale et par conséquent, à la fois la distribution spatiale et la distribution statistique de la VR d'origine.

Une autre façon de décorréler $z^{(1)}(\cdot)$ consiste à permuter aléatoirement les valeurs parmi les supports. La VR $z^{(2)}(\cdot)$ qui résulte des permutations des valeurs de $z^{(1)}(\cdot)$ est non autocorrélée (Fig. 3.1.2a & 3.1.2c), mais la distribution statistique est conservée (Fig. 3.1.2b). Il est possible d'utiliser une approche permutationnelle pour imposer à $z^{(2)}(\cdot)$ une structure d'autocorrélation spatiale donnée par un modèle, grâce à une procédure d'optimisation combinatoire, mais le coût de cette opération réciproque est bien plus élevé que celui de la décorrélation puisque cela revient à faire diminuer l'entropie d'un système.

Enfin, toute manipulation de la distribution statistique modifie l'autocorrélation spatiale ; par exemple, la transformation des valeurs de $z^{(1)}(\cdot)$ par anamorphose (Annexe E) conduit à une VR $z^{(3)}(\cdot)$ différente (Fig. 3.1.3a, 3.1.3b & 3.1.3c). Dans ce cadre, des procédures de corrélation/décorrélation sont logiquement possibles mais certainement très difficiles à concevoir. De telles procédures semblent *a priori* sans utilité et ne présentent qu'un intérêt théorique.

Il apparaît donc de façon évidente que toute modification de la distribution spatiale ou de la distribution statistique entraîne une modification de l'autocorrélation spatiale, conformément à sa définition. En outre, la modification de l'autocorrélation spatiale par permutation des valeurs parmi les supports montre que les VR $z^{(1)}(\cdot)$ et $z^{(2)}(\cdot)$ sont différentes alors qu'en contexte a-spatial elles ne peuvent pas être distinguées puisque leurs distributions statistiques sont identiques[11].

Plus précisément, pour une VR définie de façon discrète sur N supports, les relations logiques entre les deux types de distributions sont :

- une distribution spatiale donnée $\Rightarrow$ une seule distribution statistique possible,
- une distribution statistique donnée $\Rightarrow$ $N!$ distributions spatiales possibles.

En définissant de façon opératoire l'autocorrélation spatiale, par exemple au sens du variogramme (3.10), alors ces relations deviennent :

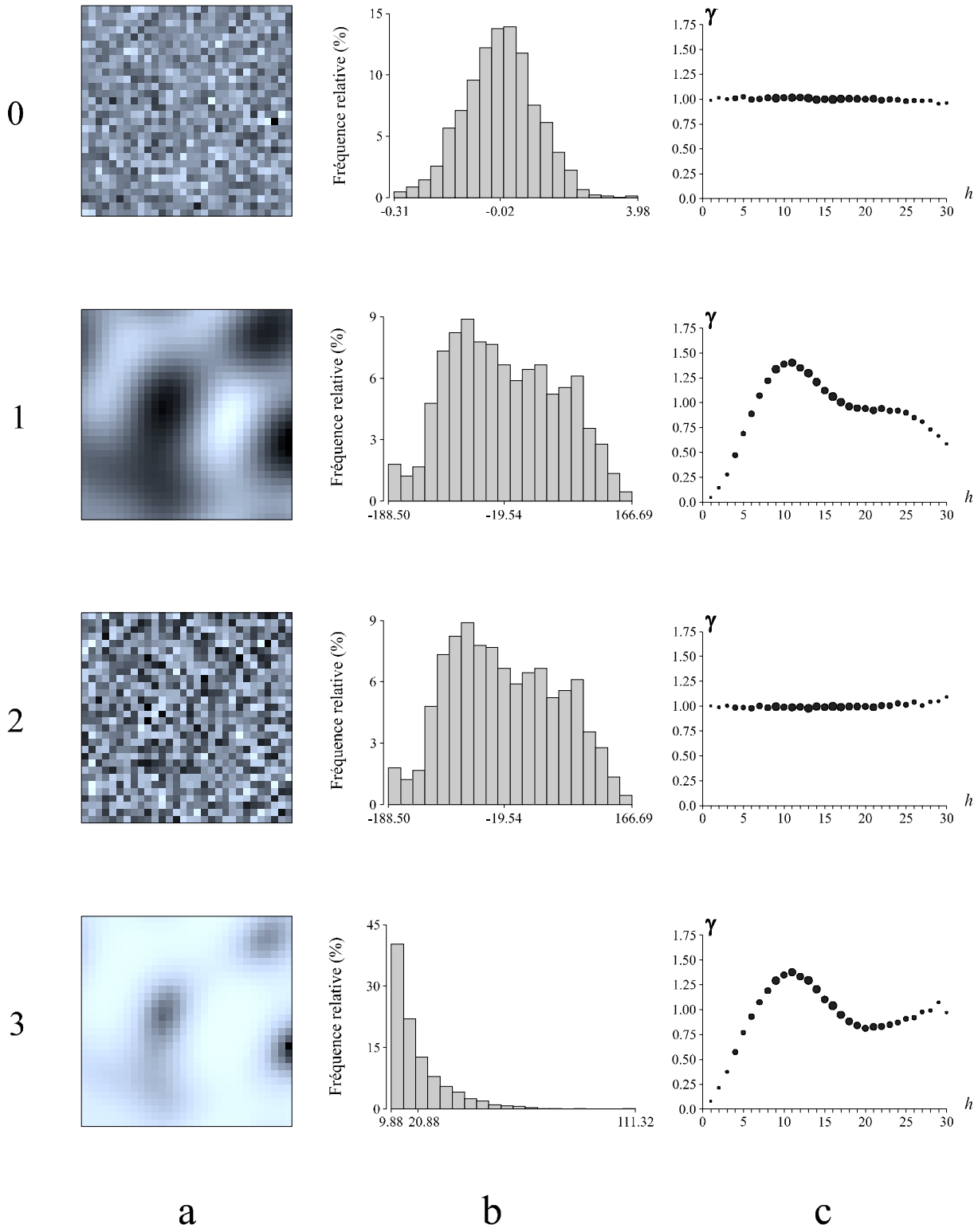
- un variogramme expérimental donné $\Rightarrow$ une distribution statistique, définie à une constante près,
- une distribution statistique donnée $\Rightarrow$ $N!$ variogrammes possibles.

La conclusion de cette section visant à clarifier la nature de l'autocorrélation spatiale peut s'énoncer sous la forme de la proposition suivante :

Proposition 1 *Il n'existe pas de relation bijective entre, d'une part, la distribution statistique des valeurs d'une variable régionalisée $z(\cdot)$, et d'autre part, la distribution spatiale de ces valeurs ou l'autocorrélation spatiale de $z(\cdot)$.*

[10] On retrouve les valeurs d'origine, aux imprécisions numériques près.

[11] Les méthodes qui n'exploitent que la distribution statistique pour résumer les structures spatiales échouent forcément à distinguer deux distributions spatiales différentes d'un même ensemble de valeurs (Jumars *et al.* 1977).



Figure˜3.1: Relations entre la distribution spatiale, la distribution statistique, et l'autocorrélation spatiale des variables régionalisées (0), (1), (2) et (3). (a) Image de la distribution spatiale. (b) Histogramme des valeurs. (c) Variogramme standardisé. Détails dans le texte.

3.5 Analyse de l'autocorrélation spatiale

L'analyse de l'autocorrélation spatiale en écologie ou en biologie évolutive a été introduite indépendamment par Jumars *et al.* (1977) et Sokal & Oden (1978a, 1978b). Par la suite, des articles de synthèse tels que ceux de Legendre & Fortin (1989) et de Rossi *et al.* (1992) ont mis l'accent sur les outils provenant, respectivement, de la géographie statistique (Cliff & Ord 1981), et d'un certain courant de la géostatistique (Isaaks & Srivastava 1988). De nombreux domaines sont concernés par l'analyse de l'autocorrélation spatiale, notamment :

- la génétique des populations, discipline au sein de laquelle l'analyse de la répartition spatiale des fréquences de gènes (ou éléments génétiques assimilables) est reconnue comme hautement informative, et où le nombre d'études utilisant des statistiques d'autocorrélation spatiale ne cesse d'augmenter (Sokal & Oden 1978a, 1978b, Sokal & Wartenberg 1983, Sokal *et al.* 1989a, 1989b, Slatkin & Arter 1991, Epperson 1990, 1995, Epperson & Li 1996, 1997, Epperson & Alvarez-Buylla 1997, Hossaert-McKey *et al.* 1996, McFadden & Aydin 1996, Doligez & Joly 1997, Mahy & Nève 1997, Caujapé-Castells & Pedrola-Monfort 1997, Petit *et al.* 1997, Sokal *et al.* 1997, Sokal & Thomson 1998, Epperson *et al.* 1999, etc.),
- l'écologie des populations d'arthropodes (Pont 1986, Schotzko & O'Keeffe 1989, Rossi *et al.* 1992, Midgarden *et al.* 1993, Liebhold *et al.* 1993),
- l'écologie des populations de nématodes (Webster & Boag 1992a, 1992b, Delaville *et al.* 1996, Rossi *et al.* 1996, Rossi & Quenehervé 1998),
- l'écologie microbienne des sols (Dandurand *et al.* 1995, 1997),
- la phytopathologie (Larkin *et al.* 1995).

L'objet de cette section est de porter un regard critique sur la pratique de l'analyse de l'autocorrélation spatiale. Pour simplifier, nous considérons que les données concernent une variable quantitative unique. Dans un premier temps, nous discutons de l'interprétation des tests de l'autocorrélation spatiale. Par la suite, nous traitons exclusivement des études locales au moyen des fonctions d'autocorrélation, ce qui nous amène à examiner :

- le choix de la fonction,
- le choix entre une analyse omnidirectionnelle et une analyse directionnelle,
- les tests de signification,
- l'identification de la portée.

Enfin, nous faisons quelques recommandations générales quant à la façon de mener à bien l'analyse de l'autocorrélation spatiale.

3.5.1 Interprétation des résultats des tests

Il est évident que le résultat d'un test d'autocorrélation spatial est à la fois sensible à la définition opératoire utilisée (Upton & Fingleton 1985) et conditionnel à l'échelle spatiale à laquelle on se situe (Qi & Wu 1996). En particulier, pour une statistique calculée en utilisant un graphe de voisinage, la manière dont on connecte les supports conditionne la

nature de l'autocorrélation spatiale testée[12] (Gabriel & Sokal 1969, Sokal & Oden 1978a). La question est alors de savoir ce que l'on doit inférer si l'autocorrélation s'avère statistiquement significative en utilisant une certaine définition opératoire (*e.g.*, une certaine matrice de poids $\mathbf{W}$) et non significative avec une autre définition (Gatrell 1979). La réponse est simplement qu'un résultat plus ou moins significatif met en évidence l'adéquation plus ou moins bonne du modèle de variation spatial imposé *a priori* par $\mathbf{W}$, avec les données sous étude. Dans cette optique, plutôt que de tester différents modèles de variation spatiale, il est possible de chercher directement la matrice $\mathbf{W}$ qui maximise la statistique utilisée (*e.g.*, le I de Moran) (Kooijman 1976, Boots & Dufournaud 1994). Cependant, il reste ensuite à interpréter la signification du modèle de variation spatiale obtenu, ce qui peut s'avérer assez difficile.

Afin d'illustrer les problèmes d'interprétation soulevés par les tests de l'autocorrélation spatiale, nous donnons un exemple de deux formes opératoires distinctes du test de Mantel. Sokal (1979b) a exhumé le test de Mantel (1967) afin de tester de façon globale des modèles de variation spatiale en biologie, écologie, systématique, etc. Auparavant, Royaltey *et al.* (1975) avaient conçu un test du même type que celui de Mantel (1967), tout en étant plus compliqué. En pratique, le test de Royaltey-Astrachan-Sokal est rarement utilisé (*e.g.*, Costa *et al.* 1992), et nous ne le mentionnons que pour mémoire.

La définition opératoire de la statistique de Mantel (3.2) nécessite de construire une matrice de proximités entre supports ($\mathbf{X}$), ainsi qu'une matrice de similarités entre valeurs ($\mathbf{Y}$). Pour une variable quantitative, nous proposons d'utiliser une similarité entre valeurs de type *covariance centrée* $Y_{ij} = (z_i - \bar{z})(z_j - \bar{z})$, avec $\bar{z}$ la moyenne des valeurs. Le calcul de $\mathbf{Y}$ étant effectué, le résultat du test de Mantel dépend de la définition retenue pour la matrice de proximité $\mathbf{X}$. Par exemple, Douglas & Endler (1982) définissent la proximité d'après un graphe de Gabriel valué par la distance euclidienne, et imposent la plus grande distance pour les couples de supports qui ne sont pas connectés dans le graphe. Il existe beaucoup d'autres façons de définir la proximité des supports d'après un graphe de voisinage. Dans ce qui suit, nous ne considérons pas un graphe de voisinage mais faisons référence à deux proximités basées sur la distance euclidienne h_{ij} entre les supports x_i et x_j :

1. la proximité $X_{ij} = 1 - h_{ij} / \max(h_{ij})$ qui varie linéairement avec h_{ij},
2. la proximité $X_{ij} = 1/h_{ij}$ qui ne varie pas linéairement avec h_{ij} et met l'accent sur les faibles distances.

Les *p-values* obtenues à l'issue des tests utilisant respectivement les proximités 1 et 2 n'ont aucune raison d'être du même ordre de grandeur, et les conclusions peuvent diverger, parce que le modèle de variation spatiale testé est différent :

- variation spatiale du type cline ou gradient pour 1,
- variation spatiale à faible distance pour 2.

Considérons par exemple une structure spatiale pseudo-périodique formée d'agrégats de valeurs élevées, séparés par des zones de faibles valeurs. Dans ce cas, le test utilisant la

[12] De ce point de vue, l'analyse de l'autocorrélation spatiale pose le même type de difficultés que la mesure de l'agrégation pour laquelle il existe plusieurs méthodes mesurant en fait des choses différentes (*cf.* Pielou 1969, p. 90).

définition 1 peut s'avérer incapable de détecter une autocorrélation spatiale significative tandis que celui utilisant la définition 2 peut donner un résultat significatif. En effet, à basse distance — au sein des agrégats ou des zones intermédiaires — les valeurs sont semblables, mais en considérant globalement toutes les distances, l'autocorrélation spatiale positive intra-agrégat et intra-zone est compensée par de l'autocorrélation spatiale négative entre les agrégats et les zones intermédiaires. Le résultat peut donc être non significatif du point de vue du premier test, tout en étant significatif pour le second.

Il convient d'ajouter à la difficulté d'interprétation des tests d'autocorrélation bien établis, le problème posé par les tests dont la validité statistique est suspecte. Un exemple récent est fourni par le test d'autocorrélation spatiale proposé en écologie par Koenig & Knops (1998) sous le prétexte que l'analyse de l'autocorrélation est peu développée, ce qui témoigne pour le moins d'une méconnaissance du sujet traité. Koenig (1998) applique ce nouveau test à l'abondance des oiseaux de Californie, pour 88 espèces documentées par le BBS (*Breeding Bird Survey*) et pour 79 espèces documentées par le CBC (*Christmas Bird Count*). Koenig (1998) conclut que 87 espèces sur les 88 du BBS ne présentent pas d'autocorrélation significative, pour les quatre classes de distances considérées. En outre, seulement 27 espèces sur les 79 du CBC montrent de l'autocorrélation significative. Ces résultats sont suspects. En effet, Koenig (1998) conclut par exemple que l'espèce *Molothrus ater* ne montre pas d'autocorrélation spatiale significative alors que Maurer (1994, Fig. 5.4, p. 87), également à partir des données du BBS, cartographie son abondance par krigeage universel (Section 6.2.1.5, p. 138), ce qui indique que le variogramme pour cette espèce montre de l'autocorrélation spatiale. Toujours à partir des données du BBS, il est surprenant que Koenig (1998) ne trouve pas d'autocorrélation significative pour *Tyrannus verticalis* (*Western Kingbird*) alors que le variogramme de *Tyrannus tyrannus* (*Eastern Kingbird*) atteste de la présence d'une forte structure d'autocorrélation spatiale (Maurer 1994, Fig. 4.1, p. 60). Il en est de même pour les piverts dont aucune espèce ne présente d'autocorrélation spatiale significative d'après Koenig (1998), alors que les variogrammes des deux piverts figurés dans Villard & Maurer (1996) témoignent du contraire. En ajoutant aux résultats surprenants obtenus par Koenig (1998) le fait que la description — non formalisée — du test de Koenig & Knops (1998) est assez inintelligible, nous considérons, sans examen plus approfondi, que le test utilisé n'est tout simplement pas valide.

3.5.2 Choix de la fonction d'autocorrélation

Nous avons mentionné plusieurs types de fonctions d'autocorrélation, essentiellement le c de Geary (3.9) ou le variogramme (3.10), le I de Moran (3.16) ou la fonction de covariance (3.17), et les fonctions de covariance (3.19) ou de corrélation (3.22) dites *non ergodiques*. Toutes ces fonctions ont été utilisées en pratique, le I de Moran étant d'utilisation courante en biologie évolutive et les fonctions géostatistiques (variogramme, covariance non ergodique) plus particulièrement utilisées en écologie des populations (Annexe G). L'utilisation privilégiée du I de Moran en biologie évolutive, et du variogramme ou de la covariance non ergodique en écologie des populations est pure contingence : toutes les fonctions peuvent être utilisées, indépendamment de la discipline.

Dans une analyse de l'autocorrélation locale, la principale difficulté consiste à interpréter la forme de la fonction obtenue. Bien souvent, l'interprétation canonique d'un type de fonction suppose certaines caractéristiques absentes des données, notamment :

- la symétrie de la distribution statistique,
- une certaine forme d'homogénéité spatiale, *i.e.* l'absence de tendance marquée, de variations locales de la moyenne ou de la variance, ou d'*outliers* spatiaux[13].

3.5.2.1 Asymétrie de la distribution

Les données écologiques exhibent des distributions statistiques allant de l'asymétrie positive extrême telle que celle de la loi exponentielle négative, à la symétrie de la loi normale. Une approche souvent utilisée consiste à réduire l'asymétrie de la distribution, par exemple par transformation logarithmique (*e.g.*, Legendre *et al.* 1997), ou au moyen d'une transformation de Box-Cox (*e.g.*, Rossi & Quenehervé 1998). Dans cette optique, l'anamorphose gaussienne constitue sans aucun doute la meilleure technique (Annexe E). Cependant, dans la Section 3.4 nous avons montré que toute modification de la distribution statistique entraîne une modification de l'autocorrélation spatiale. En conséquence, en toute rigueur les conclusions de l'analyse d'autocorrélation spatiale ne concernent plus la variable d'origine mais la nouvelle variable résultant de la transformation. Comme il n'existe pas de transformation réciproque applicable à la fonction d'autocorrélation spatiale de la variable transformée, en toute rigueur il n'est pas possible d'étendre les résultats à la variable d'origine. En pratique, la situation est la suivante :

- soit la transformation a peu d'impact sur le résultat de l'analyse, ce qui signifie que ladite transformation n'était pas indispensable,
- soit la transformation change profondément le résultat de l'analyse, ce qui montre bien que ce n'est plus la variable d'origine qui est étudiée.

Plutôt que de transformer les données, nous recommandons d'utiliser des fonctions de type covariance ou I de Moran, moins sensibles à une distribution asymétrique que ne le sont le variogramme ou le c de Geary (Section 3.2.4, p. 46).

3.5.2.2 Présence d'une tendance

Le prérequis d'homogénéité spatiale est souvent mis en défaut du fait de la variation de la moyenne locale, voire même de la variance locale. Le problème est particulièrement sérieux dans le cas du variogramme parce que la présence d'une tendance affecte considérablement son interprétation (Starks & Fang 1982). Le variogramme est un résumé incomplet de la structure spatiale et peut conduire à des erreurs d'interprétation lorsque les moyennes locales et/ou les variances locales changent (Rossi *et al.* 1992).

L'approche classique pour traiter ce type de problème consiste à éliminer la tendance par régression, puis à étudier l'autocorrélation spatiale du résidu (*e.g.*, Bocquet-Appel & Sokal 1989). A nouveau, ce n'est plus la variable d'origine qui est étudiée mais un résidu qui peut fort bien ne plus présenter d'intérêt du point de vue phénoménologique.

[13] Les *outliers* spatiaux ne sont pas des valeurs aberrantes ni nécessairement des valeurs extrêmes dans la distribution statistique des données, mais des valeurs extrêmes par rapport aux valeurs voisines.

Il est préférable de ne pas modifier les données mais d'utiliser une fonction d'autocorrélation qui tienne compte explicitement des différences de moyennes locales — covariance non ergodique (3.19) — voire même des différences de variances locales — corrélation non ergodique (3.22).

3.5.2.3 Présence d'outliers spatiaux

Les *outliers* spatiaux ont un effet destructurant sur les fonctions d'autocorrélation, perturbant l'analyse au point de conclure faussement à une faible structure spatiale (Brockmans & Murray 1997), voire même à l'absence d'autocorrélation spatiale (Cook & Coles 1997). Par exemple, en analysant des données[14] décrivant les dégâts occasionnés par le puceron *Diuraphis noxia* dans dix champs de blé (Colorado, USA), nous avons constaté que la suppression de seulement 3% d'*outliers* spatiaux — en leur substituant la valeur moyenne — pouvait totalement changer les résultats de l'analyse.

Afin de réduire les difficultés causées par les *outliers* spatiaux et permettre une analyse plus fine, une approche consiste à décomposer la variable quantitative en une série de k indicatrices $\{i(x, z_j) \mid j = 1, \ldots, k\}$ définies par un ensemble de valeurs seuil $\{z_j \mid j = 1, \ldots, k\}$ (Carr *et al.* 1985, Halvorson *et al.* 1995) :

$$i(x, z_j) = \begin{cases} 1 & \text{si } z(x) \geq z_j \\ 0 & \text{sinon} \end{cases} \tag{3.33}$$

Evidemment, cette "dissection" de la variable d'origine se traduit par une augmentation du temps consacré à l'analyse, mais ce dernier reste généralement largement inférieur au temps consommé par la collecte des données.

Il est sans doute plus simple d'identifier les *outliers* spatiaux au moyen d'outils d'analyse exploratoire (Sections 7.1.2 & 7.1.3), ou tout simplement d'après une représentation cartographique des données, puis de réaliser les analyses avec et sans les *outliers* spatiaux présumés, afin d'en apprécier l'impact.

3.5.3 Analyse omnidirectionnelle vs. directionnelle

Le calcul de fonctions directionnelles constitue le summum du raffinement dans l'analyse de l'autocorrélation spatiale. En contrepartie, ce type d'analyse requiert des données suffisamment abondantes (Annexe G). Ainsi, il est recommandé de disposer d'au moins $n = 300$ données afin de calculer des variogrammes directionnels (Oliver *et al.* 1989b).

Néanmoins, avec seulement $n = 100$ données il est possible de calculer une fonction pour quelques directions principales, en utilisant des classes d'angles suffisamment larges. Typiquement, les variogrammes directionnels sont calculés pour quatre directions, à 0°, 45°, 90° et 135° (Annexe G).

[14] Données communiquées par le Dr. Xavier Fauvergue.

3.5.4 Tests de signification

Interpréter une fonction d'autocorrélation spatiale consiste essentiellement à comparer les valeurs qui la composent. D'une façon tout à fait similaire à la corrélation entre deux variables (*cf.* Rodriguez 1982), il ne faudrait pas interpréter les différences numériques entre les valeurs de la fonction d'autocorrélation sans faire un test de signification. Ce test peut porter sur chaque valeur de la fonction ou sur la fonction dans son ensemble.

3.5.4.1 Test de signification locale

Lorsque l'ensemble des données n'est pas de très grande taille ($n < 1000$), comme c'est souvent le cas en écologie, une fonction d'autocorrélation peut être soumise à un test de randomisation sans aucune difficulté en termes de temps de calcul (Section 3.3.4.1). A toute fonction d'autocorrélation, il est donc généralement possible en pratique d'associer un p-gramme calculé sur la base d'un minimum de 10^3 permutations aléatoires. Les *p-values* du p-gramme permettent de juger objectivement de l'intensité de l'autocorrélation pour chaque distance ou classe de distances. L'interprétation de la structure d'autocorrélation spatiale devrait donc s'effectuer davantage par examen du p-gramme que par examen de la fonction d'autocorrélation elle-même. Dans ce contexte, il n'y a aucun sens à imposer un seuil de signification arbitraire (*e.g.*, $\alpha = 0.05$). En revanche, il faut faire porter son attention sur la continuité de la variation des *p-values* au sein du p-gramme. En effet, la variation progressive des *p-values* témoigne objectivement de l'augmentation ou de la diminution progressive de l'intensité de l'autocorrélation spatiale en fonction de la distance.

Le p-gramme se révèle être un outil qui facilite considérablement l'interprétation des faibles structures spatiales, soit que le phénomène étudié est lui-même faiblement structuré, soit que l'échelle de l'étude ne correspond pas exactement à celle du phénomène.

3.5.4.2 Test de signification globale

La pratique du test global d'une fonction d'autocorrélation au moyen de la correction de Bonferroni a été diffusée en écologie par Sokal & Thomson (1987) et en biologie évolutique par Sokal *et al.* (1987) (Annexe G).

Si un test de signification globale s'avère vraiment nécessaire à la discussion, plutôt que d'utiliser la correction de Bonferroni (ou celle de Šidák), nous recommandons de tester la corrélation sérielle au sein de la fonction ou du p-gramme associé à la fonction (Section 3.3.4.3, p. 53).

3.5.5 Identification de la portée

Au-delà de la question de savoir si l'autocorrélation est significative ou pas, ce qui intéresse le plus l'écologiste c'est de mesurer la portée de l'autocorrélation afin de déterminer l'échelle de la dépendance spatiale (Schlesinger *et al.* 1996, Koenig & Knops 1998). Lorsque l'autocorrélation positive décroît avec la distance puis s'annule complètement, la portée est définie de façon stricte comme la distance à partir de laquelle l'autocorrélation est non

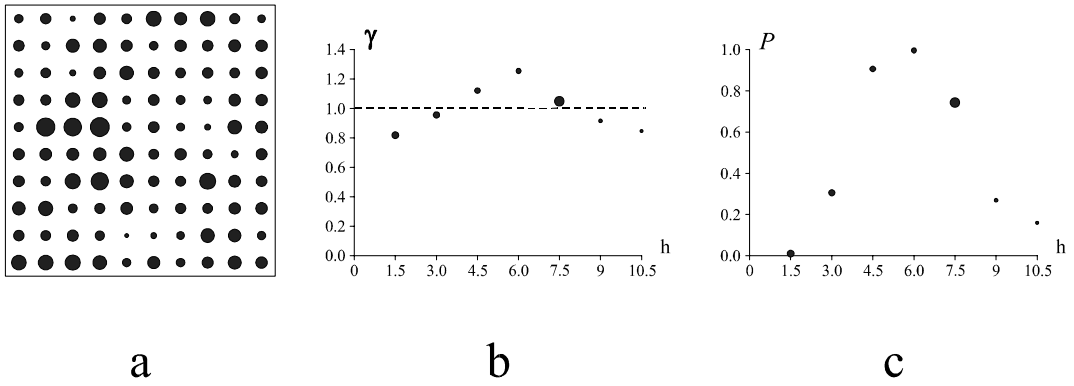
significative. Lorsque l'autocorrélation décroît, s'annule, puis change de signe, la notion de portée perd de sa signification.

Classiquement, les fonctions issues de la géostatistique sont considérées comme utiles pour mesurer la portée de l'autocorrélation spatiale (Rossi *et al.* 1992). En fait, le corrélogramme du I de Moran peut tout aussi bien être utilisé qu'un variogramme ou que la covariance non ergodique, mais dans tous les cas, la définition opératoire de la portée constitue un problème délicat. Il convient tout d'abord de remarquer que la définition opératoire de la portée dépend du type de fonction d'autocorrélation, et qu'un résultat différent peut être obtenu selon que l'on utilise le c de Geary (ou le variogramme) ou le I de Moran (ou la covariance) : la notion de portée n'échappe évidemment pas au problème de la définition opératoire de l'autocorrélation spatiale elle-même. Par exemple, dans leur étude de la teneur en cuivre et en zinc dans les sédiments du bassin du Yangtze (Chine), Zhang & Selinus (1997) mettent en évidence une portée d'environ 1000 km avec le variogramme et de 500 km avec le corrélogramme du I de Moran.

Nous proposons — indépendamment de Walker *et al.* (1997) — d'utiliser le p-gramme afin de déterminer le plus objectivement possible la portée de l'autocorrélation spatiale. A nouveau, le choix d'un seuil arbitraire α au-delà duquel l'autocorrélation spatiale serait jugée non significative n'a pas de sens. En effet, le choix de α influence évidemment la définition de la portée (Walker *et al.* 1997). Il est plus judicieux d'examiner la variation des *p-values* au sein du p-gramme : lorsque plusieurs *p-values* successives plaident en faveur de l'hypothèse H_0, alors c'est que la portée de l'autocorrélation a été atteinte ou a été légèrement dépassée.

3.5.6 Exemple

Afin d'illustrer l'analyse et le test de l'autocorrélation spatiale, nous considérons un jeu de données composé d'une grille de supports ponctuels 10×10, de pas $\Delta = 1.5$ unités. La représentation graphique des données $\{z_i \mid i = 1, \ldots, 100\}$ témoigne de la présence d'une structure spatiale (Fig. 3.2.a).



Figure˜3.2: Exemple d'analyse de l'autocorrélation spatiale. (a) Représentation cartographique des données réparties selon une grille 10×10. La taille des cercles est proportionnelle aux valeurs. (b) Variogramme standardisé. (c) p-gramme du variogramme, obtenu à partir de 10^5 permutation aléatoires des données.

Cette structure peut être testée globalement à l'aide du Z de Mantel (3.2), du c de Geary (3.8), ou du I de Moran (3.15), puis analysée localement en calculant, par exemple, le semivariogramme standardisé (3.9).

Effectuons un test de Mantel en utilisant la promixité spatiale $X_{ij} = 1/h_{ij}$ et la proximité statistique $Y_{ij} = (z_i - \bar{z})(z_j - \bar{z})$. Un test de randomisation portant sur 10^5 permutations aléatoires donne $p = 0.01163$, ce qui témoigne de la présence d'autocorrélation positive à faible distance. Le c de Geary et le I de Moran sont calculés à l'ordre 1 dans le graphe de voisinage 4-connexe associé à la grille des supports. Les valeurs observées $c_{obs} \simeq 0.82$ et $I_{obs} \simeq 0.21$ sont testées sur la base de 10^5 permutations aléatoires, ce qui conduit aux *p-values* respectives $p = 0.00984$ et $p = 0.00184$. Les tests du c de Geary et du I de Moran montrent que les valeurs voisines sont positivement autocorrélées.

Le variogramme standardisé est calculé jusqu'à une distance maximale $h_{\max} = 10.5$ unités (Fig. 3.2.b). Cette fonction montre que l'autocorrélation spatiale est positive à faible distance ($h = 1.5$), s'annule, change de signe ($h = 6$), puis s'annule à nouveau et redevient positive ($h = 10.5$). Afin de tester chaque valeur de la fonction, un p-gramme est calculé sur la base de 10^5 permutations aléatoires (Fig. 3.2.c). La *p-value* en $h = 1.5$ vaut $p = 0.01000$, ce qui correspond bien à la *p-value* du c de Geary calculé entre voisins 4-connexes ($p = 0.00984$). L'autocorrélation négative est significative en $h = 6$ ($p = 1 - 0.99596 = 0.00444$), mais l'autocorrélation positive n'est pas significative en $h = 10.5$ ($P = 0.15883$). S'il fallait définir la portée de l'autocorrélation positive d'après le p-gramme, nous la situerions approximativement entre $h = 3.0$ et $h = 4.5$.

Enfin, s'il s'avérait utile de tester globalement le variogramme standardisé, la procédure de Bonferroni utilisée avec $\alpha = 0.05$ conclurait au caractère significatif puisque la plus petite *p-value* ($p = 0.00444$) est inférieure au seuil corrigé $\alpha' = \alpha/7 \simeq 0.00714$. Les tests de permutation de la corrélation sérielle portant sur la fonction ou sur le p-gramme donnent des résultats similaires, en accord avec la procédure de Bonferroni ($p = 0.04643$ et $p = 0.05615$).

3.5.7 Recommandations

L'analyse des données est relativement bon marché tandis que les données écologiques sont coûteuses à acquérir. Dans ces conditions, il semble logique de consacrer au moins autant d'efforts à l'analyse qu'à la collecte des données. Ainsi, Jumars *et al.* (1977) considèrent que la meilleure approche consiste à utiliser conjointement plusieurs méthodes applicables de façon valide au problème à traiter. En effet, si différentes méthodes raisonnables conduisent à des résultats convergents, alors il devient possible d'accorder quelque crédit à l'analyse de l'autocorrélation spatiale.

Nous déconseillons de se limiter à une étude de l'autocorrélation globale lorsque les données permettent d'effectuer une étude locale au moyen d'une fonction d'autocorrélation. Dans le cas d'une étude globale, il faut être attentif à la définition opératoire du Z de Mantel, du c de Geary, ou du I de Moran, afin de comprendre quel est le modèle de variation spatiale sous-jacent. Nous conseillons d'examiner l'impact sur les résultats des différents choix opératoires qui sont faits, ainsi que d'éventuels *outliers*.

Dans le cas des fonctions d'autocorrélation, il est évidemment redondant d'utiliser à la fois le c de Geary et le variogramme, ou à la fois le I de Moran et la covariance. En revanche, il est judicieux de comparer les résultats obtenus avec le c de Geary d'une part, et le I de Moran de l'autre — ou leurs équivalents respectifs. En effet, il n'est pas exceptionnel que le variogramme et le I de Moran conduisent à des conclusions différentes (*e.g.*, Midgarden *et al.* 1993). Actuellement, la démarche qui semble la plus judicieuse consiste à calculer simultanément le variogramme standardisé (c de Geary), la covariance et la corrélation non ergodiques, afin de procéder aux comparaisons décrites Section 3.3.3 (p. 50) (Liebhold *et al.* 1993, Sharov *et al.* 1996). Enfin, le calcul du p-gramme associé à une fonction d'autocorrélation devrait être systématiquement effectué afin de permettre une interprétation fine et objective.

Chapitre 4

Modèles géostatistiques

> *"Given the results of a survey, what is the error of predicting the results of another similar survey? "* (Osborne 1942)

En géostatistique, il est possible de distinguer classiquement deux niveaux de modélisation des phénomènes régionalisés :

- le modèle primaire, objet de la *géostatistique transitive*,
- le modèle probabiliste, objet de la *géostatistique intrinsèque*.

Les liens qui existent entre la géostatistique transitive et la géostatistique intrinsèque ne sont pas évoqués ici puisque la géostatistique transitive est hors de notre propos (*cf.* Matheron 1965, 1978, Chauvet 1994, El Bahi 1981) et dans ce qui suit, nous considérons toujours un domaine d'étude $D \subset \mathbb{R}^2$ de géométrie fixée *a priori*.

Comme il n'est pas possible de traiter de tous les aspects de la géostatistique intrinsèque dans le cadre d'un seul chapitre, nous nous contentons de présenter le passage du modèle primaire ou modèle probabiliste, ainsi que les éléments fondamentaux de géostatistique qui sont nécessaires à la compréhension des chapitres suivants. Des compléments peuvent être trouvés notamment dans Matheron (1965, 1969, 1978, 1982), David (1977), Journel & Huijbregts (1978), Delfiner & Matheron (1980), Journel (1989), Isaaks & Srivastava (1989), Deutsch & Journel (1992), Christakos (1992), Chauvet (1994), Rivoirard (1991, 1994) et Goovaerts (1997).

4.1 Modèle primaire

En géostatistique, le premier niveau de modélisation mathématique considère que les valeurs $\{z_i \mid i = 1, \ldots, n\}$ observées aux temps $\{t_i \mid i = 1, \ldots, n\}$ sur un ensemble de supports $\{x_i \mid i = 1, \ldots, n\}$ dans un domaine D sont structurées selon une fonction déterministe $z(x, t)$. Le concept de fonction est utilisé uniquement pour indiquer qu'en un point x, et à un instant t, il ne peut y avoir qu'une seule valeur $z(x, t)$. Nous considérons implicitement que les valeurs sont mesurées ou observées au même instant t de sorte que nous ne traitons pas la dimension temporelle : nous écrivons simplement $z(x)$ pour faire référence à la valeur en x, et $z(\cdot)$ pour désigner la fonction elle-même.

Jusqu'à présent, aucune hypothèse n'est faite quant aux propriétés mathématiques de $z(\cdot)$ telles que sa continuité ou sa dérivabilité. En fait, la fonction $z(\cdot)$ n'est pas nécessairement une fonction analytique mais est définie plus généralement comme une *variable régionalisée* (VR) (Matheron 1965). Les manipulations effectuées sur les variables régionalisées peuvent également être appliquées au cas particulier des fonctions analytiques, mais la réciproque n'est pas nécessairement vraie. C'est précisément parce que les phénomènes régionalisés naturels ne pouvaient pas être modélisés de façon satisfaisante par des fonctions analytiques que Matheron a proposé la *théorie des variables régionalisées* ou *géostatistique* (Matheron 1965).

4.2 Modèle probabiliste

Afin de traduire la variation spatiale de la variable régionalisée sous la forme d'un modèle probabiliste (ou stochastique), il faut considérer :

1. que chaque valeur $z(x)$ est une réalisation d'une variable aléatoire (VA) $Z(x)$ pour tout point $x \in D$,
2. que les variables aléatoires $\{Z(x) \mid x \in D\}$ ne sont pas indépendantes les unes des autres mais liées entre elles par une structure de corrélation.

L'ensemble des variables aléatoires $\{Z(x) \mid x \in D\}$ constitue une *fonction aléatoire* (FA) notée $Z(\cdot)$, qui est le pendant probabiliste de la variable régionalisée $z(\cdot)$. Le modèle de fonction aléatoire étant défini à la fois dans un espace topologique et dans un espace probabilisé, il est parfois qualifié de *topo-probabiliste* (Matheron 1982, Chauvet 1994). L'espace probabilisé est un triplet $(\Omega, \mathcal{A}, P)$ où, en termes intuitifs (Chauvet 1994) :

- Ω est l'inventaire des états possibles du système étudié,
- $\mathcal{A}$ est l'ensemble des événements, relatifs aux états du système, ayant un sens,
- P est une loi de probabilité, c'est-à-dire intuitivement une règle de valuation des événements de $\mathcal{A}$.

Considérer que la variable régionalisée $z(\cdot)$ est une réalisation de la fonction aléatoire $Z(\cdot)$ consiste à identifier $z(x)$ et $Z(x, \omega)$ où $x \in D$ et ω est un événement de $\mathcal{A}$ de probabilité $P(\omega)$. En pratique, nous notons plus simplement $Z(x)$, et une première source de confusion — relativement fréquente dans la littérature — consiste à nommer *variable régionalisée* la fonction aléatoire elle-même (*e.g.*, Hamlett *et al.* 1986, Ord 1988, Pettitt & McBratney 1993, Donald 1994, Stein 1994, van Groenigen & Stein 1998, Hoosbeek *et al.* 1998, Oliver *et al.* 1998). Un abus de langage encore plus prononcé consiste à identifier directement la FA au phénomène régionalisé lui-même (*e.g.*, le pH d'un sol), ce qui n'a évidemment aucun sens (*e.g.*, McBratney & Webster 1983a).

Dès lors que le principe de la modélisation topo-probabiliste est retenu, l'objectif central de la géostatistique est de proposer des modèles de FA qui puissent rendre compte de façon suffisamment fine de la variation spatiale exhibée par la VR, tout en autorisant des opérations mathématiques les moins restrictives possibles, ces deux objectifs étant généralement contradictoires. Le choix d'un type de fonction aléatoire étant fait, la géostatistique propose d'opérer dans l'espace probabilisé, puis d'exprimer les résultats de ces opérations sur le même plan que la VR.

Ce va-et-vient constitue toute la puissance de l'approche topo-probabiliste, mais est source de :

- difficultés de compréhension de la démarche,
- difficultés d'interprétation des résultats en termes objectifs.

Ces difficultés seront en partie aplanies dans les chapitres suivants, et il n'est ici question que de la nature des modèles eux-mêmes et presque pas des problèmes posés par leur utilisation pratique.

4.2.1 Fonctions aléatoires

Soit $Z(\cdot)$ une fonction aléatoire et $s = \{x_i \mid i = 1, \ldots, n\}$ un ensemble de supports dans $\mathbb{R}^2$. A tout ensemble s correspond un vecteur de variables aléatoires comportant n composantes, soit $\mathbf{Z} = [Z(x_1), Z(x_2), \ldots, Z(x_n)]^{\mathrm{T}}$. D'un point de vue probabiliste, le vecteur $\mathbf{Z}$ est caractérisé par sa fonction de répartition conjointe (Journel & Huijbregts 1978, Deutsch & Journel 1992) :

$$F_{x_1,x_2,\ldots,x_n}(z_1, z_2, \ldots, z_n) = \Pr(Z(x_1) \leq z_1, Z(x_2) \leq z_2, \ldots, Z(x_n) \leq z_n) \tag{4.1}$$

L'ensemble de toutes les fonctions de répartition (4.1), pour tout entier n et tout ensemble s dans $\mathbb{R}^2$ constitue la *loi spatiale* de la fonction aléatoire $Z(\cdot)$. Une première simplification du modèle consiste à distinguer deux VA $Z(x_1)$ et $Z(x_2)$ uniquement sur la base de leurs deux premiers moments : espérance et variance/covariance. L'espérance d'une VA en un point x est généralement considérée comme une fonction de x :

$$\mathrm{E}[Z(x)] = m(x) \tag{4.2}$$

avec $m(\cdot)$ la fonction de dérive, ou plus simplement la *dérive* de la fonction aléatoire $Z(\cdot)$. Lorsqu'elle est définie, la variance de $Z(x)$ est également exprimée comme une fonction de x :

$$\mathrm{Var}[Z(x)] = \mathrm{E}\left[\{Z(x) - m(x)\}^2\right] \tag{4.3}$$

et si la variance est définie, la covariance entre deux VA $Z(x_1)$ et $Z(x_2)$ l'est également :

$$\mathrm{Cov}[Z(x_1), Z(x_2)] = \mathrm{E}[\{Z(x_1) - m(x_1)\}\{Z(x_2) - m(x_2)\}] \tag{4.4}$$

4.2.1.1 Stationnarité

Il existe plusieurs acceptions du terme *stationnarité* selon qu'il s'agit de la variable régionalisée ou de la fonction aléatoire qui la modélise. Dans le cas de la VR, nous parlons d'*homogénéité spatiale* pour décrire l'absence de tendance marquée, à l'échelle considérée, et nous réservons le terme de *stationnarité* aux FA.

Il est nécessaire d'invoquer une forme de stationnarité de la FA parce qu'il n'est pas possible d'inférer la loi spatiale ou ses moments à partir d'une seule réalisation d'une FA. Le problème de l'inférence des paramètres d'une FA à partir d'une VR est analogue à celui de l'inférence des paramètres d'une VA à partir d'une seule valeur. Ne disposant que d'une seule VR, pour réaliser l'inférence des paramètres de la FA, il faut trouver

l'information nécessaire parmi les valeurs $\{z(x_i) \mid i = 1, \ldots, n\}$. Faire l'hypothèse de la stationnarité revient à compenser l'absence de plusieurs réalisations de la FA par une forme de redondance de l'information au sein d'une seule réalisation (*i.e.,* la VR). Il convient toutefois de distinguer plusieurs formes de stationnarité d'une FA (Myers 1989).

Stationnarité stricte Définie sans ambiguïté, une FA est une fonction aléatoire stationnaire (FAST) si pour n fini, et pour tout vecteur inter-support $\mathbf{h}$, la fonction de répartition conjointe de $\{Z(x_i) \mid i = 1, \ldots, n\}$ est la même que celle de $\{Z(x_i + \mathbf{h}) \mid i = 1, \ldots, n\}$: on parle d'invariance de la loi spatiale par translation (Journel & Huijbregts 1978). La stationnarité stricte ne contient aucune hypothèse concernant les espérances, variances ou covariances, qui peuvent éventuellement ne pas être définies (Myers 1989, Chauvet 1994). Modéliser une VR par une FAST correspond à une hypothèse irréaliste parce que beaucoup trop forte vis-à-vis de l'homogénéité spatiale de la VR. En outre, comme la géostatistique linéaire — la plus largement utilisée en pratique — ne fait référence qu'aux deux premiers moments de la FA, il suffit d'affaiblir les hypothèses en n'imposant pas la stationnarité des moments au-delà de l'ordre 2 (Journel & Huijbregts 1978).

Stationnarité d'ordre 2 Une fonction aléatoire est dite *stationnaire à l'ordre 2* (FAST-2) si la covariance existe et ne dépend que du vecteur inter-support $\mathbf{h}$, ce qui implique que l'espérance et la variance existent et ne dépendent pas de x (Journel & Huijbregts 1978, Myers 1989) soit :

$$\begin{aligned} \mathrm{E}[Z(x)] &= m & (4.5)\\ \mathrm{Var}[Z(x)] &= \mathrm{E}\left[\{Z(x) - m\}^2\right] = \mathrm{C}(\mathbf{0}) & (4.6)\\ \mathrm{Cov}[Z(x), Z(x+\mathbf{h})] &= \mathrm{E}[Z(x) \cdot Z(x+\mathbf{h})] - m^2 = \mathrm{C}(\mathbf{h}) & (4.7) \end{aligned}$$

où $\mathbf{0}$ désigne le vecteur nul. Il faut noter qu'une FAST n'est pas nécessairement une FAST-2 et réciproquement (Myers 1989). La stationnarité d'ordre 2 est souvent qualifiée de *stationnarité* ou de *stationnarité au sens large* (*e.g.*, Matérn 1960) : comme cette terminologie peut prêter à confusion, nous ne l'utilisons pas et parlons toujours de *stationnarité à l'ordre 2*.

Hypothèse intrinsèque Dans le cas des VR qui présentent une variation spatiale qui n'apparaît pas bornée, au moins au sein du domaine d'étude D, il n'est pas réaliste d'employer une FAST-2 et il convient d'affaiblir encore davantage l'hypothèse de stationnarité. Une fonction aléatoire est *intrinsèque à l'ordre 0* (FAI-0) si ses accroissements d'ordre 1 sont stationnaires d'ordre 2, autrement dit, si les espérances et les variances des incréments $Z(x+\mathbf{h}) - Z(x)$ existent et ne dépendent pas de x, soit :

$$\begin{aligned} \mathrm{E}[Z(x+\mathbf{h}) - Z(x)] &= m & (4.8)\\ \mathrm{Var}[Z(x+\mathbf{h}) - Z(x)] &= \mathrm{E}\left[\{Z(x+\mathbf{h}) - Z(x) - m\}^2\right] = 2\gamma(\mathbf{h}) & (4.9) \end{aligned}$$

avec $\gamma(\mathbf{h})$ une fonction nommée *demi-variogramme*, ou selon l'usage de plus en plus répandu, *variogramme* (Isaaks & Srivastava 1989, p. 65, Chauvet 1994, p. 54). Afin de simplifier les développements mathématiques, on considère généralement que la dérive

constante est nulle, soit $m = 0$ (Matheron 1965, p. 123, 130, Chauvet 1994, p. 54), ce qui donne la définition du variogramme la plus fréquemment rencontrée dans la littérature :

$$\gamma(\mathbf{h}) = \frac{1}{2} \mathrm{E}\left[\{Z(x+\mathbf{h}) - Z(x)\}^2\right] \tag{4.10}$$

Il convient de noter que dans le modèle intrinsèque, il ne figure plus aucune hypothèse d'existence concernant l'espérance et *a fortiori* la variance des VA elles-mêmes, de sorte que tous les développements mathématiques s'effectuent, en toute rigueur, uniquement sur les accroissements. Cependant, la classe des FAI-0 inclut celle des FAST-2 de sorte que toute FAST-2 est aussi FAI-0. En traitant des FAST-2 sous la forme de FAI-0, il est licite de manipuler directement les VA, et la covariance est définie et reliée au variogramme par la relation (Journel et Huijbregts 1978) :

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}) \tag{4.11}$$

En revanche, toutes les FAI-0 ne sont pas nécessairement des FAST-2, et dans ce cas on parle de *fonctions aléatoires intrinsèques strictes*, et il n'y a plus aucun sens à manipuler les espérances et les variances des VA.

L'accroissement d'ordre 1 $Z(x+\mathbf{h}) - Z(x)$ filtre une dérive constante, *i.e.* un polynôme de degré 0. Ce mécanisme peut se généraliser en considérant des incréments d'ordre supérieur à 1, ce qui conduit à la définition des *fonctions aléatoires intrinsèques d'ordre k* (FAI-k). Les FAI-k ont la propriété de filtrer des polynômes de degré k, ce qui permet de modéliser une VR présentant une tendance, pour $k > 0$. Sans entrer dans les détails de la théorie des FAI-k (*cf.* Matheron 1973, Delfiner & Matheron 1980, Cressie 1991), le résultat essentiel à retenir est que l'affaiblissement progressif de l'hypothèse de stationnarité s'accompagne d'une abstraction croissante du modèle topo-probabiliste tout en autorisant la modélisation de classes de phénomènes plus vastes.

Remarque 2 *Les différentes formes de stationnarités affaiblies sont des propriétés mathématiques concernant exclusivement les FA et par conséquent, il est hors de propos de chercher à tester une hypothèse de stationnarité à partir d'une VR (Journel 1985, Myers 1989). La pratique qui consiste à nommer* variable régionalisée *la fonction aléatoire elle-même contribue certainement à entretenir la confusion concernant les hypothèses de stationnarité en géostatistique, de sorte que certains auteurs mentionnent l'hypothèse intrinsèque ou la stationnarité d'ordre 2 des données elles-mêmes (e.g., Phinn et al. 1996, Brannan & Hamlett 1998, Chang et al. 1998), ce qui n'a rigoureusement aucun sens.*

4.2.1.2 Ergodicité

L'ergodicité est une seconde hypothèse introduite dans le cadre de la modélisation probabiliste. Une FA est dite ergodique si ses paramètres peuvent être inférés à partir d'une seule réalisation, autrement dit, si les espérances peuvent être estimées par des moyennes spatiales (Cressie 1988a, Chauvet 1993). L'ergodicité établit le passage entre la loi de probabilité de la FA et sa structure spatiale qui seule sera "observable" à travers ce qui est considéré dans le modèle comme une de ses réalisations possibles, *i.e.* la VR (Chauvet 1994). L'ergodicité considère que les moyennes spatiales tendent vers les espérances dans

le modèle lorsque le domaine D devient infiniment grand, ce qui s'écrit, dans le cas d'une FAST-2 et de la moyenne globale (Chauvet 1994) :

$$\lim_{D\to\infty} \frac{1}{[D]} \int_D Z(x)\, dx = \mathrm{E}[Z(x)] \tag{4.12}$$

avec $[D]$ l'aire de D. En pratique, cela revient à calculer :

$$z_D = \frac{1}{[D]} \int_D z(x)\, dx \tag{4.13}$$

pour D aussi grand que possible, et à considérer z_D comme un estimateur de $\mathrm{E}[Z(x)]$. Pour que l'ergodicité constitue une hypothèse objective, il faut que le domaine soit grand par rapport à la portée de l'autocorrélation spatiale (Matheron 1978, p. 106, Deutsch & Journel 1992, p. 127).

Enfin, il est important de noter que l'ergodicité n'a de raison d'être que dans la mesure où le problème posé est celui de l'inférence des paramètres d'une FA à partir d'une seule réalisation. Dès lors que le point de vue se porte sur la VR elle-même et que la FA est considérée comme un intermédiaire opératoire, il n'est plus nécessaire d'invoquer l'ergodicité du modèle (Journel 1985). Ce point de vue *non ergodique* est illustré par la suite.

4.2.1.3 Isotropie

Une variable régionalisée $z(\cdot)$ et la fonction aléatoire $Z(\cdot)$ qui la modélise sont dites *isotropes* lorsque leur structure de corrélation spatiale ne dépend pas de la direction, *e.g.* dans le cas d'une FAI-0, lorsque le variogramme ne dépend pas du vecteur inter-support $\mathbf{h}$ mais seulement de la distance entre supports h. Dans le cas contraire — lorsque la variabilité spatiale diffère sensiblement selon la direction — la VR et la FA sont dites *anisotropes*. Il existe plusieurs types d'anisotropies (revues dans Journel & Huijbregts 1978, Zimmerman 1993). Pour simplifier, nous considérons toujours que la VR et la FA sont isotropes.

4.2.1.4 Fluctuation

Le variogramme théorique $\gamma(h)$ est défini comme un moment d'ordre 2 d'une fonction aléatoire, *i.e.* comme une intégrale stochastique portant sur une infinité de réalisations de la FA. Il est également possible de définir un *variogramme local* comme une intégrale d'espace pour une réalisation donnée. Le variogramme local sur D s'écrit (Matheron 1965, Journel & Huijbregts 1978) :

$$\gamma_D(h) = \frac{1}{2[D \cap \tau_h D]} \int_{D\cap\tau_h D} \{z(x+h) - z(x)\}^2\, dx \tag{4.14}$$

avec τ_h l'opérateur de translation par les vecteurs de module h. Dans le cadre du modèle topo-probabiliste, $\gamma_D(h)$ est une réalisation d'une VA dont l'espérance n'est autre que le variogramme théorique :

$$\gamma(h) = \mathrm{E}[\gamma_D(h)] \tag{4.15}$$

et $\gamma_D(h)$ fluctue d'une réalisation à l'autre autour de $\gamma(h)$ avec une *variance de fluctuation* $\mathrm{Var}[\gamma_D(h) - \gamma(h)]$, la différence $\gamma_D(h) - \gamma(h)$ étant désignée sous le nom d'*erreur de fluctuation*. Matheron (1965) montre que la variance de fluctuation est faible pour les faibles distances mais qu'elle croît rapidement lorsque h augmente.

De la même façon que le variogramme local fluctue autour du variogramme théorique, d'une réalisation à l'autre, il est possible de considérer la fluctuation de n'importe quelle caractéristique définie au niveau de chaque réalisation. La *fluctuation* désigne donc plus généralement la variabilité induite par le modèle probabiliste.

4.2.2 Fonctions structurales

A chaque classe de fonction aléatoire correspond une *fonction structurale* qui caractérise la structure de corrélation entre les variables aléatoires qui composent la FA[1] :

- covariance pour les FAST-2,
- variogramme pour les FAI-0,
- covariance généralisée pour les FAI-k.

Pour la cohérence interne du modèle probabiliste, les fonctions structurales doivent tout d'abord satisfaire certaines contraintes mathématiques — d'où découlent une partie de leurs propriétés — dont la conséquence immédiate est de restreindre le nombre des modèles analytiques utilisables en pratique.

4.2.2.1 Contraintes mathématiques

Soit $Z(\cdot)$ une FAST-2 d'espérance m et de covariance $C(h)$. Soit Y une combinaison linéaire finie du type :

$$Y = \sum_{i=1}^{n} \lambda_i Z(x_i) \tag{4.16}$$

avec λ_i un pondérateur quelconque. La combinaison linéaire Y est une VA et sa variance ne doit jamais être négative, ce qui impose (Journel & Huijbregts 1978, Armstrong & Jabin 1981) :

$$\mathrm{Var}[Y] = \sum_i \sum_j \lambda_i \lambda_j C(x_i, x_j) \geq 0 \tag{4.17}$$

avec la notation $C(x_i, x_j) \equiv C(h_{ij})$ où h_{ij} est la distance $d(x_i, x_j)$. Une fonction de covariance $C(h)$ qui garantit que la contrainte (4.17) est respectée, est dite *semi-définie positive*[2]. Soit $\mathbf{C}$ la matrice de covariance dont les entrées sont $c_{ij} = \mathrm{Cov}(Z(x_i), Z(x_j)) = C(x_i, x_j)$. Pour que $\mathbf{C}$ soit une matrice semi-définie positive (Annexe D), il suffit que $C(h)$ respecte la contrainte (4.17).

[1] Dans la littérature écologique, il y a parfois confusion entre les FA proprement dites et leurs fonctions structurales respectives (*e.g.*, Fortin *et al.* 1989).

[2] D'une façon générale, les fonctions semi-définies positives s'avèrent importantes dans les procédures d'interpolation spatiale (*cf.* Myers 1988).

Le même type de contrainte doit être imposé au variogramme dans le cas où $Z(\cdot)$ est une FAI-0, ce qui s'écrit (Journel & Huijbregts 1978, Armstrong & Jabin 1981) :

$$\mathrm{Var}[Y] = -\sum_i \sum_j \lambda_i \lambda_j \gamma(x_i, x_j) \geq 0 \tag{4.18}$$

avec la contrainte supplémentaire $\sum_i \lambda_i = 0$. Un variogramme $\gamma(h)$ qui garantit que la contrainte (4.18) est respectée est dit *conditionnellement semi-défini négatif*. Dans le cas des FAI-k ($k > 0$), les covariances généralisées qui respectent la contrainte du type (4.17) sont dites *conditionnellement semi-définies positives d'ordre k* (Delfiner & Matheron 1980, p. 18).

Dans le cas d'un modèle de FA faisant également référence à un certain type de distribution multivariée (*e.g.*, la distribution lognormale multivariée), il ne suffit pas que la covariance soit semi-définie positive, et d'autres contraintes mathématiques doivent être imposées (Armstrong 1992, Journel 1992). Seul le modèle gaussien multivarié peut être construit de façon mathématiquement consistante avec n'importe quelle covariance semi-définie positive (Armstrong 1992, Journel 1992).

4.2.2.2 Propriétés mathématiques

La covariance $C(h)$ présente les propriétés mathématiques suivantes (Journel & Huijbregts 1978, Chauvet 1994) :

- Symétrie : $C(h) = C(-h)$
- Positivité : $\int C(t,u)\,\lambda(dt)\,\lambda(du) \geq 0$
- Respect de l'inégalité de Cauchy-Schwarz[3] : $|C(h)| \leq C(0)$

La covariance décroît généralement à partir de sa valeur à l'origine $C(0)$ lorsque h croît, avec la limite :

$$\lim_{h\to\infty} C(h) = 0 \tag{4.19}$$

et en pratique, il est possible de poser $C(h) = 0$ lorsque $h \geq a$. La distance a est la *portée* de la covariance et représente la *transition* entre l'état où il existe de l'autocorrélation spatiale ($h < a$) et l'état d'absence d'autocorrélation ($h \geq a$).

La définition du variogramme comme la demi-variance d'incréments conduit aux propriétés suivantes (Journel & Huijbregts 1978, Chauvet 1994) :

- $\gamma(0) = 0$
- $\gamma(h) \geq 0$
- Symétrie : $\gamma(h) = \gamma(-h)$
- La fonction $\gamma(\cdot,\cdot)$ est de type semi-négatif conditionnel ; pour toute mesure λ vérifiant $\int \lambda(dt) = 0$ on a $\int \gamma(t,u)\,\lambda(dt)\,\lambda(du) \leq 0$

[3] L'inégalité de Cauchy-Schwarz affirme que si ϕ est un produit scalaire sur un espace verctoriel E, on a pour tout élément $(x,y) \in E^2$ la relation $[\phi(x,y)]^2 \leq \phi(x,x) \cdot \phi(y,y)$.

- $\gamma(\cdot)$ étant de type semi-négatif conditionnel, on montre que $\gamma(\cdot)$ croît moins vite que h^2 soit :

$$\lim_{h\to\infty} \frac{\gamma(h)}{h^2} = 0 \tag{4.20}$$

Si $\gamma(h)$ est borné, la FA est une FAST-2 et la fonction $C(h)$ existe. Ce type de variogramme caractérise la *transition* entre l'état où il existe de l'autocorrélation spatiale et l'état d'absence d'autocorrélation, avec la limite :

$$\lim_{h\to\infty} \gamma(h) = C(0) \tag{4.21}$$

La continuité et la régularité de la FA $Z(\cdot)$ — et par conséquent la régularité spatiale de la VR $z(\cdot)$ qu'elle modélise — sont reliées au comportement à l'origine du variogramme. Par ordre décroissant de continuité et de régularité, on peut considérer (Journel & Huijbregts 1978) :

1. une allure parabolique : $\gamma(\cdot)$ est deux fois dérivable à l'origine et la FA est dérivable en moyenne quadratique, ce qui est caractéristique d'une forte régularité spatiale,
2. une allure linéaire : $\gamma(\cdot)$ n'est pas dérivable à l'origine mais reste continu en $h = 0$, et la FA est continue en moyenne quadratique[4] mais pas dérivable,
3. une discontinuité à l'origine : $\gamma(h)$ ne tend pas vers 0 lorsque $h \to 0$, bien que par définition $\gamma(0) = 0$. La FA n'est plus continue en moyenne quadratique en 0 mais pour $h > 0$. On parle d'*effet de pépite* pour indiquer que cette discontinuité à l'origine peut traduire l'existence d'une structure dont l'échelle est inférieure à celle qui est considérée,
4. un aspect plat : $\gamma(0) = 0$ et $\gamma(h) = C(0)$ pour $h > \varepsilon$. On parle d'*effet de pépite pur* pour indiquer l'absence d'autocorrélation spatiale.

4.2.2.3 Modèles analytiques

Les calculs faisant intervenir la fonction de covariance $C(h)$ ou le variogramme $\gamma(h)$ nécessitent que ces fonctions soient représentées par des modèles analytiques. Ces modèles de fonctions structurales doivent nécessairement respecter les contraintes (4.17) ou (4.18) : de tels modèles sont dits *autorisés*. Dans le cas des FAST-2, il est équivalent de manipuler la covariance $C(h)$, la corrélation $\rho(h)$ ou le variogramme $\gamma(h)$ puisqu'il est possible de passer d'une fonction à l'autre selon :

$$C(h) = C(0) - \gamma(h) \tag{4.22}$$

$$\rho(h) = \frac{C(h)}{C(0)} = 1 - \frac{\gamma(h)}{C(0)} \tag{4.23}$$

et il suffit donc de considérer les modèles d'une seule fonction, *e.g.* le variogramme. Parmi les modèles de variogrammes autorisés[5] dans $\mathbb{R}^2$ (revues dans Journel & Huijbregts 1978,

[4] Une FA $Z(\cdot)$ est dite continue en moyenne quadratique si $\lim \mathsf{E}\left[\{Z(x+h) - Z(x)\}^2\right] = 0$ lorsque $h \to 0$, $\forall x$ (Journel & Huijbregts 1978, p. 38).

[5] A noter qu'un modèle linéaire à seuil est autorisé dans $\mathbb{R}$ uniquement, et qu'il peut donc être utilisé dans le cas de transects (*e.g.*, Debouzie *et al.* 1996), mais qu'il ne devrait pas être employé dans le cas du plan comme c'est parfois le cas (*e.g.*, Donald 1994).

McBratney & Webster 1986, Cressie 1991, Jian *et al.* 1996), nous avons retenu les modèles périodique, gaussien, cubique, pentasphérique, sphérique, et exponentiel (Fig. 4.1, Annexe F). Tous ces modèles $\gamma(h;\theta)$ sont des fonctions d'un vecteur de trois paramètres $\theta = (c_0, c, a)^{\mathrm{T}}$, où c_0 est la pépite, $c_0 + c$ le seuil, et a la portée du variogramme. Lorsque nous faisons référence à un modèle particulier, nous l'exprimons sous la forme d'un opérateur à trois paramètres, toujours donnés dans le même ordre, par exemple :

- pour le modèle exponentiel : Expo (c_0, c, a),
- pour le modèle périodique : Perio (c_0, c, a), etc.

Au contraire des modèles pentasphérique, sphérique, et exponentiel, les modèles périodique, gaussien et cubique présentent un comportement parabolique à l'origine (tangente horizontale), ce qui traduit une forte régularité spatiale de la VR (Chapitre 11).

Le nombre de modèles simples classiquement utilisés est limité à environ une dizaine de fonctions, mais il est possible de construire des modèles complexes puisque toute combinaison linéaire du type :

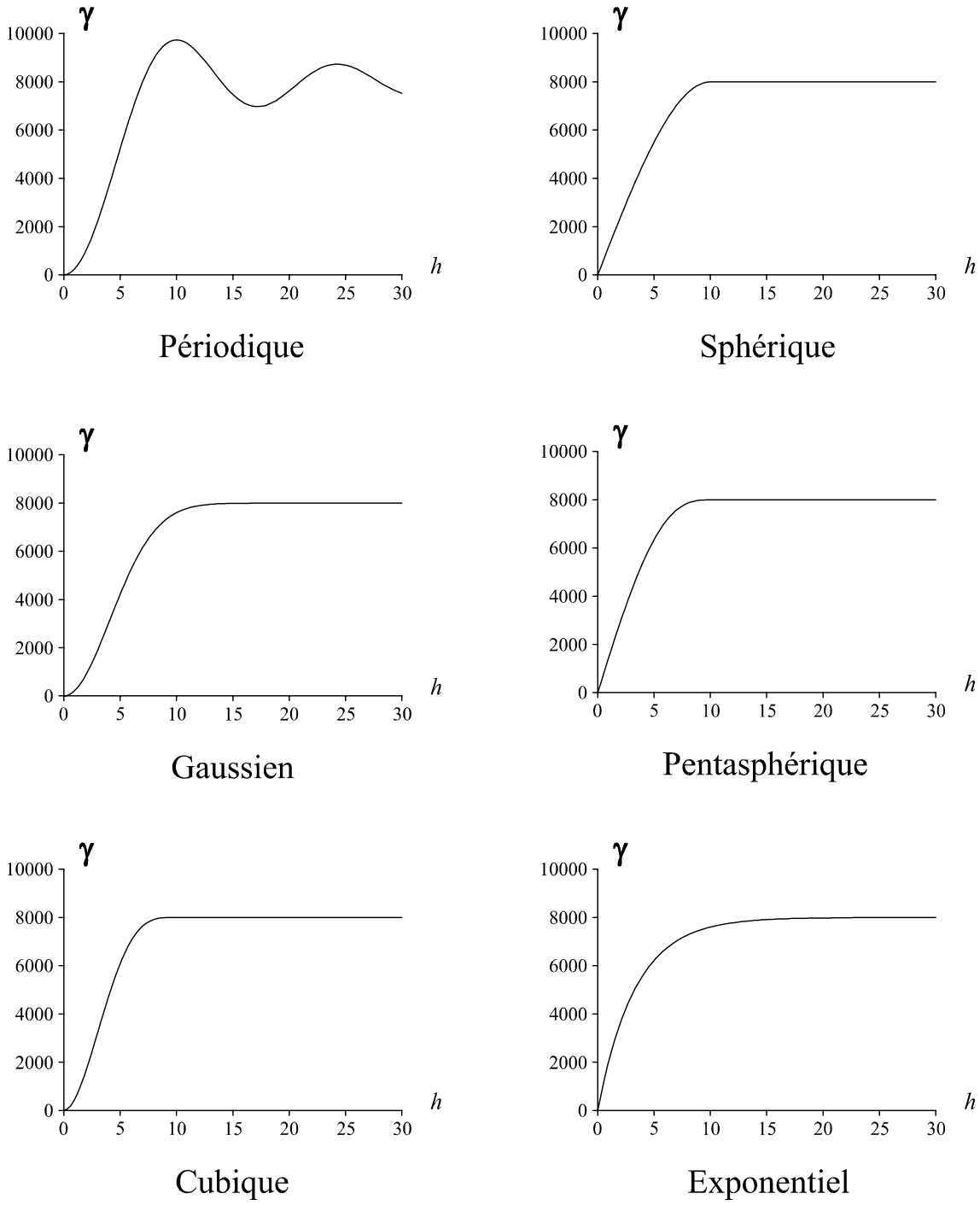
$$\gamma(h) = \sum_{j=1}^{m} \lambda_j \gamma_j(h) \tag{4.24}$$

constitue un modèle autorisé, pourvu que les $\gamma_j(h)$ soient eux-mêmes des modèles autorisés et que $\lambda_j \geq 0 \; \forall j$ (Journel & Huijbregts 1978, p. 162). En pratique, ces modèles dits *emboîtés* comportent rarement plus de $m = 3$ modèles simples.

Dans Aubry (1996a, p. 17, 25) nous avions suggéré de construire des modèles par parties en utilisant les modèles classiques, mais de tels modèles composites s'avèrent généralement inutilisables dans les calculs parce qu'ils ne sont pas toujours conditionnellement semi-définis négatifs. Dans le cas des modèles linéaires par parties, Armstrong & Jabin (1981) notent que certains sont valides et d'autres pas, et qu'il est difficile de tester si un modèle particulier est autorisé ou pas. Armstrong & Diamond (1984) suggèrent de vérifier la validité d'un modèle grâce à une procédure fondée sur le théorème de Bochner, mais cette méthode semble cependant très difficile à utiliser en pratique. Shapiro & Botha (1991, *op. cit.* Barry & Ver Hoef 1996) introduisent une famille de modèles de variogrammes autorisés fondée sur une série de cosinus ou une combinaison linéaire de fonctions de Bessel. Barry & Ver Hoef (1996) proposent de modéliser les variogrammes à seuil par des modèles linéaires par parties grâce à une méthode de moyenne mobile. Nous n'avons pas évalué ces propositions et nous nous contentons des six modèles autorisés mentionnés plus haut.

4.3 Simulation du modèle probabiliste

La simulation du modèle topo-probabiliste consiste à générer numériquement un ensemble de L réalisations $\left\{z^{(\ell)}(\cdot) \mid \ell = 1, \ldots, L\right\}$ de la fonction aléatoire $Z(\cdot)$ sur un ensemble fini de supports $\{x_i \mid i = 1, \ldots, N\}$. Le modèle mathématique laisse la place à un *modèle numérique*. Le passage de l'univers mathématique à l'univers numérique n'est pas sans conséquences sur la fiabilité et la généralité des résultats et en toute rigueur, il n'est plus possible d'effectuer des démonstrations au sein de l'univers numérique. En revanche, de



Figure˜4.1: Six modèles de variogrammes, de paramétrage identique : $c_0 = 1$, $c = 7999$ et $a = 10$.

très nombreux résultats, très difficiles ou même impossibles à obtenir dans l'univers mathématique, s'obtiennent presque sans effort dans le monde numérique, pour peu que l'on dispose d'une puissance de calcul importante. Cet avantage compense largement la perte de généralité des résultats et explique l'effort croissant consacré à la simulation, domaine encore en pleine expansion actuellement (Dowd 1992, Deutsch & Journel 1992, Goovaerts 1999). Afin de limiter notre exposé, nous considérons ici exclusivement la simulation de FAST-2 gaussiennes[6].

Il convient tout d'abord de distinguer deux types de simulations des FA selon que les réalisations sont conditionnées ou pas par des valeurs : on parle de *simulation non conditionnelle*, ou au contraire de *simulation conditionnelle*. Hormis la simulation séquentielle que nous ne considérons pas ici (*cf.* Gómez-Hernández & Srivastava 1990, Dowd 1992, Deutsch & Journel 1992, de Cesare & Posa 1995, Chu 1996, Soares 1998), la plupart des méthodes de simulation conditionnelle des FA sont des extensions des méthodes de simulation non conditionnelle, ce qui nous autorise à examiner d'abord les méthodes de simulation des FA (sous entendue *non conditionnelle*) proprement dites, puis les méthodes de conditionnement par les données.

Il existe de nombreuses méthodes de simulation des FA qui peuvent être classées en plusieurs catégories, selon différents critères tels que la précision des résultats, la généralité ou la flexibilité, la complexité algorithmique (en temps de calcul ou en espace mémoire), ou encore la facilité d'implémentation et d'utilisation. Cependant, plusieurs méthodes peuvent être combinées entre elles de sorte qu'il n'est pas toujours facile de savoir dans quelle catégorie il convient de classer la méthode hybride résultante. Néanmoins, en fonction du type d'approche, il est possible de distinguer les méthodes :

- spectrales,
- des bandes tournantes,
- de factorisation de la matrice de covariance.

L'objet de cette section est de présenter de façon relativement détaillée l'approche que nous avons retenue dans le cadre de l'écologie statistique, et de justifier ce choix, ce qui nécessite au préalable de mentionner les principales approches alternatives — sans toutefois entrer dans les détails très techniques de ces méthodes — et de comparer leurs avantages et défauts respectifs.

4.3.1 Méthode spectrale

La plus ancienne méthode de simulation d'une FA est l'approche spectrale. Les réalisations basées sur la méthode spectrale sont généralement obtenues en sommant une série finie de cosinus dont les coefficients ont des phases et des amplitudes uniformément distribuées et proportionnelles à la fonction de densité spectrale associée à la corrélation de la FA (Dietrich & Newsam 1993). Mantoglou & Wilson (1982) recommandent d'utiliser la méthode spectrale de Shinozuka & Jan (1972). La méthode spectrale exposée dans Borgman *et al.* (1984) est également souvent citée (*e.g.*, Davis 1987a, Dietrich 1995, Dietrich & Newsam 1997).

[6] Pour la simulation des FAI-k, voir notamment Matheron (1973) et Dimitrakopoulos (1990). Pour la simulation des FA non gaussiennes, voir Lee & Ellis (1997) et Bourgault (1997).

4.3.2 Méthode des bandes tournantes

Un peu plus récente que l'approche spectrale, la méthode des bandes tournantes (Matheron 1973, Journel & Huijbregts 1978), consiste à réduire le problème de la simulation bidimensionnelle (2D) sur $D \subset \mathbb{R}^2$ à celui de la simulation unidimensionnelle (1D). Des simulations 1D indépendantes sont réalisées sur M lignes $\{\ell_i \mid i = 1, \ldots, M\}$ dont les vecteurs directeurs unitaires sont uniformément répartis sur le cercle unité centré dans D. Autrement dit, l'angle θ_i formé par la ligne ℓ_i et l'axe (O, x) est distribué uniformément entre 0 et 2π. Chaque ligne est discrétisée en un certain nombre de points qui sont vus comme les centres de segments de droite, tous les points d'un segment ayant la valeur du point central. En limitant chaque extrémité d'un segment sur ℓ_i par une droite perpendiculaire, on définit une *bande* perpendiculaire à la ligne dont la largeur est celle du segment. Comme les lignes tournent autour du centroïde de D, les bandes perpendiculaires aux lignes tournent également, d'où la terminologie des *bandes tournantes*. La valeur de la réalisation 2D $z(x_i)$ au point $x_i \in D$ est obtenue comme une combinaison linéaire, étendue aux M lignes, des valeurs des bandes qui passent par x_i.

La principale difficulté théorique est d'obtenir la covariance 1D $C^1(h)$ qui permet de simuler des réalisations 2D ayant la covariance désirée $C^2(h)$ (Gneiting 1998). Initialement, seuls quelques modèles de covariance 2D pouvaient être simulés (*e.g.*, sphérique, exponentiel), généralement en utilisant une méthode de moyenne mobile pour les simulations 1D (Journel & Huijbregts 1978, p. 505). Brooker (1985) montre notamment comment obtenir la covariance $C^1(h)$ pour le modèle sphérique en 2D. Mantoglou & Wilson (1982) ont proposé que les réalisations 1D soient effectuées dans le domaine des fréquences, ce qui permet de simuler tout modèle de covariance $C^2(h)$ en exprimant la fonction de densité spectrale de la corrélation 1D à partir de la fonction de densité spectrale radiale de la corrélation 2D. Dietrich (1995) recommande plutôt une implémentation de la méthode des bandes tournantes dans le domaine spatial, ce qui nécessite toutefois de trouver les covariances $C^1(h)$ correspondant à un nombre suffisant de modèles de covariances $C^2(h)$. Gneiting (1998) fournit les fonctions de corrélation 1D pour différentes classes de fonctions de corrélation 2D et différents paramétrages, mais pas pour les modèles usuels utilisés dans ce mémoire (Annexe F).

4.3.3 Simulation par factorisation de la matrice de covariance

Dans ce qui suit, nous détaillons les méthodes fondées sur la décomposition (ou *factorisation*) de la matrice $\mathbf{C}$ des variances/covariances des VA, et nous examinons successivement :

- la structure de $\mathbf{C}$,
- la simulation utilisant la décomposition de Cholesky de $\mathbf{C}$,
- la simulation utilisant la racine carrée de $\mathbf{C}$.

4.3.3.1 Structure de la matrice de covariance

Soit $\mathbf{C}$ la matrice de covariance $N \times N$ associée aux VA $\{Z(x_i) \mid i = 1, \ldots, N\}$ d'une FAST-2. Si les supports des VA sont situés aux noeuds d'une grille, en particulier une

grille carrée $n \times n = N$, la matrice de covariance $\mathbf{C}$ de taille $N \times N$, symétrique et semi-définie positive, possède en plus une structure particulière dite *Toeplitz par bloc* (*block Toeplitz structure*) :

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_1 & \mathbf{C}_2^{\mathrm{T}} & \cdots & \mathbf{C}_n^{\mathrm{T}} \\ \mathbf{C}_2 & \mathbf{C}_1 & & \mathbf{C}_{n-1}^{\mathrm{T}} \\ \vdots & & \ddots & \vdots \\ \mathbf{C}_n & \mathbf{C}_{n-1} & \cdots & \mathbf{C}_1 \end{bmatrix} \tag{4.25}$$

où chaque bloc $\mathbf{C}_i$ de taille $n \times n$, symétrique et semi-défini positif, est lui-même de structure Toeplitz. Ainsi, la matrice $\mathbf{C}$ est entièrement déterminée par sa première ligne (ou première colonne) de blocs, chaque bloc étant lui même entièrement déterminé par sa première ligne (ou première colonne) de valeurs. Cette structure provient :

- du fait que l'hypothèse de stationnarité d'ordre 2 (étendue aux incréments dans le cas des FAI-k) implique que la covariance ne dépend que de la distance entre les supports,
- de la régularité de la répartition spatiale des supports au sein d'une grille.

Un exemple de petite taille permet d'illustrer la structure Toeplitz par bloc. Considérons une grille 3×3 de maille carrée unité : la matrice 9×9 des distances entre les noeuds de la grille s'écrit :

$$\begin{bmatrix} 0 & 1 & 2 & 1 & \sqrt{2} & \sqrt{5} & 2 & \sqrt{5} & \sqrt{8} \\ 1 & 0 & 1 & \sqrt{2} & 1 & \sqrt{2} & \sqrt{5} & 2 & \sqrt{5} \\ 2 & 1 & 0 & \sqrt{5} & \sqrt{2} & 1 & \sqrt{8} & \sqrt{5} & 2 \\ 1 & \sqrt{2} & \sqrt{5} & 0 & 1 & 2 & 1 & \sqrt{2} & \sqrt{5} \\ \sqrt{2} & 1 & \sqrt{2} & 1 & 0 & 1 & \sqrt{2} & 1 & \sqrt{2} \\ \sqrt{5} & \sqrt{2} & 1 & 2 & 1 & 0 & \sqrt{5} & \sqrt{2} & 1 \\ 2 & \sqrt{5} & \sqrt{8} & 1 & \sqrt{2} & \sqrt{5} & 0 & 1 & 2 \\ \sqrt{5} & 2 & \sqrt{5} & \sqrt{2} & 1 & \sqrt{2} & 1 & 0 & 1 \\ \sqrt{8} & \sqrt{5} & 2 & \sqrt{5} & \sqrt{2} & 1 & 2 & 1 & 0 \end{bmatrix} \tag{4.26}$$

et peut se condenser en un vecteur $\left(0, 1, 2, 1, \sqrt{2}, \sqrt{5}, 2, \sqrt{5}, \sqrt{8}\right)^{\mathrm{T}}$ et un algorithme de reconstruction de toute la matrice. La structure Toeplitz par bloc de $\mathbf{C}$ s'avère donc particulièrement intéressante en pratique. En effet :

- elle permet d'économiser de la place mémoire puisqu'il suffit de stocker N valeurs au lieu de N^2 dans le cas général,
- elle présente de nombreuses symétries qui peuvent être exploitées afin d'accélérer les algorithmes d'analyse numérique (*e.g.*, Zimmerman 1989a, 1989b, Dietrich 1993a).

4.3.3.2 Décomposition de Cholesky

Soit $\mathbf{z} = (z_1, \ldots, z_N)^{\mathrm{T}}$ les valeurs d'une réalisation $z^{(\ell)}(\cdot)$ localisées sur les supports $\Omega = \{x_1, \ldots, x_N\}$. Le vecteur $\mathbf{z}$ est une réalisation du vecteur aléatoire $\mathbf{Z} = (Z_1, \ldots, Z_N)^{\mathrm{T}}$ dont la matrice de covariance est $\mathbf{C}$. La matrice $\mathbf{C}$ étant symétrique et semi-définie positive, elle peut être décomposée en un produit $\mathbf{C} = \mathbf{AB}$, notamment en utilisant la décomposition de Cholesky $\mathbf{C} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$. Si $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_N)^{\mathrm{T}}$ est un vecteur de valeurs indépendantes telles

que $\omega_i \sim \mathcal{N}(0,1)$, alors $\mathbf{z} = \mathbf{L}\boldsymbol{\omega}$ est tel que $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, autrement dit, la réalisation simulée $z^{(\ell)}(\cdot)$ respecte le modèle de covariance donné par $\mathbf{C}$ (Ripley 1981, Davis 1987a, Alabert 1987, Oliver 1995). L'algorithme de Cholesky est donné notamment dans Golub & van Loan (1983, p. 89) et sa complexité est généralement $O(N^3)$ en temps de calcul. Si la simulation s'effectue aux noeuds d'une grille $n \times n$, la structure Toeplitz par bloc de $\mathbf{C}$ peut être exploitée de façon à réduire la complexité de l'algorithme en $O(N^{2.5})$ voire même $O(N^2)$ (Dietrich 1993a).

4.3.3.3 Racine carrée

En dehors de la décomposition $\mathbf{C} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$ (Ripley 1981, Davis 1987a) il est également possible d'utiliser la décomposition $\mathbf{C} = \mathbf{B}^2$ (Davis 1987b). En effet, pour une matrice symétrique semi-définie positive $\mathbf{C}$, il existe une matrice orthogonale[7] $\mathbf{Q}$ et une matrice diagonale $\boldsymbol{\Lambda}$ telles que :

$$\mathbf{C} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{\mathrm{T}} \tag{4.27}$$

où les colonnes de $\mathbf{Q}$ sont les vecteurs propres de $\mathbf{C}$ et $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ avec λ_i la valeur propre associée au vecteur propre $\mathbf{e}_i$ (Davis 1987b). Comme $\mathbf{C}$ est semi-définie positive, $\lambda_i \geq 0$ pour $i = 1, 2, \ldots, N$. En posant :

$$\boldsymbol{\Lambda}^{1/2} = \mathrm{diag}\left(\lambda_1^{1/2}, \lambda_2^{1/2}, \ldots, \lambda_N^{1/2}\right) \tag{4.28}$$

la matrice

$$\mathbf{B} = \mathbf{Q}\boldsymbol{\Lambda}^{1/2}\mathbf{Q}^{\mathrm{T}} \tag{4.29}$$

est telle que

$$\mathbf{B}^2 = \mathbf{Q}\boldsymbol{\Lambda}^{1/2}\mathbf{Q}^{\mathrm{T}}\mathbf{Q}\boldsymbol{\Lambda}^{1/2}\mathbf{Q}^{\mathrm{T}} = \mathbf{Q}\boldsymbol{\Lambda}^{1/2}\mathbf{I}\boldsymbol{\Lambda}^{1/2}\mathbf{Q}^{\mathrm{T}} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{\mathrm{T}} = \mathbf{C} \tag{4.30}$$

ce qui établit l'existence de $\mathbf{B}$ si $\mathbf{C}$ est symétrique et semi-définie positive (Davis 1987b). En conséquence, la simulation peut s'effectuer comme $\mathbf{z} = \mathbf{C}^{1/2}\boldsymbol{\omega}$.

La simulation des réalisations d'une loi normale multivariée $\mathcal{N}(\mathbf{0}, \mathbf{C})$ exploitant directement la diagonalisation de $\mathbf{C}$ et la définition (4.29) (Hurst & Knop 1972), n'est envisageable en pratique que si N est relativement petit.

Späth (1967) donne un algorithme de complexité $O(N^3)$pour calculer la racine carrée $\mathbf{B} = \mathbf{C}^{1/2}$, mais sa vitesse de convergence le rend inutilisable en pratique compte tenu des dimensions de $\mathbf{C}$ (généralement $N > 500$).

Le calcul de $\mathbf{B}$ peut également s'effectuer en utilisant une fonction de matrice telle que $\mathbf{B} = f(\mathbf{C})$. En effet, si $\mathbf{C} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{\mathrm{T}}$, la fonction réelle continue $f(\mathbf{C})$ est définie par (Golub & van Loan 1983, Davis 1987b) :

$$\mathbf{C} = \mathbf{Q}\mathbf{F}\mathbf{Q}^{\mathrm{T}} \tag{4.31}$$

avec $\mathbf{F} = \mathrm{diag}(f(\lambda_1), f(\lambda_2), \ldots, f(\lambda_N))$. Afin d'éviter de calculer les valeurs propres, il est possible d'utiliser une approximation $g(x)$ définie sur l'intervalle $[\lambda_{\min}, \lambda_{\max}]$, avec $\lambda_{\min}$ et $\lambda_{\max}$ des bornes des valeurs propres, respectivement inférieure et supérieure. En effet, si $g(x) \simeq f(x)$ sur $[\lambda_{\min}, \lambda_{\max}]$ alors $g(\mathbf{C}) \simeq f(\mathbf{C})$ (Golub & van Loan 1983). Le problème consiste à déterminer les bornes $\lambda_{\min}$ et $\lambda_{\max}$ de façon suffisamment précise et efficace, puis à proposer une approximation de la fonction $f : x \mapsto \sqrt{x}$.

[7] Une matrice Q est orthogonale si $\mathsf{Q}\mathsf{Q}^{\mathsf{T}} = \mathsf{Q}^{\mathsf{T}}\mathsf{Q} = \mathsf{I}$.

Bornes des valeurs propres En principe, les bornes $\lambda_{\min}$ et $\lambda_{\max}$ peuvent être déterminées par la seconde méthode d'Ostrovskiĭ (Sobol *et al.* 1973) mais cette approche nous est apparue difficile à maîtriser. Comme $\mathbf{C}$ est semi-définie positive, le choix de $\lambda_{\min} = 0$ est immédiat. Une approximation précise de la plus grande valeur propre peut être obtenue grâce à la méthode de la puissance (*power method*) (Golub & van Loan 1983, p. 209, Dietrich & Newsam 1995). Cette méthode est assez simple et donne de très bons résultats, mais l'expérience nous a montré que la convergence est trop lente de sorte que cette approche se révèle inutilisable en pratique pour des valeurs de N courantes (*e.g.*, $N = 900$). Enfin, Davis (1987b) propose de déterminer rapidement[8] $\lambda_{\max}$ selon :

$$\lambda_{\max} = \max_i \sum_{j=1}^{N} |k_{ij}| \tag{4.32}$$

en vertu du théorème du cercle de Gershgorin (Golub & van Loan 1983, p. 200). Cette méthode est également celle employée par Späth (1967) dans son algorithme calculant $\mathbf{C}^{1/2}$. En pratique, nous avons constaté que l'utilisation de $\lambda_{\max}$ déterminée par (4.32) à la place de la plus grande valeur propre ne modifie pratiquement pas le résultat de la simulation.

Approximation polynomiale L'approximation de la fonction $f : x \mapsto \sqrt{x}$ par un polynôme de degré m :

$$\sqrt{x} \simeq \sum_{i=0}^{m} a_i x^i \tag{4.33}$$

avec $\{a_i \mid i = 0, \ldots, m\}$ les coefficients du polynôme, permet de calculer approximativement la racine carrée de $\mathbf{C}$ comme (Davis 1987b) :

$$\mathbf{C}^{1/2} \simeq \sum_{i=0}^{m} a_i \mathbf{C}^i \tag{4.34}$$

et la réalisation $\mathbf{z}$ est calculée comme :

$$\mathbf{z} = \mathbf{C}^{1/2}\boldsymbol{\omega} \simeq \sum_{i=0}^{m} a_i \mathbf{C}^i \boldsymbol{\omega} \tag{4.35}$$

les puissances de $\mathbf{C}$ étant calculées successivement grâce à l'algorithme :

1. $i \leftarrow 0$, $\mathbf{z} \leftarrow \mathbf{0}$.
2. Si $i = 0$ alors $\mathbf{y}_i \leftarrow \boldsymbol{\omega}$ sinon $\mathbf{y}_i \leftarrow \mathbf{C}\mathbf{y}_{i-1}$.
3. $\mathbf{z} \leftarrow \mathbf{z} + a_i \mathbf{y}_i$.
4. Si $i = m$ alors FIN, sinon aller en 2.

Pour approximer $f : x \mapsto \sqrt{x}$, Davis (1987b) propose d'utiliser une approximation polynômiale minimax[9] plutôt qu'un développement limité de Taylor qui est sujet aux

[8] Le calcul de $\lambda_{\max}$ peut être accéléré si C présente une structure Toeplitz par bloc.

[9] Un polynôme minimax $g(x)$ est un polynôme qui, parmi les polynômes de même degré, est le plus proche de la fonction $f(x)$ à approximer (Press *et al.* 1989).

instabilités numériques et qui requiert en outre davantage de termes pour atteindre la même précision. Cependant, le polynôme minimax est très difficile à calculer alors que l'approximation de Chebyshev est presque identique tout en étant de calcul aisé (Press *et al.* 1989). En conséquence, nous utilisons l'approximation de Chebyshev afin d'approximer la fonction $f(x)$ sur $[0, \lambda_{\max}]$, un choix identique ayant été fait indépendamment par Dietrich & Newsam (1995), pour les mêmes raisons[10]. Après standardisation des valeurs propres par $\lambda_{\max}$, le problème peut se ramener à l'approximation de $f(x)$ sur $[0, 1]$.

Pour Davis (1987b) ce sont les valeurs propres les plus élevées qui sont importantes et l'auteur concentre la qualité de son approximation au voisinage de 1. Dietrich & Newsam (1994) affirment au contraire que la prépondérance des faibles valeurs propres — surtout dans le cas des covariances "lisses" (*e.g.*, covariance gaussienne) sans pépite — nécessite que l'approximation soit plus précise au voisinage de 0. Dans le doute, nous cherchons une bonne approximation sur l'intervalle $[0, 1]$. Afin de calculer les coefficients du polynôme de Chebyshev, nous disposons d'une part d'une définition analytique des coefficients donnée par Dietrich & Newsam (1994), et d'autre part, de la procédure numérique de Press *et al.* (1989). A nombre de coefficients constant, l'expérience montre que les coefficients de Dietrich & Newsam (1994) permettent effectivement une meilleure approximation au voisinage de 0, mais que la procédure de Press *et al.* (1989) conduit à une approximation globalement plus précise sur l'ensemble de l'intervalle $[0, 1]$. La méthode de calcul des coefficients étant fixée, il reste à choisir le degré du polynôme de Chebyshev.

Afin de déterminer la qualité de l'approximation en fonction du degré du polynôme de Chebyshev, nous avons comparé une réalisation de référence — obtenue en diagonalisant la matrice $\mathbf{C}$ et en utilisant la définition (4.29) — à celles obtenues par la méthode de l'approximation polynomiale pour différentes valeurs de m. Sans détailler nos résultats, il apparaît qu'une valeur $m = 5$ constitue un minimum absolu tandis qu'au-delà de $m = 30$, des problèmes numériques apparaissent en relation avec la détermination des coefficients du polynôme de Chebyshev par la procédure de Press *et al.* (1989). Pour leur version de l'approximation de Chebyshev, Dietrich & Newsam (1994) estiment que pour obtenir une erreur inférieure à 1 % il faut au moins $m = 32$. En conséquence, nous considérons que la valeur $m = 30$ constitue un degré maximal, conduisant à une très bonne précision.

Méthodes d'accélération Afin d'accélérer le calcul des produits matrice-vecteur, il est possible de recourir à des convolutions en utilisant la FFT (*Fast Fourier Transform*) (Davis 1987b, Dietrich & Newsan 1995). Par ailleurs, dans le cas d'une matrice de covariance Toeplitz par bloc, Dietrich & Newsam (1993), ainsi que Chan & Wood (1997) proposent une méthode basée sur l'enchâssement de la matrice de covariance $\mathbf{C}$ au sein d'une matrice $\mathbf{S}$ plus large, définie positive, présentant une structure de blocs circulants. Les auteurs utilisent une diagonalisation par FFT afin de calculer $\mathbf{S}^{1/2}$ de façon efficace. Toujours grâce à la FFT, le produit matrice-vecteur $\mathbf{q} = \mathbf{S}^{1/2}\boldsymbol{\omega}$ est effectué rapidement, et la réalisation $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ est obtenue comme un sous-vecteur de $\mathbf{q} \sim \mathcal{N}(\mathbf{0}, \mathbf{S})$ (Dietrich & Newsam 1993). Cette méthode est censée être à la fois exacte et de coût d'exécution comparable à celui de la méthode spectrale, mais elle n'est pas exempte de certaines limitations et son utilisation nous est apparue relativement compliquée, notamment parce qu'il faut assurer que la matrice $\mathbf{S}$ est bien définie positive.

[10] Tal-Ezer (1989, 1991) étudie spécifiquement l'approximation polynomiale des fonctions de matrices, mais nous n'avons pas évalué ses propositions.

Enfin, Oliver (1995) propose de factoriser analytiquement la covariance 2D en l'exprimant sous la forme d'un produit de convolution d'une fonction et de sa transposée. Cette approche évite d'avoir à factoriser une grande matrice de covariance et autorise l'application de la méthode des moyennes mobiles. La méthode des moyennes mobiles était déjà classiquement utilisée pour les simulations 1D (*e.g.*, Journel & Huijbregts 1978) mais on estimait encore récemment qu'elle ne pouvait pas être étendue à la simulation 2D (*e.g.*, Dietrich & Newsam 1993). Cependant, dans l'espace bidimensionnel, il s'avère très difficile de déterminer analytiquement la racine carrée de certains modèles simples (*e.g.*, modèle sphérique) et *a fortiori* des modèles emboîtés. Dans le cas d'un modèle emboîté, il est toutefois possible de contourner cette difficulté en calculant la somme de réalisations indépendantes obtenues avec chacun des modèles simples (Oliver 1995).

4.3.4 Performances des méthodes

Les méthodes de simulation mentionnées peuvent comporter de nombreuses variantes de sorte que toute comparaison rigoureuse devrait faire référence à des implémentations particulières, clairement identitifées, plutôt qu'à des intitulés (*e.g.*, méthode spectrale, méthode des bandes tournantes) qui englobent en fait des procédures dont les performances peuvent s'avérer très différentes. Comme nous n'avons pas implémenté toutes les méthodes, nous nous contentons ici d'une comparaison générale, tant du point de vue de l'exactitude, de la généralité ou de l'efficacité, en exploitant la littérature. Il faut toutefois noter que les avis des mêmes auteurs (*e.g.*, Dietrich & Newsam) peuvent être contradictoires d'un article à l'autre selon la méthode dont ils font la promotion.

Nous abordons séparément la question de la précision, de la généralité et de l'efficacité même si, en pratique, le choix d'une méthode résulte évidemment d'un compromis entre ces trois paramètres.

4.3.4.1 Précision

La méthode spectrale et la méthode les bandes tournantes ne sont qu'asymptotiquement exactes et il faut de nombreuses harmoniques dans la méthode spectrale, et suffisamment de lignes dans la méthode des bandes tournantes pour atteindre une précision correcte (Dietrich 1993a). Mantoglou & Wilson (1982) indiquent que la méthode spectrale de Mejia & Rodriguez-Iturbe (1974) nécessite, selon la portée, de 630 à 40000 harmoniques pour obtenir la même précision que celle atteinte par leur méthode de bandes tournantes avec seulement 8 lignes. Mantoglou & Wilson (1982) estiment qu'une précision très satisfaisante est obtenue en utilisant 16 lignes alors que Brooker (1985) en utilise 100.

Dietrich & Newsam (1995) notent que la méthode spectrale peut faire apparaître des artefacts et que son utilisation nécessite par conséquent un fin réglage des paramètres. D'après Journel (1994a), la méthode des bandes tournantes est également sujette aux artefacts.

La simulation par factorisation de la matrice de covariance (*e.g.*, décomposition de Cholesky) est une méthode exacte en ce qu'elle reproduit exactement le modèle de covariance imposé. En termes de précision, la méthode utilisant la décomposition de Cholesky peut donc être vue comme une référence (*benchmark*) dont la limite ne provient que du

générateur de nombres pseudo-aléatoires (Annexe B) et de l'arithmérique de l'ordinateur utilisé (Brooker & Stewart 1994).

Sur la base de 50 réalisations, et en termes de fidélité de la reproduction du variogramme théorique, Brooker & Stewart (1994) comparent la précision de la méthode des bandes tournantes (avec 15 ou 50 bandes), de la méthode de la décomposition de Cholesky ($\mathbf{z} = \mathbf{L}\boldsymbol{\omega}$), et de la méthode de la racine carrée ($\mathbf{z} = \mathbf{C}^{1/2}\boldsymbol{\omega}$) avec une approximation polynomiale de 8 et 15 coefficients. La méthode des bandes tournantes s'avère médiocre lorsque la portée est petite vis-à-vis du domaine. En outre, elle a tendance à introduire une anisotropie artefactuelle. Ces défauts peuvent être réduits en augmentant le nombre de bandes de 15 à 50. La méthode de l'approximation polynomiale de la racine carrée donne de bons résultats pour les portées assez petites mais les performances se dégradent lorsque la portée augmente, ces défauts étant cependant moins prononcés avec un polynôme de degré $m = 15$.

4.3.4.2 Généralité

La méthode spectrale nécessite que les supports des réalisations à simuler soient répartis selon une grille, ce qui n'est pas le cas pour la méthode des bandes tournantes. La méthode spectrale et la méthode les bandes tournantes sont souvent considérées comme essentiellement restreintes aux FAST-2 isotropes (Dietrich 1993a, 1996). Cependant, Matheron (1973) et Dimitrakopoulos (1990) considèrent la simulation des FAI-k par la méthode des bandes tournantes, et Mantoglou & Wilson (1982) estiment que leur méthode des bandes tournantes peut aisément être étendue aux fonctions anisotropes. Par ailleurs, Dietrich (1995) mentionne à son tour l'extension de la méthode des bandes tournantes aux cas anisotropes (Mantoglou 1987), et aux FAI-k (Dimitrakopoulos 1990). Dietrich (1995) et Dietrich & Newsam (1997) utilisent la méthode de l'enchâssement de la matrice de covariance (Dietrich & Newsam 1993) afin de simuler les réalisations 1D utilisées dans la méthode des bandes tournantes, ce qui évite notamment d'avoir à connaître la fonction de densité spectrale de la covariance 1D.

Enfin, la méthode basée sur la factorisation de la matrice de covariance (*e.g.*, décomposition de Cholesky) est tout à fait générale et permet de simuler tout modèle de covariance, simple ou emboîté, isotrope ou anisotrope, sur des supports qui ne sont pas nécessairement disposés selon une grille. La flexibilité de cette approche constitue donc un de ses principaux avantages (Alabert 1987, Dietrich 1993b)

4.3.4.3 Efficacité

La méthode spectrale et la méthode des bandes tournantes sont des méthodes rapides (Dietrich & Newsam 1996), la méthode des bandes tournantes de Mantoglou & Wilson (1982) ayant cependant l'avantage de converger plus rapidement que la méthode spectrale. La méthode de l'enchâssement de la matrice de covariance (Dietrich & Newsam 1993) est censée être d'une efficacité comparable à celle de la méthode spectrale.

Les méthodes utilisant la factorisation exacte de la matrice de covariance figurent parmi les moins rapides, et leur temps d'exécution devient prohibitif dans le cas des grandes simulations. L'utilisation d'une factorisation approximative permet toutefois de traiter des problèmes de plus grande taille, surtout en faisant appel à la FFT.

4.3.5 Conditionnement des simulations par les données

Soit la variable régionalisée étudiée $z^{(0)}(\cdot)$ définie dans un domaine $D \subset \mathbb{R}^2$. Cette variable régionalisée est connue uniquement sur les supports $\Omega_1 = \{x_i \mid i = 1, \ldots, n_1\}$ et les valeurs $\left\{z^{(0)}(x_i) \mid i = 1, \ldots, n_1\right\}$ sont désignées comme les *données*. Soit un ensemble de supports $\Omega_2 = \{x_i \mid i = 1, \ldots, n_2\}$ inclus dans D pour lesquels on cherche à simuler L réalisations $\left\{z_c^{(\ell)}(\cdot) \mid \ell = 1, \ldots, L\right\}$ de la fonction aléatoire $Z(\cdot)$, conditionnées par les données. Il est possible de distinguer deux types de conditionnement par les données, selon qu'il est réalisé *a posteriori* ou *a priori*.

4.3.5.1 Conditionnement a posteriori

Historiquement (Journel & Huijbregts 1978), les premières simulations conditionnelles ont été obtenues en conditionnant *a posteriori* des simulations non conditionnelles obtenues par la méthode des bandes tournantes, en imposant que les réalisations respectent les données, *i.e.* $z_c^{(\ell)}(x) = z^{(0)}(x)$ pour tout $x \in \Omega_1$. Soit $z^{(0)}(x_0)$ la valeur de la variable régionalisée $z^{(0)}(\cdot)$ en $x_0 \in \Omega_2$ et $z_k^{(0)}(x_0)$ celle obtenue par krigeage (Section 6.2) à partir des données localisées sur Ω_1. Ces deux valeurs diffèrent d'une erreur $z^{(0)}(x_0) - z_k^{(0)}(x_0)$ indépendante de $z_k^{(0)}(x_0)$ (David 1977, Journel & Huijbregts 1978) :

$$z^{(0)}(x_0) = z_k^{(0)}(x_0) + \left[z^{(0)}(x_0) - z_k^{(0)}(x_0)\right] \tag{4.36}$$

Dans ce contexte, le principe de la simulation conditionnelle consiste à remplacer l'erreur $z^{(0)}(x_0) - z_k^{(0)}(x_0)$ propre à la réalisation $z^{(0)}(\cdot)$ par un terme isomorphe, indépendant de $z_k^{(0)}(x_0)$, obtenu à partir d'une réalisation simulée $z^{(\ell)}(\cdot)$, isomorphe mais indépendante de $z^{(0)}(\cdot)$, ce qui donne, pour $x \in \Omega_2$ (Journel & Huijbregts 1978, Davis 1987a) :

$$z_c^{(\ell)}(x) = z_k^{(0)}(x) + \left[z^{(\ell)}(x) - z_k^{(\ell)}(x)\right] \tag{4.37}$$

avec $z_c^{(\ell)}(\cdot)$ une réalisation conditionnelle, $z_k^{(0)}(\cdot)$ une estimation par krigeage à partir des données localisées sur Ω_1, $z^{(\ell)}(\cdot)$ une réalisation non conditionnelle, et enfin $z_k^{(\ell)}(\cdot)$ une estimation par krigeage à partir de valeurs simulées sur Ω_1. Comme le krigeage est un interpolateur exact, pour $x \in \Omega_1$ on a $z^{(\ell)}(x) = z_k^{(\ell)}(x)$ et $z^{(0)}(x) = z_k^{(0)}(x)$ d'où $z_c^{(\ell)}(x) = z^{(0)}(x)$: autrement dit, la réalisation conditionnelle $z_c^{(\ell)}$ respecte strictement les données. L'expression (4.37) peut également être exprimée sous la forme :

$$z_c^{(\ell)}(x) = z^{(\ell)}(x) + \left[z_k^{(0)}(x) - z_k^{(\ell)}(x)\right] \tag{4.38}$$

où l'écart $z_k^{(0)}(x) - z_k^{(\ell)}(x)$ est estimé par krigeage à partir de la différence $z^{(0)}(x) - z^{(\ell)}(x)$ pour $x \in \Omega_1$ (Davis 1987a), ce qui ne nécessite qu'un krigeage au lieu des deux krigeages qui figurent dans l'expression (4.37). La forme (4.38) permet en outre de mettre en évidence la nature du conditionnement puisqu'il apparaît clairement qu'il s'agit d'ajouter à une réalisation non conditionnelle $z^{(\ell)}(\cdot)$ une estimation de l'écart entre $z^{(\ell)}(\cdot)$ et la variable régionalisée sous étude $z^{(0)}(\cdot)$. Il en découle que le conditionnement de la simulation se traduit par une réduction considérable des fluctuations des réalisations.

4.3.5.2 Conditionnement a priori

Il est évident que la méthode (4.38) permet de conditionner *a posteriori* une réalisation non conditionnelle obtenue par n'importe quelle méthode de simulation, y compris dans le cas des réalisations obtenues par la méthode de l'approximation polynomiale de $\mathbf{C}^{1/2}$ (*e.g.*, Davis 1987b). Cependant, dans le cas des méthodes exactes basées sur la factorisation de la matrice de covariance, il est possible de conditionner la simulation *a priori*, ce qui évite notamment d'avoir à utiliser un programme de krigeage. Dans ce contexte, le conditionnement peut être défini dans un cadre statistique multivarié classique (Solow 1994, Dietrich & Newsam 1995).

Considérons le modèle gaussien multivarié $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ avec $\mathbf{y} = [y(x) \mid x \in \Omega]^{\mathrm{T}}$ une réalisation du vecteur aléatoire $\mathbf{Y} = [Y(x) \mid x \in \Omega]^{\mathrm{T}}$ et Ω un ensemble de supports dans D. La partition des supports de Ω en $\Omega_1 \cup \Omega_2$ entraîne une partition du vecteur de VA $\mathbf{Y}$ en deux sous-vecteurs $\mathbf{Y}_1$ et $\mathbf{Y}_2$, associée à celle du vecteur des espérances $\boldsymbol{\mu}$ en $\boldsymbol{\mu}_1$ et $\boldsymbol{\mu}_2$, et à celle de la matrice de covariance conjointe $(n_1 + n_2) \times (n_1 + n_2)$:

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}^{\mathrm{T}} & \mathbf{C}_{22} \end{bmatrix} \tag{4.39}$$

avec $\mathbf{C}_{11}$ la matrice de covariance $n_1 \times n_1$ entre les VA de $\mathbf{Y}_1$, $\mathbf{C}_{22}$ la matrice de covariance $n_2 \times n_2$ entre les VA de $\mathbf{Y}_2$, et $\mathbf{C}_{12}$ la matrice de covariance $n_1 \times n_2$ entre les VA de $\mathbf{Y}_1$ et $\mathbf{Y}_2$. En vertu de l'hypothèse de stationnarité d'ordre 2, les espérances des VA de $\mathbf{Y}_1$ et de $\mathbf{Y}_2$ sont constantes, soit $\mathrm{E}[Y_1(x)] = \mu_1$ et $\mathrm{E}[Y_2(x)] = \mu_2$. Etant donnée une réalisation $\mathbf{y}_1$ de $\mathbf{Y}_1$, la distribution gaussienne multivariée de $\mathbf{Y}_2$ est d'espérance conditionnelle (Easley *et al.* 1991, Solow 1994, Dietrich & Newsam 1995) :

$$\mathrm{E}[\mathbf{Y}_2 \mid \mathbf{y}_1] = \boldsymbol{\mu}_{\mathbf{Y}_2|\mathbf{y}_1} = \mu_2 + \mathbf{C}_{12}^{\mathrm{T}} \mathbf{C}_{11}^{-1} (\mathbf{y}_1 - \mu_1) \tag{4.40}$$

où $(\mathbf{y}_1 - \mu_1)$ est le vecteur des données centrées, et de covariance conditionnelle :

$$\mathbf{C}_{\mathbf{Y}_2|\mathbf{y}_1} = \mathbf{C}_{22} - \mathbf{C}_{12}^{\mathrm{T}} \mathbf{C}_{11}^{-1} \mathbf{C}_{12} \tag{4.41}$$

En conséquence, une réalisation $\mathbf{y}_2$ de $\mathbf{Y}_2$ conditionnée par $\mathbf{y}_1$ peut être calculée en utilisant la décomposition de Cholesky $\mathbf{C}_{\mathbf{Y}_2|\mathbf{y}_1} = \mathbf{L}_c \mathbf{L}_c^{\mathrm{T}}$ selon :

$$\mathbf{y}_2 = \boldsymbol{\mu}_{\mathbf{Y}_2|\mathbf{y}_1} + \mathbf{L}_c \boldsymbol{\omega} \tag{4.42}$$

ou, en utilisant la factorisation $\mathbf{C}_{\mathbf{Y}_2|\mathbf{y}_1} = \mathbf{B}_c^2$ selon :

$$\mathbf{y}_2 = \boldsymbol{\mu}_{\mathbf{Y}_2|\mathbf{y}_1} + \mathbf{B}_c \boldsymbol{\omega} \tag{4.43}$$

avec $\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}_{n_2}, \mathbf{I}_{n_2})$. Dans le cas de la décomposition de Cholesky, il est possible d'exprimer la simulation conditionnelle en termes des décompositions de Cholesky des sous-matrices de $\mathbf{C}$ (Davis 1987a, Alabert 1987, Dietrich 1993b). En faisant référence à la distribution gaussienne multivariée, il est nécessaire de transformer les données $\mathbf{z}_1$ afin d'avoir des valeurs distribuées approximativement selon la loi normale $\mathcal{N}(0, 1)$, puis d'effectuer la transformation réciproque pour chaque réalisation conditionnelle (Alabert 1987). En géostatistique, ces transformations s'effectuent généralement grâce à l'anamorphose de la fonction de répartition (Annexe E).

Easley *et al.* (1991) traitent la simulation conditionnelle dans le domaine des fréquences plutôt que dans le domaine spatial afin d'exploiter les avantages de la FFT en termes de temps d'exécution et d'occupation de la mémoire. Cependant, cette méthode impose que les supports soient disposés selon une grille régulière. La même restriction apparaît lorsque Dietrich & Newsam (1996) étendent la méthode de l'enchâssement de la matrice de covariance à la simulation conditionnelle. En effet, la méthode de simulation non conditionnelle nécessite que la matrice de covariance soit Toeplitz par bloc et que les blocs soient eux-mêmes Toeplitz.

4.3.6 Choix d'une méthode pour l'écologie statistique

Dans le cadre de l'écologie statistique, il est nécessaire de pouvoir simuler de façon précise de très nombreuses réalisations (*e.g.*, 10^4 voire 10^5 réalisations) afin d'obtenir des résultats suffisamment fiables. La méthode doit être générale de sorte que l'on puisse simuler un large panel de modèles de covariances, simples ou emboîtés, isotropes ou anisotropes, sur des supports qui ne sont pas nécessairement disposés selon une grille. En outre, il est préférable que la méthode soit assez facile à utiliser de sorte que les efforts puissent se concentrer davantage sur les objectifs que sur les outils. La simulation basée sur la factorisation de la matrice de covariance satisfait à ces prérequis (Dietrich & Newsam 1994). De plus, cette approche permet de générer directement des réalisations conditionnées, ce qui est préférable au conditionnement *a posteriori* utilisant le krigeage (Dietrich & Newsam 1995). Cette méthode est simple à implémenter et est particulièrement recommandée lorsqu'il est nécessaire de simuler de très nombreuses réalisations sur un ensemble de supports pas trop grand (Deutsch & Journel 1992, Goovaerts 1997). En conséquence, nous considérons que la simulation utilisant la décomposition de Cholesky représente actuellement la meilleure méthode pour l'écologie statistique, au moins pour des simulations portant au maximum sur 900 ou 1000 points.

Lorsque le nombre de supports augmente, le coût de la décompostition de Cholesky $\mathbf{C} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$ et du calcul des réalisations $\mathbf{z} = \mathbf{L}\boldsymbol{\omega}$ devient rapidement prohibitif et il convient de préférer l'approximation polynomiale de la racine carrée $\mathbf{C}^{1/2}$, jusqu'à un nombre de points maximal qui dépend essentiellement de l'implémentation, de l'ordinateur utilisé et de la précision requise. Cette approche alternative présente l'inconvénient d'être approximative, mais les bornes de l'erreur d'approximation sont disponibles et la méthode s'avère relativement simple à implémenter et numériquement stable (Dietrich & Newsam 1995).

Remarque 3 *Il est important de souligner que les factorisations* $\mathbf{C} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$ *et* $\mathbf{C} = \mathbf{B}^2$ *diffèrent en ce que la première correspond à une décomposition asymétrique de la covariance tandis que la seconde correspond à une décomposition symétrique (Oliver 1995). Il est licite d'utiliser l'une ou l'autre mais les fonctions sous-jacentes sont différentes, ce qui conduit à deux définitions opératoires différentes de la simulation exacte*[11]*. Bien qu'en pratique les résultats s'avèrent peu différents, dans un souci de cohérence, il convient de comparer les résultats obtenus à partir de la même méthode.*

[11] Il en est donc de même pour l'opération réciproque, *i.e.* la décorrélation.

Compte tenu du caractère exact et général de la méthode utilisant la décomposition de Cholesky, et de la taille modeste des grilles que nous simulons par la suite (grilles 30×30), nos réalisations sont simulées selon $\mathbf{z} = \mathbf{L}\boldsymbol{\omega}$ avec $\omega_i \in \boldsymbol{\omega}$ tel que $\omega_i \sim \mathcal{N}(0,1)$. Les ω_i sont générés en utilisant une variante de la méthode de Box-Müller (Ripley 1983, 1987, p. 62, Press *et al.* 1989, pp. 224-226).

Il faut toutefois identifier un certain nombre de problèmes numériques tels que le mauvais conditionnement ou l'éventuelle singularité de la matrice de covariance.

4.3.6.1 Conditionnement

Les modèles périodique et gaussien donnent des matrices de covariance mal conditionnées de sorte qu'il faut toujours introduire une petite pépite dans le modèle, *e.g.* décomposer un seuil $c_0 + c = 8000$ en $c_0 = 1$ et $c = 7999$ au lieu de $c_0 = 0$ et $c = 8000$. Si les points de Ω sont proches, et si les VA correspondantes sont très corrélées, il faut s'attendre à ce que la matrice de covariance soit très mal conditionnée. Dans le cas de l'algorithme de Cholesky, le mauvais conditionnement se traduit par une tentative de calcul de la racine carrée d'un nombre négatif (Dietrich 1993b). Ce problème survient généralement lorsque la portée est grande par rapport à la taille de la grille (Dietrich 1993a).

Dans le cas d'une grille de pas $\Delta = 1$ ou $\Delta = 0.5$, grâce à l'utilisation d'une petite pépite et de flottants codés sur 10 octets, nous n'avons jamais rencontré ce problème, même pour des portées relativement élevées par rapport au domaine.

4.3.6.2 Singularité

Soit Ω un ensemble de supports et $\mathbf{C}$ la matrice de covariance associée. Si des supports de Ω sont proches les uns des autres, alors $\mathbf{C}$ peut s'avérer impossible à inverser. A l'extrême, si des supports sont confondus alors $\mathbf{C}$ est singulière et son inverse n'est pas définie. Dans ce cas, il est possible de recourir à l'inverse généralisée de Moore-Penrose (Easley *et al.* 1991).

En particulier, dans le cadre de la simulation conditionnelle rien n'interdit que des points de Ω_2 coïncident avec des points de Ω_1 mais dans ce cas, la matrice de covariance est singulière (Dietrich 1993b). Pour la rendre régulière, et si $c_0 \simeq 0$, le plus simple est de décaler les points incriminés, même très légèrement. Par exemple, si Ω_2 est une grille 30×30 de pas $\Delta = 0.5$ d'origine $\left(\frac{1}{2}\Delta, \frac{1}{2}\Delta\right)$, autrement dit, centrée dans un domaine D carré de 15 unités de côté et d'origine $(0,0)$, et si Ω_1 est une grille 10×10 également centrée dans D, alors il suffit de décaler l'origine de Ω_2 en $\left(\frac{1}{2}\Delta \pm \varepsilon, \frac{1}{2}\Delta \pm \varepsilon\right)$ avec $\varepsilon = 10^{-6}$, ce qui a évidemment un impact négligeable sur la validité de la simulation conditionnelle, du moins tant que $c_0 \simeq 0$. Pour c_0 quelconque, la solution la plus simple est de forcer $\gamma(0) = \varepsilon$, avec par exemple $\varepsilon = 10^{-4}$.

4.4 Régularisation

Une variable régionalisée $z(\cdot)$ est définie comme une fonction de supports ponctuels $x \in D$ alors qu'en pratique les données sont mesurées ou observées sur des supports surfaciques $v(x)$, x désignant alors le centre de gravité de v, ou plus généralement le centroïde de v. La valeur de la VR sur v est par conséquent une intégrale d'espace :

$$z_v(x) = \frac{1}{[v]} \int_v z(y)\, dy \tag{4.44}$$

avec $[v]$ l'aire de v, et y un point décrivant v. En géostatistique, cette fonction est nommée la *régularisée* de $z(\cdot)$ sur $v(x)$ et notée $z_v(\cdot)$. Le passage de $z(\cdot)$ à $z_v(\cdot)$ est connu principalement sous le nom de *régularisation* (Matheron 1965, Journel & Huijbregts 1978) ou *intégration locale* (Matérn 1960, pp. 59-62).

Afin d'illustrer le phénomène de régularisation, nous considérons une VR définissant initialement une image de résolution 120×120, avec un pixel de largeur $\Delta = 1$. L'image initiale (Fig. 4.2.1a) est générée par simulation d'une FA de variogramme périodique. Une pyramide (Section 2.3.4.1) comportant quatre niveaux est obtenue en calculant la moyenne locale dans chaque quadrant de 2×2 pixels, ce qui produit successivement une image de résolution 60×60 ($\Delta = 2$) (Fig. 4.2.2a), une image de résolution 30×30 ($\Delta = 4$) (Fig. 4.2.3a), et enfin une image de résolution 15×15 ($\Delta = 8$) (Fig. 4.2.4a). La régularisation se traduit par un phénomène de lissage, autrement dit, par une diminution de la variance des valeurs (Zhang *et al.* 1990). En conséquence, la pépite et le seuil du variogramme décroissent progressivement lorsque la résolution diminue, *i.e.* lorsque Δ augmente (Fig. 4.2.1b à 4.2.4b).

Lorsque la taille du support est négligeable devant celle du domaine, il est raisonnable de traiter $z_v(\cdot)$ comme s'il s'agissait de $z(\cdot)$ et par conséquent, $Z_v(\cdot)$ comme s'il s'agissait de $Z(\cdot)$. En revanche, si $[v]$ n'est pas négligeable devant $[D]$, le problème qui se pose est d'obtenir le variogramme $\gamma(h)$ défini pour $Z(\cdot)$ à partir du variogramme $\gamma_v(h)$ défini pour $Z_v(\cdot)$. Si $Z(\cdot)$ est une FAST-2 d'espérance m, alors $Z_v(\cdot)$ est également une FAST-2, de même espérance (Journel & Huijbregts 1978). En revanche, la régularisation réduit la variabilité de sorte que le seuil du variogramme $\gamma_v(h)$ est inférieur à celui de $\gamma(h)$. On montre que $\gamma_v(h)$ et $\gamma(h)$ sont liés par la relation (Journel & Huijbregts 1978, p. 78) :

$$\gamma_v(h) = \bar{\gamma}(v, \tau_h v) - \bar{\gamma}(v, v) \tag{4.45}$$

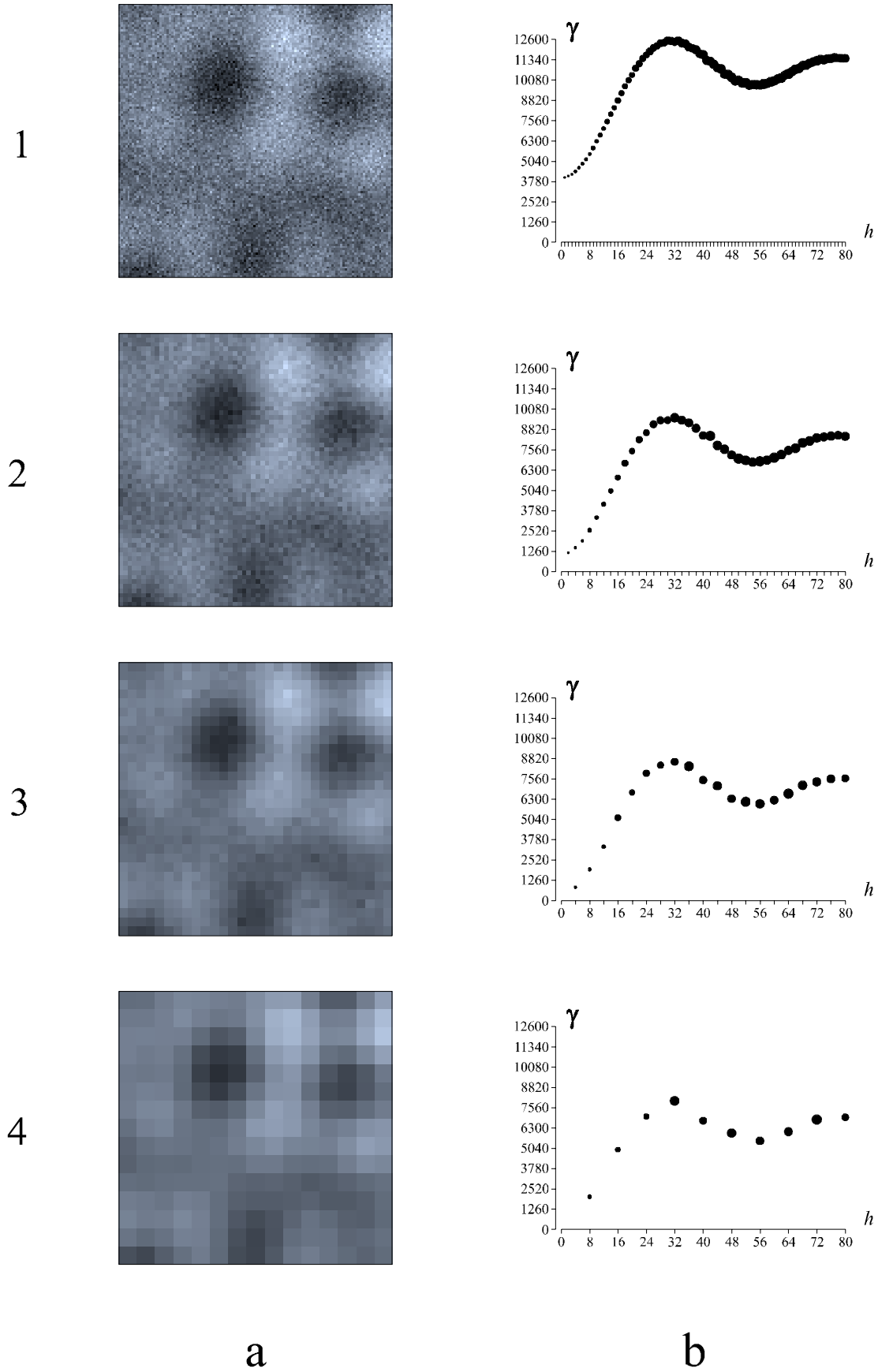
avec τ_h l'opérateur de translation par h. Pour des distances h grandes par rapport aux dimensions de v, on a $\bar{\gamma}(v, \tau_h v) \simeq \gamma(h)$, d'où l'approximation :

$$\gamma_v(h) \simeq \gamma(h) - \bar{\gamma}(v, v) \tag{4.46}$$

Le variogramme dérégularisé $\gamma(h)$ peut être obtenu à partir de $\gamma_v(h)$ en proposant une succession de variogrammes candidats et en retenant celui pour lequel l'approximation (4.46) est vérifiée, au moins pour $h \gg [v]$. Dans cette procédure, il n'est pas nécessaire de calculer $\bar{\gamma}(v, v)$ avec une grande précision, et la méthode de Cauchy-Gauss exposée dans Journel & Huijbregts (1978, pp. 98-108) donne des résultats très satisfaisants.

Dans ce qui suit, nous considérons toujours que les supports sont ponctuels et ne faisons plus référence à la dérégularisation du variogramme[12].

[12] Exemples dans He *et al.* (1994), Aubry (1996b) et Bellehumeur *et al.* (1997).



Figure˜4.2: Régularisation du variogramme pour une pyramide d'images à quatre niveaux. (1) Résolution 120×120. (2) Résolution 60×60. (3) Résolution 30×30. (4) Résolution 15×15. (a) Image. (b) Variogramme local.

4.5 Variance d'extension

Soient deux domaines v et v' sur lesquelles sont définies les régularisées $Z(v)$ et $Z(v')$. Notons $\sigma_E^2(v, v')$ la variance de la différence $Z(v) - Z(v')$, soit :

$$\sigma_E^2(v, v') = \mathrm{Var}\left[Z(v) - Z(v')\right] \tag{4.47}$$

La variance (4.47) est nommée *variance d'extension* de v à v' et peut s'interpréter comme la variance de l'erreur commise en étendant les valeurs de $Z(v)$ au support v'. Dans le cadre d'une FAST-2, la variance d'extension se calcule selon :

$$\sigma_E^2(v, v') = \bar{C}(v, v) + \bar{C}(v', v') - 2\bar{C}(v, v') \tag{4.48}$$

avec la simplification d'écriture :

$$\bar{C}(v, v') = \frac{1}{[v]\,[v']} \int_v \int_{v'} C(x, y)\, dxdy \tag{4.49}$$

où $[v]$ est l'aire du domaine v. Dans le cadre d'une FAI-0, la variance d'extension se calcule de façon analogue selon :

$$\sigma_E^2(v, v') = 2\bar{\gamma}(v, v') - \bar{\gamma}(v, v) - \bar{\gamma}(v', v') \tag{4.50}$$

avec une simplication d'écriture analogue à la précédente :

$$\bar{\gamma}(v, v') = \frac{1}{[v]\,[v']} \int_v \int_{v'} \gamma(x, y)\, dxdy \tag{4.51}$$

Si v et v' se réduisent à des supports ponctuels x et $x+h$, alors la variance d'extension n'est pas autre chose que le variogramme $2\gamma(h)$: le variogramme représente donc la plus simple des variances d'extension (Journel & Huijbregts 1978, Chauvet 1994). En considérant à présent que $Z(v)$ "estime" $Z(v')$, alors la variance d'extension est une *variance d'erreur d'estimation*.

4.6 Variance d'erreur d'estimation

Soit une variable régionalisée $z(\cdot)$ définie sur un domaine D. Soit z_D la moyenne globale définie comme l'intégrale d'espace :

$$z_D = \frac{1}{[D]} \int_D z(x)\, dx \tag{4.52}$$

Si la VR est connue par n valeurs de supports ponctuels $\{z(x_i) \mid i = 1, \ldots, n\}$, alors z_D peut être estimée par la combinaison linéaire :

$$z_D^* = \sum_{i=1}^{n} \lambda_i z(x_i) \tag{4.53}$$

avec $\sum_i \lambda_i = 1$. L'erreur d'estimation de la moyenne globale peut être définie comme $z_D^* - z_D$. En considérant que la VR $z(\cdot)$ est une réalisation d'une FA $Z(\cdot)$, la géostatistique probabilise *ipso facto* l'erreur d'estimation $z_D^* - z_D$ en $Z_D^* - Z_D$.

Par définition, la variance d'erreur d'estimation $\mathrm{Var}[Z_D^* - Z_D]$ est la variance de l'erreur d'estimation $z_D^* - z_D$ pour une infinité de réalisations de la FA $Z(\cdot)$ sur D. Soit un ensemble fini de L réalisations indépendantes $\left\{z^{(\ell)}(\cdot) \mid \ell = 1, \ldots, L\right\}$ sur un ensemble de N supports $\Omega = \{x_i \mid i = 1, \ldots, N\}$ et $z_D^{*(\ell)}$ une estimation de $z_D^{(\ell)}$ calculée sur la base des valeurs $\left\{z^{(\ell)}(x) \mid x \in \Omega_1\right\}$ avec $\Omega_1 \subset \Omega$. La variance d'erreur d'estimation peut être estimée par :

$$\widehat{\mathrm{Var}}[Z_D^* - Z_D] = \frac{1}{L-1} \sum_{\ell=1}^{L} \left\{ z_D^{*(\ell)} - z_D^{(\ell)} - \widehat{\mathrm{E}}[Z_D^* - Z_D] \right\}^2 \tag{4.54}$$

avec l'estimateur de l'espérance :

$$\widehat{\mathrm{E}}[Z_D^* - Z_D] = \frac{1}{L} \sum_{\ell=1}^{L} \left\{ z_D^{*(\ell)} - z_D^{(\ell)} \right\} \tag{4.55}$$

Autrement dit, la variance $\mathrm{Var}[Z_D^* - Z_D]$ peut être calculée numériquement par simulation non conditionnelle de $Z(\cdot)$ sur D. Il est cependant plus simple de formuler la variance d'erreur d'estimation globale directement comme une fonction du variogramme $\gamma(h)$ (Matheron 1965, Journel & Huijbregts 1978) :

$$\sigma_E^2 = 2 \sum_{i=1}^{n} \lambda_i \bar{\gamma}(x_i, D) - \bar{\gamma}(D, D) - \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \gamma(x_i, x_j) \tag{4.56}$$

avec

$$\bar{\gamma}(x_i, D) = \frac{1}{[D]} \int_D \gamma(x_i, x)\, dx \tag{4.57}$$

et

$$\bar{\gamma}(D, D) = \frac{1}{[D]^2} \int_D \int_D \gamma(x, x')\, dx dx' \tag{4.58}$$

L'intégrale d'espace $\bar{\gamma}(x_i, D)$ représente le variogramme moyen entre le point x_i et tous les autres points décrivant le domaine D. La double intégrale d'espace $\bar{\gamma}(D, D)$ représente le variogramme moyen sur D.

Lorsque l'estimation globale s'effectue par combinaison linéaire d'estimations locales, l'erreur d'estimation globale résulte également d'une combinaison linéaire d'erreurs locales. Cependant, les erreurs locales sont généralement corrélées et leurs variances ne peuvent pas être combinées aussi simplement que les estimations elles-mêmes (Journel & Huijbregts 1978).

4.6.1 Combinaison de variances d'erreurs locales

Soit D un domaine partitionné en N supports surfaciques $\{v_i \mid i = 1, \ldots, N\}$. Pour simplifier, nous considérons que D est partitionné par une tessellation régulière (Section 2.1.2.3, p. 15), *i.e.* les tuiles ont même géométrie v. Considérons en outre que la VR est estimée sur chaque tuile à partir de n valeurs de supports ponctuels $\{z(x_i) \mid i = 1, \ldots, n\}$. La moyenne globale z_D peut être estimée par la moyenne des estimations locales, soit :

$$z_D^* = \frac{1}{N} \sum_{i=1}^{N} z_{v_i}^* \tag{4.59}$$

et l'erreur d'estimation globale est composée d'erreurs d'estimations locales :

$$z_D^* - z_D = \frac{1}{N} \sum_{i=1}^{N} \left(z_{v_i}^* - z_{v_i} \right) \tag{4.60}$$

avec

$$z_{v_i} = \frac{1}{[v_i]} \int_{v_i} z(x)\, dx \tag{4.61}$$

$$z_{v_i}^* = \sum_{\alpha=1}^{n} \lambda_i^\alpha z(x_\alpha) \tag{4.62}$$

où λ_i^α est le poids affecté à la donnée $z(x_\alpha)$ dans l'estimation de z_{v_i}. En probabilisant, on obtient immédiatement :

$$Z_D^* - Z_D = \frac{1}{N} \sum_{i=1}^{N} \left(Z_{v_i}^* - Z_{v_i} \right) \tag{4.63}$$

et la variance d'erreur d'estimation s'écrit (Crozel & David 1985) :

$$\sigma_E^2 = \mathrm{E}\left[(Z_D^* - Z_D)^2 \right] = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathrm{E}\left[\left(Z_{v_i}^* - Z_{v_i} \right) \left(Z_{v_j}^* - Z_{v_j} \right) \right] \tag{4.64}$$

soit :

$$\sigma_E^2 = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left[2 \sum_{\alpha=1}^{n} \lambda_i^\alpha \bar{\gamma}(x_\alpha, v_j) - \bar{\gamma}(v_i, v_j) - \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} \lambda_i^\alpha \lambda_j^\beta \gamma(x_\alpha, x_\beta) \right] \tag{4.65}$$

Cette expression peut se simplifier (Crozel & David 1985) pour se ramener à la forme (4.56). En effet, du fait de la linéarité des estimateurs, Z_D^* peut s'exprimer directement comme une combinaison des VA $Z(x_\alpha)$, soit :

$$Z_D^* = \sum_{\alpha=1}^{n} \lambda_D^\alpha Z(x_\alpha) \tag{4.66}$$

avec λ_D^α le poids moyen affecté à $Z(x_\alpha)$, calculé selon :

$$\lambda_D^\alpha = \frac{1}{N} \sum_{i=1}^{N} \lambda_i^\alpha \tag{4.67}$$

La variance d'erreur d'estimation globale s'écrit plus simplement (Crozel & David 1985) :

$$\sigma_E^2 = 2\sum_{\alpha=1}^{n} \lambda_D^\alpha \bar{\gamma}\left(x_\alpha, D\right) - \bar{\gamma}\left(D, D\right) - \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} \lambda_D^\alpha \lambda_D^\beta \gamma\left(x_\alpha, x_\beta\right) \tag{4.68}$$

4.6.2 Composition de termes de ligne et de section

Soit $D \subset \mathbb{R}^2$ un domaine rectangulaire et $s \subset D$ un ensemble de n supports présentant une direction de plus grande densité, *i.e.* un ensemble de n_ℓ lignes (ou transects) parallèles, $\{\ell_i \mid i = 1, \ldots, n_\ell\}$. Chaque ligne ℓ_i est discrétisée par n_i supports $\{s_j \mid j = 1, \ldots, n_i\}$, soit :

$$n = \sum_{i=1}^{n_\ell} n_i \tag{4.69}$$

A chaque ligne ℓ_i il est possible d'associer un rectangle d'influence ou *section* s_i de façon à former une tessellation de D, la ligne ℓ_i constituant la ligne médiane de s_i. En estimant la valeur moyenne z_{s_i} pour chaque section s_i par la valeur moyenne z_{ℓ_i} de chaque ligne ℓ_i, et la valeur moyenne z_{ℓ_i} par combinaison linéaire des valeurs $\{z_j \mid j = 1, \ldots, n_i\}$, deux types d'erreur d'extension s'ajoutent (Matheron 1965) :

- une erreur d'extension des points aux lignes,
- une erreur d'extension des lignes aux sections (et donc à D).

Si les erreurs peuvent être considérées — au moins en première approximation — comme indépendantes, et en probabilisant toutes les quantités mises en jeu, la variance d'extension des points aux lignes σ_ℓ^2 s'ajoute à la variance d'extension des lignes aux sections σ_s^2 ce qui donne $\sigma_E^2 = \sigma_\ell^2 + \sigma_s^2$. L'intérêt de cette composition consiste à décomposer un calcul posé dans $\mathbb{R}^2$ en deux calculs effectués dans $\mathbb{R}$, donc plus simples (Matheron 1965). A notre connaissance, en écologie cette technique a été exclusivement utilisée pour traiter les campagnes acoustiques effectuées en halieutique (*e.g.*, Petitgas 1991, 1993, Simard *et al.* 1993, Guiblin 1997).

4.6.3 Calcul numérique

Considérons l'estimation globale sur v à partir de n valeurs de supports ponctuels. La variance d'erreur d'estimation (4.56) peut s'écrire $\sigma_E^2 = 2s_1 - s_2 - s_3$ avec :

$$s_1 = \sum_{i=1}^{n} \lambda_i \frac{1}{[v]} \int_v \gamma\left(x_i, x\right) dx \tag{4.70}$$

$$s_2 = \frac{1}{[v]^2} \int_v \int_v \gamma\left(x, x'\right) dx dx' \tag{4.71}$$

$$s_3 = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma\left(x_i, x_j\right) \tag{4.72}$$

Matheron (1965) fait remarquer que la double somme discrète s_3 est proche du terme mixte s_1 qui à son tour est proche de la double intégrale s_2, de sorte que le calcul numérique de σ_E^2 fait intervenir des différences secondes entre termes d'ordres de grandeur très voisins. Le calcul numérique des termes doit donc s'effectuer avec suffisamment de précision (Crozel & David 1985), et en tout cas avec des méthodes de précision analogue, sous peine d'obtenir une valeur de σ_E^2 négative (Isaaks & Srivastava 1989, p. 523). Le terme s_3 qui dérive des données est calculé de façon exacte comme :

$$s_3 = 2\sum_{i<n}\lambda_i \sum_{i<j}\lambda_j \gamma\left(x_i, x_j\right) + \sum_i \lambda_i^2 \gamma\left(x_i, x_i\right) \tag{4.73}$$

puisque $\gamma\left(x_i, x_j\right) = \gamma\left(x_j, x_i\right)$. Comme $\gamma\left(x_i, x_i\right) = 0$, le terme de droite disparaît, ce qui donne :

$$s_3 = 2\sum_{i<n}\lambda_i \sum_{i<j}\lambda_j \gamma\left(x_i, x_j\right) \tag{4.74}$$

Pour des domaines de géométries simples (segments, rectangles, carrés) et certains modèles de variogrammes, le calcul des termes s_1 et s_2 peut s'effectuer grâce à des *fonctions auxiliaires* (Matheron 1965, Clark 1976, Journel & Huijbregts 1978, pp. 103-123, Webster & Burgess 1984) ou grâce à des formules d'approximation du type MacLaurin (Chadœuf *et al.* 1998).

L'utilisation des fonctions auxiliaires ou d'abaques (Journel & Huijbregts 1978, pp. 103-144), bien qu'encore mentionnée dans la littérature récente (*e.g.*, He *et al.* 1994, Bellehumeur & Legendre 1997, Bellehumeur *et al.* 1997, Myers 1997, pp. 324-326), est à la fois trop peu générale, peu pratique et obsolète. En effet, en supposant que le variogramme $\gamma\left(h\right)$ soit valide pour v entier, le calcul de σ_E^2 doit pouvoir s'effectuer pour n'importe quelle géométrie de v, et n'importe quel modèle autorisé de $\gamma\left(h\right)$. Chauvet (1994) fait également remarquer qu'il n'est plus guère intéressant d'avoir recours à des abaques ou à des fonctions préalablement tabulées lorsque les moyens informatiques permettent un calcul effectif au cas par cas. Le calcul des termes s_1 et s_2 nécessite donc de choisir une méthode d'intégration.

Les méthodes d'intégration numérique classiques (*e.g.*, méthodes de Simpson ou de Romberg) étendues à $\mathbb{R}^2$ sont coûteuses en temps de calcul et surtout peu pratiques à utiliser lorsque la géométrie de v n'est pas simple, *i.e.* pas rectangulaire. Nous avons utilisé dans Aubry & Debouzie (1999a) une procédure générale utilisant des approximations discrètes calculées par intégration de Monte-Carlo (Rubinstein 1981). Ainsi, les termes s_1 et s_2 sont approximés respectivement par :

$$s_1^{'} = \sum_{i=1}^{n}\lambda_i \frac{1}{n^{'}}\sum_{j=1}^{n^{'}}\gamma\left(x_i, x_j\right) \tag{4.75}$$

$$s_2^{'} = \frac{1}{n^{''}n^{'''}}\sum_{k=1}^{n^{''}}\sum_{\ell=1}^{n^{'''}}\gamma\left(x_k, x_\ell\right) \tag{4.76}$$

avec $\left\{x_j \mid j = 1, \ldots, n^{'}\right\}$, $\left\{x_k \mid k = 1, \ldots, n^{''}\right\}$ et $\left\{x_\ell \mid \ell = 1, \ldots, n^{'''}\right\}$ trois semis de points dans v. Soit $\mathcal{P}$ un prédicat topologique permettant de tester si $x \in \overset{\circ}{v}$ ou $x \in \overline{v}$, avec $\overset{\circ}{v}$ l'intérieur de v et $\overline{v}$ son adhérence (Section 2.2.2.1).

Quelle que soit la géométrie de v, il est possible de répartir les semis de points dans le rectangle minimum englobant v (Section 2.2.2.3), puis de restreindre ces semis à v par application de $\mathcal{P}$ de sorte que l'intégration de Monte-Carlo permet de calculer σ_E^2 pour tout v. La qualité des approximations — faible erreur d'approximation et stabilité des résultats — dépend alors :

- de la position relative des points dans v,
- de la quantité de points dans v,
- du générateur de nombres pseudo-aléatoires.

L'expérience acquise lors d'une étude de cas particulièrement exigeante dans laquelle la portée du variogramme est grande vis-à -vis du domaine (Aubry & Debouzie 1999a) nous permet de discuter succinctement l'influence de chacun des trois facteurs mentionnés.

4.6.3.1 Position relative des points

La discrétisation de v s'effectue de façon classique par une grille de points centrée dans v (David 1977, p. 192, Journel & Huijbregts 1978, p. 95, Simard *et al.* 1992, Myers 1997, pp. 322-324). Cependant, puisqu'il s'agit d'intégrer le variogramme $\gamma(h)$, *i.e.* une fonction de la distance, nous avons attiré l'attention dans Aubry & Debouzie (1999a) sur le fait que la précision de l'intégration de Monte-Carlo dépend grandement de la façon dont est approximée la fonction densité de probabilité, notée $fxv(h)$, des distances entre les supports x des données et v (calcul du terme s_1), ainsi que la fonction densité de probabilité, notée $fv(h)$, des distances dans v (calcul du terme s_2). Matérn (1960, p. 21) exprime explicitement la covariance entre deux VA $Z(v_i)$ et $Z(v_j)$ d'un processus stochastique isotrope défini dans $\mathbb{R}^n$ comme :

$$\mathrm{Cov}\left[Z(v_i), Z(v_j)\right] = \sigma^2 \int_{\mathbb{R}^n} \rho(h)\, dF(h) \tag{4.77}$$

avec $\rho(h)$ la fonction de corrélation, σ^2 la variance et $F(h)$ la fonction de répartition des distances h entre deux points choisis aléatoirement dans v_i et v_j. Matérn (1960, p. 23) en déduit immédiatement la méthode de Monte-Carlo suivante : si $\left\{x_i, x_i^{'} \mid i = 1, \ldots, m\right\}$ constituent m paires de points tirés au hasard dans v_i et v_j, alors l'expression :

$$\frac{\sigma^2}{m} \sum_{i=1}^{m} \rho\left(x_i, x_i^{'}\right) \tag{4.78}$$

est une estimation sans biais de (4.77).

Il apparaît immédiatement que la discrétisation par un semis de points répartis selon une grille constitue en principe **la pire des approches** en ce qui concerne l'approximation de $fxv(h)$ et de $fv(h)$ parce que peu de distances différentes sont représentées, et selon des fréquences relatives peu représentatives de la distribution continue sous-jacente. Le semis aléatoire mentionné par Matérn (1960, p. 23) constitue évidemment un meilleur schéma de discrétisation, mais l'expérience nous a montré qu'un semis aléatoire rendu spatialement plus homogène grâce à l'utilisation d'une grille, conduit à des approximations plus précises — l'erreur est souvent plus faible, et moins variable d'un semis à l'autre — que dans le cas du semis aléatoire.

Pour calculer $s_1^{'}$ nous recommandons que $s^{'} = \left\{x_j \mid j = 1, \ldots, n^{'}\right\}$ soit un semis aléatoire stratifié par une grille, à un point par maille. De façon analogue, pour calculer $s_2^{'}$ nous recommandons que $s^{''} = \left\{x_k \mid k = 1, \ldots, n^{''}\right\}$ et $s^{'''} = \left\{x_l \mid l = 1, \ldots, n^{'''}\right\}$ soient obtenus à partir d'un semis aléatoire stratifié par une grille, à deux points par maille, un point appartenant à $s^{''}$ et l'autre à $s^{'''}$. Il convient de s'assurer que les deux points sont distincts afin d'éviter l'*effet zéro*, *i.e.* une éventuelle sous-estimation de l'intégrale $\bar{\gamma}(v, v)$ causée par de nombreuses valeurs nulles $\gamma(0) = 0$ (Journel & Huijbregts 1978, pp. 96-97).

Bellehumeur *et al.* (1997) mentionnent également une autre méthode qui consiste à générer aléatoirement les distances elles-mêmes. Nous considérons que cette approche ne peut pas donner un résultat correct puisque la distribution des distances sera — par définition de la loi uniforme — rectangulaire, et donc sans aucun rapport avec la véritable forme de $fxv(h)$ et de $fv(h)$.

4.6.3.2 Quantité de points

Afin de minimiser le temps de calcul, Journel & Huijbregts (1978) considèrent que dans les applications minières il suffit de discrétiser le domaine (ou bloc) par une grille de 6×6 points. Dans le domaine halieutique, Simard *et al.* (1993) utilisent une grille encore plus réduite, de seulement 3×3 points.

En fait, la quantité de points nécessaire afin d'obtenir une bonne approximation dépend, notamment, de la portée du variogramme par rapport à la taille du domaine d'intégration. Lorsque la portée est grande vis-à-vis du domaine, il s'avère nécessaire d'augmenter considérablement la quantité de points. En considérant un STR, il suffit d'augmenter progressivement la quantité de points en diminuant la taille de la maille élémentaire, jusqu'à ce que l'approximation ait suffisamment convergé.

4.6.3.3 Générateur de nombres pseudo-aléatoires

Pour un domaine v, un variogramme $\gamma(h)$ et une discrétisation donnés, le générateur de nombre pseudo-aléatoires constitue une source de variabilité de l'approximation. Sans considérer l'impact du choix du générateur lui-même, l'expérience montre que même le choix de la graine du générateur a un impact sur l'approximation. L'ordre de grandeur de cet impact doit être connu afin de relativiser l'écart observé entre deux variances. Cet aspect est illustré dans Aubry & Debouzie (1999a).

4.6.4 Interprétation

En examinant (4.56) il apparaît que la variance σ_E^2 dépend uniquement (Journel & Huijbregts 1978) :

- de la structure d'autocorrélation spatiale de la FA représentée par le variogramme $\gamma(h)$,
- de la position relative des points du semis s les uns par rapport aux autres,
- de l'implantation des points de s dans D,
- de la géométrie de D.

La variance σ_E^2 est dite *non conditionnée* parce qu'elle ne dépend pas des valeurs observées (Chauvet 1993). N'étant pas conditionnée par les valeurs, σ_E^2 ne peut pas constituer une mesure de la précision de l'estimation de z_D par z_D^*, *i.e.* pour une réalisation particulière (Journel 1986a, Journel & Rossi 1989, Deutsch & Journel 1992). Ainsi, σ_E^2 reflète la capacité d'un ensemble de supports à estimer précisément la moyenne d'une variable régionalisée, en tenant compte de la taille et de la position des supports, mais indépendamment des valeurs effectivement mesurées ou observées sur ces supports.

4.6.5 Variance d'erreur d'estimation conditionnée

La variance σ_E^2 ne constitue pas à proprement parler une mesure de précision de l'estimation mais plutôt une mesure d'incertitude dépendant de la configuration. Afin d'obtenir une véritable mesure de précision, il convient de définir une variance conditionnée par les valeurs comme (Goovaerts 1997) :

$$\mathrm{Var}\left[Z_D^* - Z_D \mid Z\left(x_i\right) = z\left(x_i\right) \ ; \ i = 1, \ldots, n\right] \tag{4.79}$$

que nous notons plus simplement σ_C^2. A notre connaissance, et à la différence de σ_E^2, il n'existe pas de formule analogue à (4.56) qui permettrait de calculer σ_C^2 directement à partir du variogramme et des données, et il faut par conséquent la calculer comme en (4.54) mais en conditionnant les réalisations de $Z\left(\cdot\right)$ sur D par les données $\{z\left(x_i\right) \mid i = 1, \ldots, n\}$. Cette application de la simulation conditionnelle est illustrée notamment dans le Chapitre 6 et dans Aubry & Debouzie (1999b).

Chapitre 5

Echantillonnage spatial

> *"... reliance on a realistic and valid superpopulation model can give powerful inference"* (Cassel *et al.* 1977)
>
> *"However, any estimate must depend on certain assumptions about the form of the population which is being sampling and is likely to be vitiated insofar as these assumptions are false"* (Cochran 1946)

Une étude écologique est toujours restreinte à un domaine spatial borné D. Nous considérons d'une façon générale que D est un domaine quelconque du géoïde. Toutefois, pour simplifier toutes les opérations géométriques, nous assimilons D à une région du plan. Le domaine D étant spécifié, la population statistique considérée ou *population cible* correspond à toutes les unités élémentaires localisées dans D. Il convient de distinguer, notamment (Pielou 1969) :

- les phénomènes abiotiques définis en tout point $x \in D$ (*e.g.*, chaleur, humidité),
- les phénomènes biotiques pour lesquels :
  - il est possible d'identifier des entités discrètes ou *individus* (*e.g.*, les insectes d'un champ),
  - les individus ne sont pas clairement délimités, par exemple dans le cas d'une reproduction végétative (*e.g.*, fougères) ou d'organismes coloniaux (*e.g.*, coraux).

Dans le cas des phénomènes abiotiques, il est évidemment impossible de dresser une liste des unités parce que le nombre de points $x \in D$ est infini. En ce qui concerne les phénomènes biotiques, bien que la population cible soit finie, il est souvent impossible de dresser une liste des unités élémentaires qui la composent. C'est le cas des plantes et de leurs graines dispersées sur le sol ou des insectes et de leurs oeufs, des poissons d'un lac, des nématodes d'un champ, des arbres d'une forêt tropicale, etc. Une solution générale à ce problème consiste à associer les unités de la population cible aux unités d'une autre population dont on peut en dresser la liste (Hansen *et al.* 1983). Les phénomènes écologiques se déployant dans l'espace géographique, la population cible est naturellement associée à une *population spatiale* constituée d'un ensemble de supports $\mathcal{U} \in D$. En pratique, il est évident que le caractère sessile ou hautement sédentaire des organismes se prête bien à l'échantillonnage par discrétisation de l'espace (Robson 1982), mais il peut en être de même pour des organismes capables de dispersion, selon les échelles d'espace

et de temps de l'étude. Les situations rencontrées en écologie sont trop diverses pour être discutées ici, aussi, nous considérons de façon générale que l'échantillonnage spatial est approprié.

Les phénomènes écologiques (biotiques ou abiotiques) sont appréhendés à travers un certain nombre de variables. Du fait du recours à une population spatiale, chaque variable considérée est nécessairement une variable régionalisée (VR), définie sur les supports de $\mathcal{U}$. Par extension, l'ensemble des valeurs prises par la VR sur $\mathcal{U}$ peut être nommé *population* ou plus précisément, *population de valeurs*[1].

Pour un domaine D donné, il est possible de définir différentes populations spatiales, ce qui génère *ipso facto* différentes populations de valeurs : il est par conséquent essentiel de préciser la nature de la population spatiale considérée, et en particulier, la taille des supports est très importante, et doit toujours être indiquée (David 1977).

Le domaine D étant borné, la finitude de la population dépend de la nature des supports et de la continuité spatiale de la VR. Lorsque les supports sont ponctuels, la population est infinie si la VR est spatialement continue (*e.g.*, la température à la surface du sol), et finie dans le cas contraire (*e.g.*, le diamètre des arbres d'une parcelle) (Ord 1988). Les supports surfaciques non chevauchants conduisent toujours à une population finie et peuvent être employés aussi bien pour étudier des phénomènes continus que des phénomènes discrets (Arbia 1993). L'application la plus typique de ce type de discrétisation spatiale est la méthode des quadrats, largement utilisée en écologie (Ripley 1981, Maling 1989). Les supports surfaciques peuvent être traités comme des supports ponctuels lorsque leur surface est négligeable par rapport à celle du domaine (Ripley 1981, Maling 1989, Papritz & Webster 1995b). Dans ce cas, il convient de procéder à une nouvelle distinction entre, d'une part, la population originelle finie, et d'autre part, la population opératoire qui peut être considérée comme infinie si la VR est spatialement continue.

Soit $\mathcal{U} = \{u_i \mid i = 1, \ldots, N\}$ une population spatiale finie de N unités u distinctes et identifiables. Il existe deux options pour connaître les valeurs d'une variable régionalisée $z(u)$ selon que l'on examine tout ou partie de la population spatiale $\mathcal{U}$, autrement dit, selon que l'on procède à un recensement exhaustif (*census*) ou à un échantillonnage (*survey sampling*). En écologie, il est exceptionnel de réaliser un recensement exhaustif et les populations sont le plus souvent connues par échantillonnage. Dans ce contexte, un échantillon de la population est obtenu en mesurant ou en observant la VR sur un ensemble de support $s \subset \mathcal{U}$. La répartition spatiale des supports de s dans D est désignée par la suite comme le *motif d'échantillonnage* (*sampling pattern*), et est décrite par les coordonnées des centroïdes des supports.

Il existe deux pratiques selon que le motif d'échantillonnage est choisi sur une base probabiliste ou non probabiliste. Il convient tout d'abord de décrire ces deux pratiques, puis de considérer la question de l'inférence statistique à partir d'un échantillon. Ce faisant, nous précisons les notions de dépendance et d'indépendance, spatiales et statistiques, puis nous abordons la question de l'efficacité de l'échantillonnage en fonction de la VR étudiée. Enfin nous évoquons succinctement le problème du choix d'une stratégie d'échantillonnage spatial.

[1] Par la suite, selon que nous voulons mettre l'accent sur l'échantillonnage ou sur la notion de fonction, nous utilisons *population* ou *variable régionalisée* pour désigner l'ensemble des valeurs associées à la population spatiale $\mathcal{U}$.

5.1 Echantillonnage probabiliste

Soit $s = \{s_i \mid i = 1, \ldots, n\}$ un échantillon comportant n unités prélevées, les unes après les autres, au sein d'une population $\mathcal{U}$. Deux échantillons s et s' sont dits (Hedayat & Sinha 1991) :

- *identiques* si les unités qui les composent sont les mêmes et ont été prélevées dans le même ordre, et *différents* dans le cas contraire,
- *équivalents* si les unités qui les composent sont les mêmes mais que l'ordre de tirage est différent, et *distincts* dans le cas contraire.

Pour un échantillon de taille n il y a donc $n!$ échantillons différents mais équivalents. Pour simplifier, dans ce qui suit nous ne considérons pas l'ordre de tirage des unités et par conséquent, un échantillon est vu comme un ensemble non ordonné.

Soit S l'ensemble des $\binom{N}{n}$ échantillons distincts qu'il est possible de former à partir d'une population $\mathcal{U}$ de taille N, pour une taille d'échantillon $n < N$. Les échantillons $s \in S$ peuvent être répartis dans des ensembles $S_d \subset S$ définis par différents *dispositifs d'échantillonnage probabiliste*. Par la suite, nous parlons simplement de *dispositif d'échantillonnage* en réservant ce terme à l'échantillonnage probabiliste. Un échantillon $s \in S_d$ est obtenu en suivant un *schéma d'échantillonnage* décrivant les étapes opératoires mises en oeuvre pour sélectionner les unités en respectant un dispositif d'échantillonnage donné (Cassel *et al.* 1977, Hedayat & Sinha 1991).

L'objet de cette section est de préciser les concepts de dispositif d'échantillonnage et de schéma d'échantillonnage, puis d'examiner différents dispositifs pour l'échantillonnage d'un domaine $D \subset \mathbb{R}^2$.

5.1.1 Dispositif d'échantillonnage

Formellement, un dispositif d'échantillonnage d basé sur $\mathcal{U}$ est une paire (S_d, P_d) telle que (Hedayat & Sinha 1991) :

1. la distribution de probabilités P_d sur S_d assure $P_d(s) > 0$ pour tout $s \in S_d$,
2. pour tout $u \in \mathcal{U}$, il existe au moins un échantillon $s \in S_d$ tel que $u \in s$.

La seconde condition garantit simplement que le dispositif d'échantillonnage est basé sur $\mathcal{U}$ et pas sur une sous-population.

Sous le dispositif d'échantillonnage $d = (S_d, P_d)$, il est possible de définir une probabilité d'inclusion de premier ordre $\pi_i(d)$ pour toute unité $u_i \in \mathcal{U}$ comme :

$$\pi_i(d) = \sum_{s \in S_d | u_i \in s} P_d(s) \tag{5.1}$$

la somme étant étendue à tous les échantillons $s \in S_d$ qui contiennent l'unité u_i, et une probabilité d'inclusion de second ordre (ou jointe) $\pi_{ij}(d)$ pour $u_i, u_j \in \mathcal{U}$ comme :

$$\pi_{ij}(d) = \sum_{s \in S_d | u_i, u_j \in s} P_d(s) \tag{5.2}$$

la somme étant étendue à tous les échantillons $s \in S_d$ qui contiennent simultanément les unités u_i et u_j (Horvitz & Thompson 1952, Hedayat & Sinha 1991). Les probabilités d'inclusion d'ordres supérieurs sont définies de façon identique. Les probabilités d'inclusion sont connues comme les *constantes structurelles* d'un dispositif d'échantillonnage (Cassel *et al.* 1977, Hedayat & Sinha 1991).

5.1.2 Schéma d'échantillonnage

Un schéma d'échantillonnage peut être défini de façon formelle comme un algorithme $\mathcal{A} = \mathcal{A}\{q_1(\cdot), q_2(s), q_3(\cdot \mid s)\}$ où (Hedayat & Sinha 1991, pp. 5-6) :

- $(S_1, q_1(\cdot))$ est un dispositif d'échantillonnage défini sur l'ensemble d'unités $S_1 \subseteq \{u_i \mid i = 1, \ldots, N\}$, *i.e.* sur $\mathcal{U}$ ($S_1 = \mathcal{U}$) ou un sous-ensemble strict de $\mathcal{U}$ ($S_1 \tilde{\subset} \mathcal{U}$), avec $q_1(\cdot)$ une distribution de probabilités, telle que $0 < q_1(u_i) \leq 1$ avec $u_i \in S_1$ et $\sum_{u_i \in S_1} q_1(s_i) = 1$,
- $q_2(s)$ est une probabilité associée à l'échantillon s, *i.e.* un réel tel que $0 \leq q_2(s) \leq 1$ pour tout $s \in S$,
- pour $q_2(s) > 0$, $(S_2(s), q_3(\cdot \mid s))$ est un dispositif d'échantillonnage défini sur l'ensemble d'unités $S_2(s) \subseteq \{u_j \mid u_j \in \mathcal{U} - s\}$, *i.e.* sur $\mathcal{U} - s$ ($S_2(s) = \mathcal{U} - s$) ou un sous-ensemble strict de $\mathcal{U} - s$ ($S_2(s) \tilde{\subset} \mathcal{U} - s$), avec une distribution de probabilités $q_3(\cdot \mid s)$, telle que $0 < q_3(u_j \mid s) \leq 1$ avec $u_j \in S_2(s)$ et $\sum_{u_j \in S_2(s)} q_3(u_j \mid s) = 1$.

L'algorithme $\mathcal{A} = \mathcal{A}\{q_1(\cdot), q_2(s), q_3(\cdot \mid s)\}$ destiné à construire l'échantillon s se déroule comme suit :

1. Initialiser $s \leftarrow \emptyset$.
2. Utiliser le dispositif d'échantillonnage $(S_1, q_1(\cdot))$ afin de sélectionner une première unité $s_1 \in \mathcal{U}$. Mettre à jour l'échantillon : $s \leftarrow s \cup s_1$.
3. Si $q_2(s) = 0$ alors FIN. Si $q_2(s) = 1$ alors passer à l'étape suivante, sinon, si $0 < q_2(s) < 1$, passer à l'étape suivante avec la probabilité $q_2(s)$ et stopper l'échantillonnage avec la probabilité $1 - q_2(s)$.
4. A l'itération i, utiliser le dispositif d'échantillonnage $(S_2(s), q_3(\cdot \mid s))$ pour sélectionner une autre unité s_i dans $\mathcal{U} - s$. Mettre à jour l'échantillon : $s \leftarrow s \cup s_i$ et aller à l'étape 2.

On peut démontrer que tout schéma d'échantillonnage défini par un algorithme du type $\mathcal{A}\{q_1(\cdot), q_2(s), q_3(\cdot \mid s)\}$ conduit à un dispositif d'échantillonnage unique $d = (S_d, P_d)$, et qu'il existe un seul schéma d'échantillonnage $\mathcal{A}$ sous-jacent à un dispositif d'échantillonnage d. En conséquence, il existe une bijection entre les dispositifs et les schémas d'échantillonnage (Hedayat & Sinha 1991).

5.1.3 Echantillonnage dans le plan

En vertu de la bijection qui existe entre les dispositifs et les schémas d'échantillonnage, un dispositif d'échantillonnage du plan destiné à produire des échantillons de taille n peut être défini de façon opératoire sous la forme d'une procédure de sélection de n points dans $\mathbb{R}^2$.

Parmi la multitude de procédures imaginables, nous considérons uniquement celles qui font intervenir les trois dispositifs fondamentaux suivants :

- échantillonnage aléatoire simple,
- échantillonnage systématique,
- échantillonnage stratifié, à un élément par strate.

Nous décrivons ces trois dispositifs en considérant — pour simplifier l'exposé — que D est un carré d'origine O, de côtés parallèles aux axes (O, x) et (O, y), de largeur L. Puis nous discutons des éventuelles difficultés d'implémentation des trois dispositifs fondamentaux. Enfin, nous traitons succinctement de leur combinaison.

5.1.3.1 Echantillonnage aélatoire simple

L'échantillonnage aléatoire simple (EAS) consiste à prélever au hasard et de façon indépendante n points dans D, chaque point ayant la même probabilité d'inclusion dans l'échantillon, et chaque échantillon de taille n ayant la même probabilité de sélection. En pratique, il suffit de tirer deux nombres aléatoires U_i et $U_i^{'}$ uniformément répartis dans $[0, 1]$, puis de sélectionner le point (x_i, y_i) tel que $x_i = U_i L$ et $y_i = U_i^{'} L$, en répétant cette procédure pour $i = 1, \ldots, n$.

Le dispositif aléatoire simple présente l'inconvénient de suréchantillonner et de sous-échantillonner certaines régions de D puisque les points ne sont pas répartis de façon régulière. L'intérêt de l'EAS réside dans la grande simplicité de sa théorie et dans sa capacité à bien échantillonner la distribution des distances entre les points de D.

5.1.3.2 Echantillonnage systématique

L'échantillonnage systématique (ES) consiste à tirer au hasard un point origine (x_1, y_1) dans un sous-domaine de D, puis à tirer de façon systématique tous les autres points à partir de (x_1, y_1). Considérons l'ES de taille n constitué par une grille de maille carrée de côté $\Delta = L/\sqrt{n}$. Le point (x_1, y_1) est tiré au hasard dans la maille d'origine O de côté Δ selon la procédure de l'EAS, puis les points (x_i, y_j) sont obtenus de façon non indépendante en prenant $x_i = x_1 + (i-1)\Delta$ et $y_j = y_1 + (j-1)\Delta$ pour $i, j = 1, \ldots, \sqrt{n}$. En particulier, l'échantillon centré est obtenu pour $x_1 = \Delta/2$ et $y_1 = \Delta/2$.

En totale opposition avec l'EAS, le dispositif systématique présente l'avantage d'échantillonner D de façon régulière. Mais la théorie de l'ES bidimensionnel pose encore de nombreux problèmes (Dunn & Harrison 1993), et l'ES présente l'inconvénient de mal échantillonner la distribution des distances entre les points de D.

5.1.3.3 Echantillonnage stratifié

L'intention de la statification est de combiner les avantages respectifs de l'ES et de l'EAS en assurant à la fois une répartition assez régulière des points dans D afin d'éviter les agrégats, et un bon échantillonnage de la distribution des distances entre points. La théorie de l'échantillonnage stratifié (STR) est légèrement plus compliquée que celle de l'EAS et,

le cas échéant, elle peut poser des problèmes similaires à ceux rencontrés dans le cas de l'ES (Ripley 1981). Il existe deux critères pour définir les strates (Maling 1989) :

1. le découpage géométrique de l'espace en polygones contigus, éventuellement de même taille,
2. le découpage fondé sur une connaissance préalable de la population.

Dans le cadre de l'échantillonnage du plan, seul le découpage par une tessellation a un sens puisqu'à ce stade, aucune variable n'est prise en considération. L'échantillonnage stratifié consiste alors à partitionner D en strates non chevauchantes et à tirer un EAS, indépendamment au sein de chaque strate (Matérn 1960, Cressie 1991). Le découpage de D peut s'effectuer selon une tessellation régulière dans $\mathbb{R}^2$, *i.e.* par des hexagones, des carrés ou des triangles équilatéraux (Section 2.1.2.3, p. 15). Pour simplifier, nous considérons uniquement le découpage par des carrés parce qu'il est utilisé de façon quasi-universelle (Ripley 1981). Dans le cas particulier d'un STR à un point par strate, il suffit de découper D en n mailles carrées de côtés $\Delta = L/\sqrt{n}$ et de tirer les n points de l'échantillon selon la procédure de l'EAS appliquée indépendamment à chaque maille. Les points (x_i, y_j) sont donc tels que $x_i = [(i-1) + U_{ij}]\,\Delta$ et $y_j = \left[(j-1)U'_{ij}\right]\,\Delta$ pour $i, j = 1, \ldots, \sqrt{n}$, avec U_{ij} et U'_{ij} deux nombres aléatoires uniformément répartis dans $[0, 1]$.

5.1.3.4 Facilité d'implémentation des dispositifs

En pratique, l'échantillon spatial ne concerne qu'un domaine borné $D \subset \mathbb{R}^2$. Ce domaine peut être de forme quelconque, et en particulier concave. Soit R le rectangle englobant au plus près D, de côtés L_x et L_y parallèles aux axes (O, x) et (O, y) (Section 2.2.2.3). Pour échantillonner D au sens d'un certain dispositif, il est possible d'appliquer le schéma correspondant dans R puis d'opérer une restriction aux points $(x, y) \in D$ afin de constituer un échantillon restreint à D (Section 2.2.2.1). Cependant, avec cette pratique il n'est pas toujours facile de maîtriser la taille de l'échantillon de D. Par exemple, si D est de forme irrégulière, l'implémentation d'un ES peut poser des problèmes. En effet, le tirage aléatoire de l'origine de l'ES a pour conséquence que le nombre de points $(x, y) \in D$ n'est pas constant (Brus *et al.* 1999).

En principe, l'implémentation d'un EAS s'avère alors beaucoup plus facile que celle d'un ES puisqu'il suffit de tirer au hasard des points dans R jusqu'à ce que le nombre de points $(x, y) \in D$ soit égal à n. En pratique, on considère classiquement qu'il est très coûteux ou impossible de réaliser un véritable EAS (Hansen & Hurwitz 1943). En écologie, l'implémentation d'un EAS est habituellement jugée plus difficile que celle d'un ES selon une grille (Houllier 1986, Fortin *et al.* 1989). Aujourd'hui, les facilités offertes par le GPS (*Global Positioning System*) pour localiser les points d'échantillonnage sur le terrain permettent de se passer d'un système de coordonnées selon une grille (Gerhards *et al.* 1997, Nelson *et al.* 1999). Si la précision de la localisation[2] par GPS est compatible avec l'échelle de l'étude (*e.g.*, échelle régionale), il devient assez facile de répartir les points d'échantillonnage irrégulièrement (Gerhards *et al.* 1997). Dans ces conditions, l'EAS n'est pas nécessairement plus difficile à implémenter que l'ES.

[2]Le GPS civil ne permet généralement pas une localisation horizontale beaucoup plus précise que 100 m, bien qu'elle puisse atteindre 15 à 25 m lorsque le signal est très peu dégradé. En revanche, le GPS différentiel autorise une précision de l'ordre de 2 m (Nelson *et al.* 1999).

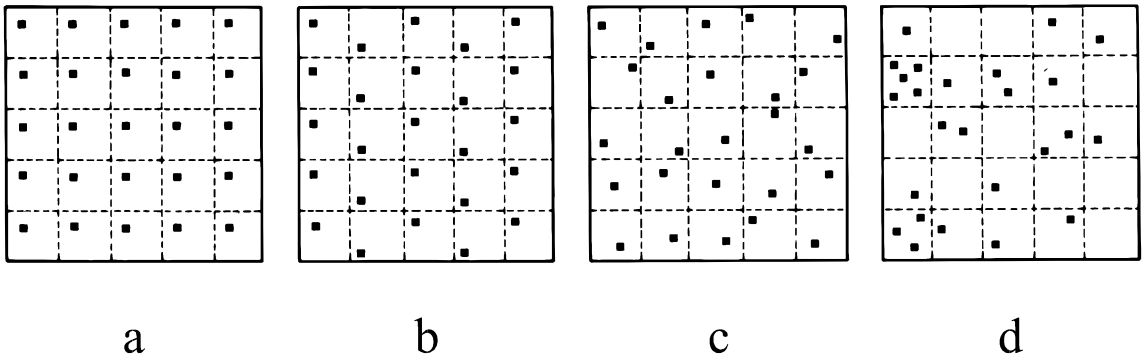
En ce qui concerne l'échantillonnage stratifié, si la géométrie de D est simple — et en particulier si D est carré — l'implémentation d'une tessellation régulière ne pose pas de difficulté. En revanche, si la géométrie de D est quelconque, il devient difficile d'implémenter un nombre de strates fixé *a priori*, de même taille, et *a fortiori*, de même géométrie. Une solution consiste à discrétiser le problème (Brus *et al.* 1999) :

- en "rasterisant" D en une population de pixels $\mathcal{U}$ (Section 2.3.4),
- puis en partitionnant $\mathcal{U}$ en strates comportant approximativement le même nombre de pixels, par minimisation de l'inertie intraclasse calculée sur les coordonnées des centres des pixels, par exemple grâce à un algorithme de type *k-means*[3] (*e.g.*, Hartigan & Wong 1979).

Compte tenu du critère minimisé, ce partitionnement conduit à une tessellation de Voronoï approximative, la méthode des *k-means* et les diagrammes de Voronoï étant en fait étroitement liés (*e.g.*, Schreiber 1991).

5.1.3.5 Combinaison des dispositifs fondamentaux

A partir des trois dispositifs fondamentaux, EAS, STR et ES (Fig. 5.1), de nombreux dispositifs concurrents peuvent être construits. En premier lieu, il est possible d'éviter l'alignement des points dans le cas de l'ES, par exemple en implémentant un échantillonnage systématique en quinconce (Fig. 5.1.b). En outre, il est également possible de décomposer l'échantillonnage bidimensionnel en deux échantillonnages unidimensionnels, *i.e.* selon les axes des abscisses et des ordonnées. Cette décomposition permet d'utiliser un dispositif différent pour chaque axe de coordonnées. Ainsi, Quenouille (1949) étudie 12 dispositifs obtenus en croisant les dispositifs EAS, ES et STR, alignés ou indépendants en abscisses et en ordonnées. Koop (1990) prolonge les travaux de Quenouille (1949) et étudie 21 dispositifs. Dans ce qui suit, nous traitons essentiellement des trois dispositifs fondamentaux, tels que nous les avons décrits précédemment.



Figure˜5.1: Quatre types de motifs d'échantillonnage spatial (d'après Scherrer 1983, Fig. 2.2, p.˜78). (a) Echantillon systématique sur une grille. (b) Echantillon systématique en quinconce. (c) Echantillon stratifié par une grille 5×5, à un élément par strate (d) Echantillon aléatoire simple.

[3] L'algorithme des *k-means* est plus connu en France sous le nom d'algorithme des *nuées dynamiques* (Diday 1971, Diday *et al.* 1982)

5.2 Echantillonnage non probabiliste

Par définition, l'échantillonnage est non probabiliste (*purposive sampling*) lorsque l'inclusion des unités de $\mathcal{U}$ dans l'échantillon s ne fait pas intervenir une règle de sélection aléatoire (Cochran 1977). L'*échantillonnage non probabiliste* est par conséquent également qualifié d'*échantillonnage non aléatoire* (Royall 1970), mais aussi d'*échantillonnage raisonné* (Frontier 1983a, Scherrer 1983), *dirigé* ou *préférentiel* (Fons *et al.* 1997, Goovaerts 1997). Par la suite, nous utilisons le terme d'*échantillonnage préférentiel* comme synonyme d'*échantillonnage non probabiliste*.

En pratique, de nombreux facteurs peuvent conduire à un échantillonnage préférentiel (Cochran 1977, p. 10, Goovaerts 1997, p. 73). Signalons seulement que :

- l'accessibilité des unités sur le terrain peut se révéler très variable, notamment dans un domaine D où la végétation est localement inextricable,
- l'expertise de l'écologiste peut être mise à profit afin d'éviter de prospecter en vain des zones de D dans lesquelles le phénomène étudié a peu de chance de se produire (*e.g.*, la présence d'un certain type d'organisme),
- l'écologiste peut explorer le domaine D et sélectionner les unités qui lui semblent les plus "typiques" en ce qui concerne le phénomène étudié,
- un modèle peut être utilisé afin d'optimiser le motif d'échantillonnage dans D en vue de l'estimation d'un certain paramètre (Chapitre 8).

L'échantillonnage préférentiel semble rare en écologie végétale et en biologie du sol, plus fréquent en écologie animale et en pédologie. En pédologie notamment, l'échantillonnage préférentiel présente l'avantage d'éliminer des facteurs indésirables en considérant uniquement des données provenant de zones particulières du domaine d'étude (Fons *et al.* 1997). En outre, il est traditionnel d'échantillonner un ou plusieurs profils choisis comme étant représentatifs au sein d'une certaine classe pédologique, et les unités échantillonnées ne sont pas réparties au hasard (Voltz *et al.* 1997).

Le recours à l'échantillonnage préférentiel n'est pas criticable en soi parce qu'il est difficile de donner une justification générale et rigoureuse de la nécessité de choisir les unités d'échantillonnage au moyen d'une procédure aléatoire (Royall 1970). Toutefois, le prix à payer en contrepartie de la grande liberté offerte par ce type d'échantillonnage, est l'impossibilité de développer des procédures d'inférence statistique qui ne fassent pas intervenir un modèle (Cochran 1977).

5.3 Inférence statistique

L'*inférence* consiste à généraliser les observations concernant un échantillon à la population dans son entier : l'inférence constitue donc une démarche de type *inductif*. Nous ne traitons pas du raisonnement inductif en général (*cf.* Haton *et al.* 1991), ni du problème particulier de la généralisation écologique des résultats issus de l'inférence statistique (*cf.* Beck 1997), mais uniquement de l'inférence statistique proprement dite, *i.e.* lorsque le lien entre l'échantillon et la population est exprimé en termes probabilistes (Dawid 1983).

Les deux principales écoles d'inférence statistique sont dites *design-based* et *model-based* (Koch & Gillings 1983). L'inférence *model-based* peut être vue comme comprenant l'inférence Bayesienne et l'inférence basée sur un modèle de superpopulation (Koch & Gillings 1983). Dans ce qui suit, nous considérons uniquement l'inférence *model-based* du point de vue non Bayesien. En effet, l'inférence statistique à partir d'un échantillon est principalement effectuée par des procédures *design-based* ou à l'aide d'un modèle de superpopulation (Cassel *et al.* 1977, Särndal 1978, Hansen *et al.* 1983, Iachan 1984, Hansen 1987). Dans ce qui suit, *inférence model-based* est par conséquent utilisé comme un raccourci de *inférence basée sur un modèle de superpopulation*.

Pour une variable régionalisée $z(\cdot)$ associée à la population spatiale $\mathcal{U}$, l'inférence statistique concerne généralement la moyenne globale ou le variogramme. Dans ce cadre, la distinction entre les approches *design-based* et *model-based* a été récemment introduite en géologie par de Gruijter & ter Braak (1990, 1992) et Brus & de Gruijter (1994), en pédologie par Brus & de Gruijter (1993, 1997), relayés par Papritz & Webster (1995a, 1995b), et enfin en écologie par Aubry & Debouzie (1999a, 1999b). Les deux types d'inférences correspondent à deux paradigmes différents à propos desquels les statisticiens s'affrontent (*e.g.*, Royall 1970, Hansen *et al.* 1983, Royall 1983, Smith 1994). Notre propos n'est pas d'essayer de démontrer la supériorité d'un paradigme par rapport à l'autre mais plutôt de les expliquer, puis de discuter de leur utilisation dans le contexte des variables régionalisées. L'inférence faisant nécessairement référence à la notion de représentativité, nous discutons également cette notion fondamentale.

5.3.1 Inférence design-based

Dans l'approche *design-based*, le motif d'échantillonnage implémenté dans un domaine D est obtenu par tirage aléatoire dans l'ensemble S_d des motifs qui peuvent être générés par le dispositif $d = (S_d, P_d)$. La source de stochasticité nécessaire à l'inférence statistique provient donc de la possibilité de randomiser le motif d'échantillonnage dans D, en appliquant le schéma d'échantillonnage correspondant à d. Dans ce contexte, les estimateurs statistiques sont construits sur la base des probabilités d'inclusion π_i et π_{ij}, conditionnellement aux valeurs de l'échantillon $\{z_i \mid i = 1, \ldots, n\}$, sans faire intervenir la structure spatiale de la population sous-jacente, les positions des supports des valeurs de l'échantillon (motif d'échantillonnage), ou la géométrie de D.

Il est facile de donner un sens concret à cette approche. Considérons par exemple une VR spatialement continue $z(\cdot)$ (*e.g.*, la température) et l'échantillonnage aléatoire simple de n points dans un domaine D. Le motif implémenté n'est qu'un motif particulier parmi une infinité de motifs possibles — puisque la population de points est infinie. La valeur estimée θ^* d'un paramètre θ de $z(\cdot)$ sur D dépend de l'échantillon implémenté, et il n'est pas possible d'utiliser de façon adéquate le résultat θ^* sans savoir à quel point cette estimation particulière est semblable à ce que donneraient d'autres échantillons prélevés selon le même dispositif (Cohran *et al.* 1954). Intuitivement, il est évident que la variance d'estimation est obtenue en imaginant que tous les motifs possibles peuvent être implémentés, indépendamment les uns des autres. Mais Royall (1983) insiste sur le fait que les résultats de l'approche *design-based* ne sont valides qu'en moyenne, sur toutes les réplications possibles, ce qui ne garantit pas que l'inférence soit valide dans le cas d'un échantillon particulier.

5.3.2 Inférence model-based

L'approche *model-based* ne tient pas compte d'un éventuel dispositif, mais du motif d'échantillonnage $s = \{s_i \mid i = 1, \ldots, n\}$ implémenté dans un domaine D, des valeurs de l'échantillon $\{z_i \mid i = 1, \ldots, n\}$, et des caractéristiques spatiales de la population sous-jacente. Cette approche est donc totalement opposée, par exemple, à l'inférence basée sur l'EAS, qui tient compte du caractère aléatoire du dispositif, mais ignore les caractéristiques spatiales de la population ou de l'échantillon. Comme l'approche *model-based* ne repose par sur la randomisation du motif d'échantillonnage, l'inférence statistique doit reposer sur une source de stochasticité différente de celle de l'approche *design-based*. Ainsi, au lieu de considérer que le motif d'échantillonnage implémenté est obtenu par tirage aléatoire dans l'ensemble des motifs possibles S_d pour un dispositif donné $d = (S_d, P_d)$, l'approche *model-based* considère que l'échantillon est prélevé dans une population obtenue par tirage aléatoire dans un ensemble de populations de structures spatiales similaires[4]. Ainsi, la population réelle étudiée est elle-même vue comme un échantillon prélevé dans une population de populations, autrement dit, dans une *superpopulation*.

En géostatistique, la superpopulation est décrite par un modèle stochastique spatial du type fonction aléatoire (Section 4.2). Considérons une population spatiale finie $\mathcal{U} = \{u_i \mid i = 1, \ldots, N\}$ et une VR $z(u)$ définie sur $\mathcal{U}$, autrement dit, une population de valeurs $\{z(u_i) \mid i = 1, \ldots, N\}$. La structure stochastique du modèle de superpopulation $Z(u)$ est entièrement caractérisée par la distribution de probabilités conjointes ξ des N variables aléatoires $\{Z(u_i) \mid i = 1, \ldots, N\}$. Les estimateurs statistiques sont construits sur la base de ξ, et ces estimateurs sont conditionnés par les valeurs de l'échantillon $\{z(s_i) \mid i = 1, \ldots, n\}$. Cette démarche est parfaitement conforme à l'avis de Royall (1983) selon lequel l'inférence doit se faire conditionnellement à l'échantillon implémenté, et pas par rapport à sa distribution d'échantillonnage. Dans ce cas, la seule source de stochasticité utilisée dans l'inférence provient du modèle de superpopulation (Royall 1970, Cassel *et al.* 1977, Iachan 1984).

5.3.2.1 Interprétation

Il est essentiel de reconnaître que le concept de superpopulation peut être interprété de plusieurs façons (Cassel *et al.* 1977, p. 81) :

1. la population est réellement tirée d'un univers plus large, et il s'agit de l'idée archétypale de superpopulation,
2. la distribution ξ est modélisée pour décrire un mécanisme stochastique ou un processus du monde réel,
3. la distribution ξ est considérée comme un modèle *a priori* d'opinion subjective, comme dans l'approche Bayesienne,
4. la distribution ξ, tout en n'étant associée ni à un processus du monde réel, ni à une opinion subjective, est utilisée simplement comme un outil mathématique.

[4] Dans le contexte de l'échantillonnage spatial en halieutique, Malinen & Peltonen (1996) utilisent la terminologie *model-based* dans un sens bien différent puisqu'ils considèrent uniquement la modélisation de la distribution statistique des données.

Dans ce mémoire, nous faisons référence exclusivement aux interprétations 2 et 4, selon le contexte. En ce qui concerne l'inférence, l'interprétation 4 est plus spécifiquement développée par la suite (Chapitre 6) et nous n'en discutons pas ici. En revanche, il est facile de donner un sens concret à l'interprétation 2, au moyen des deux exemples suivants.

Premier exemple Imaginons que D est un champ contenant une grille de quadrats dans lesquels on compte, à une époque précise de l'année, le nombre de pousses d'une plante annuelle qui dépassent une certaine taille (variable z). Si cette grille a été mise en place pour suivre l'évolution pluri-anuelle de z, alors il est parfaitement légitime de considérer une superpopulation échantillonnée dans le temps.

Second exemple Considérons à présent une grille de quadrats placée sous un arbre afin d'estimer le nombre total de fruits tombés pendant l'automne. S'il est raisonnable de penser que le type d'architecture de l'arbre conditionne en grande partie la répartition des fruits sur le sol (supposé plan), alors il est légitime de considérer une superpopulation dont chaque population correspond à la répartition spatiale des fruits tombés sous un arbre similaire en termes d'architecture, de taille, d'état phytosanitaire, de localisation géographique, etc.

5.3.3 Représentativité

L'inférence statistique, qu'elle soit *design-based* ou *model-based*, fait nécessairement référence à la notion de *représentativité*, notion vague parce que fortement polysémique (Kruskal & Mosteller 1979a, 1979b, 1979c, 1980, 1988, Frontier 1983a). Nous considérons dans ce mémoire uniquement deux types de représentativité selon que :

- l'échantillon est obtenu par échantillonnage probabiliste,
- l'échantillon reflète fidèlement les caractéristiques de la population.

Selon Scherrer (1983, p. 65), l'échantillon est qualifié d'aléatoire, ou ce qui revient au même, de *représentatif*, lorsque chaque élément de la population a une probabilité connue et différente de 0 d'appartenir à l'échantillon, alors que selon Edgington (1987, p. 5), la représentativité n'est pas liée au caractère aléatoire[5]. Cette contradiction provient évidemment d'un sens différent donné au terme *représentativité*. En fait, il n'est pas correct de parler d'*échantillon représentatif* dans le cas des échantillons tirés au hasard comme le font, notamment, Scherrer (1983) ou Legendre & McArdle (1997). En effet, ce qui rend l'échantillon aléatoire, c'est le processus aléatoire du tirage, pas la composition de l'échantillon réellement construit, de sorte qu'un échantillon aléatoire ne devrait pas être interprété comme un échantillon représentatif (Edgington 1987, Hedayat & Sinha 1991). Dans le contexte de l'inférence *design-based*, il convient donc plutôt de parler de *dispositif représentatif*, essentiellement dans le cas de l'EAS où chaque unité a la même probabilité d'être incluse dans l'échantillon (Cochran *et al.* 1954, Frontier 1983a, p. 43).

[5] Edgington (1987) traite des tests de randomisation qui ne nécessitent pas que l'échantillon soit prélevé aléatoirement.

Nous réservons le terme d'*échantillon représentatif* au second type de représentativité considéré ici, *i.e.* lorsque les caractéristiques de la population sont fidèlement *représentées* dans l'échantillon.

Dans ce mémoire consacré en grande partie à l'utilisation de modèles empruntés à la géostatistique, notre point de vue porte essentiellement sur la représentativité de l'échantillon, ce qui nous amène à illustrer cette notion par un exemple, et à discuter succinctement de la notion de niveau et de degré de représentativité d'un échantillon.

5.3.3.1 Représentativité d'un échantillon

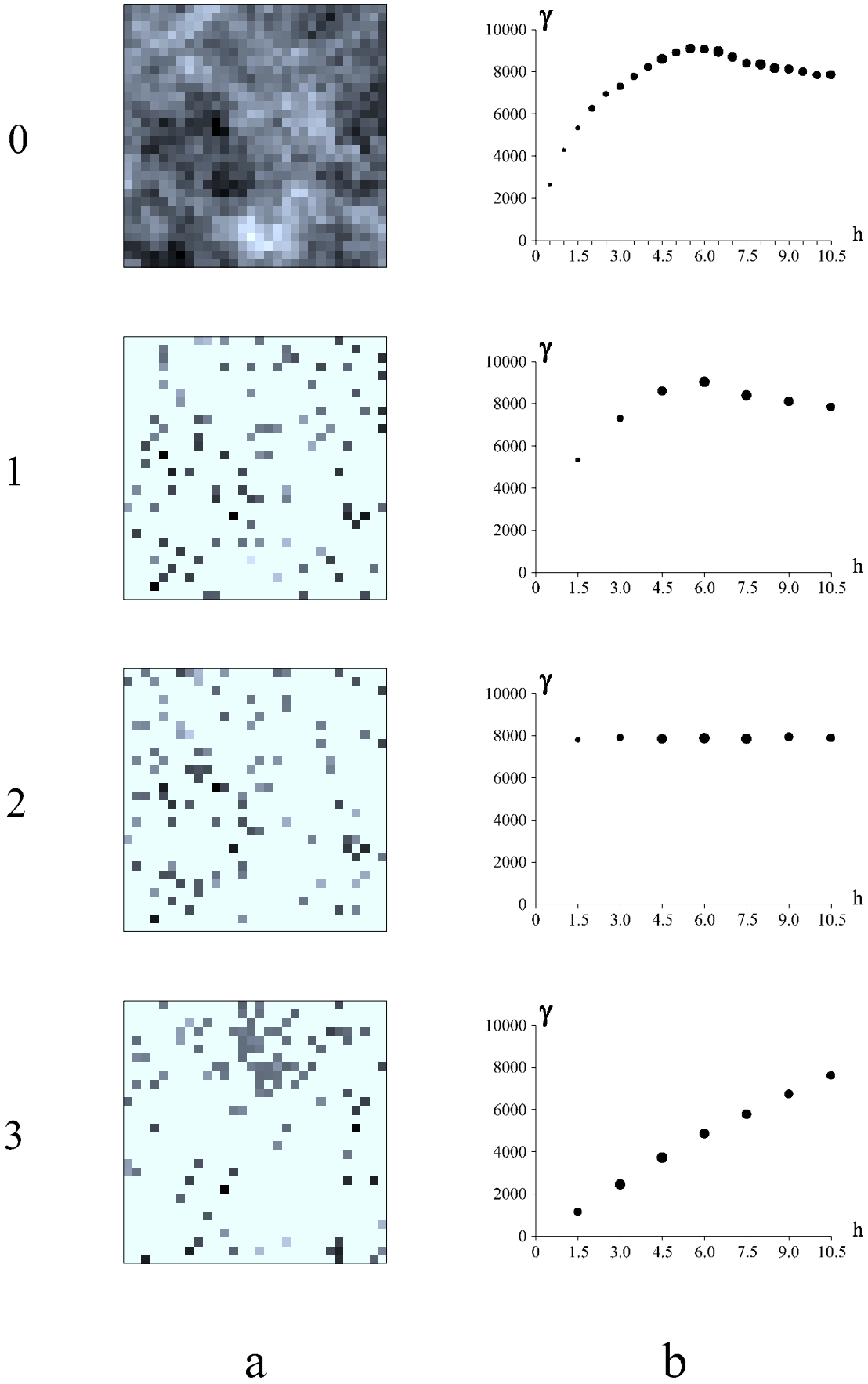
Dans le contexte des variables régionalisées, un échantillon est représentatif s'il permet, notamment, d'estimer fidèlement la structure d'autocorrélation spatiale de la population (*e.g.*, son variogramme). Armstrong (1984b) mentionne des cas où, bien que le variogramme sous-jacent soit linéaire, les variogrammes expérimentaux peuvent :

- correspondre à un effet de pépite pur, ce qui indique l'absence d'autocorrélation spatiale,
- croître plus vite que h^2, ce qui indique la présence d'une tendance.

Afin d'illustrer le problème de la représentativité de l'échantillon vis-à-vis du variogramme, nous considérons une population dont les unités sont localisées aux noeuds d'une grille 30×30 et dont les valeurs sont obtenues par simulation d'une FA de variogramme exponentiel $\text{Expo}\,(1, 7999, 5)$ (Fig. 5.2.0a). Trois échantillons $s^{(1)}$, $s^{(2)}$ et $s^{(3)}$ sont construits, chacun de taille $n = 100$. L'échantillon $s^{(1)}$ (Fig. 5.2.1a) est quasi-optimal du point de vue de la représentativité du variogramme de la population (Fig. 5.2.0b & 5.2.1b). Les échantillons $s^{(2)}$ (Fig. 5.2.2a) et $s^{(3)}$ (Fig. 5.2.3a) ne sont pas représentatifs puisque le variogramme de $s^{(2)}$ exhibe un effet de pépite pur (Fig. 5.2.2b) tandis que celui de $s^{(3)}$ n'est plus caractéristique d'un phénomène de transition (Fig. 5.2.3b).

Les trois échantillons $s^{(1)}$, $s^{(2)}$ et $s^{(3)}$ constituent des exemples extrêmes qui ne doivent rien au hasard parce qu'ils ont été obtenus au moyen d'une procédure d'optimisation combinatoire, mais ils correspondent néanmoins à trois des $\binom{900}{100}$ échantillons possibles dans le cadre de l'EAS. En termes d'échantillons représentatifs, l'EAS peut donc produire le meilleur ($s^{(1)}$) comme le pire ($s^{(2)}$ et $s^{(3)}$).

Cet exemple illustre bien la différence fondamentale entre les notions d'échantillonnage représentatif (EAS) et d'échantillon représentatif ($s^{(1)}$), ainsi que la difficulté inhérente à l'inférence d'un paramètre populationnel tel que le variogramme, à partir d'un échantillon unique.



Figure˜5.2: Représentativité de trois échantillons vis-à-vis du variogramme local. (0) Population. (1), (2) & (3), Echantillons. (a) Représentation cartograghique. (b) Variogramme.

5.3.3.2 Niveau et degré de représentativité d'un échantillon

Dans un contexte superpopulationnel, il convient de distinguer trois *niveaux de représentativité* :

- celui de la population vis-à-vis de la superpopulation (*e.g.*, variogramme local *vs.* variogramme théorique),
- celui de l'échantillon vis-à-vis de la population (*e.g.*, variogramme expérimental *vs.* variogramme local),
- celui de l'échantillon vis-à-vis de la superpopulation (*e.g.*, variogramme expérimental *vs.* variogramme théorique).

L'identification de ces différents niveaux de représentativité peut sembler évidente, toutefois elle est rarement effectuée dans la littérature, ce qui est source de malentendus.

En dehors du niveau de représentativité, il convient également de considérer le degré de représentativité d'un échantillon. Le *degré de représentativité* d_{rep} d'un échantillon pour un paramètre — au sens large (*e.g.*, variogramme, variance, distribution statistique, etc.) — peut se définir comme :

$$d_{rep} = \left[d\left(\theta^*, \theta\right) \right]^{-1} \tag{5.3}$$

avec θ la valeur du paramètre (populationnel ou superpopulationnel), θ^* la valeur observée dans l'échantillon, et $d\left(\cdot, \cdot\right)$ une distance appropriée.

Le problème est que le degré de représentativité d'un échantillon peut être élevé pour un paramètre et faible pour un autre. Dans le cadre des variables régionalisées, ceci nous amène à proposer les conjectures suivantes :

Conjecture 1 *Pour une population finie de taille N il n'existe pas d'échantillon de taille $n < N$ tel que d_{rep} soit optimal simultanément pour tous les paramètres.*

Conjecture 2 *Pour une population finie de taille N et un échantillon de taille $n < N$, $d_{rep} \to \infty$ lorsque $n \to N$, mais la forme de la convergence dépend à la fois du paramètre, de la structure spatiale de la population, et du mode d'échantillonnage.*

Conjecture 3 *Pour une superpopulation ergodique, le degré de représentativité de toutes les populations tend vers l'infini simultanément pour tous les paramètres superpopulationnels lorsque $D \to \infty$.*

5.4 Dépendance vs. indépendance

Il existe un lien très fort entre l'inférence statistique et les notions de dépendance et d'indépendance statistiques (ou stochastiques). Or, la lecture de la littérature consacrée aux données spatiales suggère que les notions de dépendance et d'indépendance sont particulièrement confuses, la dépendance spatiale des données étant même parfois confondue avec la dépendance statistique des variables aléatoires (*e.g.*, Brus & de Guijter 1997, p. 16).

Quelques citations permettent d'illustrer la confusion concernant l'indépendance des données spatiales :

> *"Thus, even though the observations are independent from the point of view of the probability of any particular geographic location to be sampled, their values at neighbouring points may not be independent from one another"* (Fortin *et al.* 1989)
>
> *"because the value at any one locality can be at least partly predicted by the values at neighboring points, these values are not stochastically independent from one another"* (Legendre 1993)
>
> *"Brus and De Gruijter (1993) showed that data which are dependent in the model-based approach still can be independent in the design-based approach and vice versa."* (Brus & de Gruijter 1994)
>
> "[with simple random sampling] *it was shown that measurements at locations within the variogram range could still be independent"* (Heuvelink 1997)
>
> *"Performing random sampling over spatial (or temporal) series does not insure that the sampling units are independent ..."* (Bellehumeur *et al.* 1997)

Cet état d'extrême confusion provient dans une large mesure d'un discours, d'une terminologie, et d'un formalisme insuffisamment explicites. L'objectif de cette section est de contribuer à éclaircir les notions fondamentales de dépendance et d'indépendance dans le domaine des données spatiales, ne serait-ce que pour fixer la terminologie que nous utilisons par la suite. Il convient en premier lieu de préciser le type de cadre mathématique employé selon que les valeurs observées sont modélisées par :

- une fonction déterministe,
- une variable aléatoire,
- un ensemble de variables aléatoires.

Ce n'est qu'une fois que le type de modélisation est précisé qu'il est possible de discuter de la signification des notions de dépendance et d'indépendance.

5.4.1 Modélisation par une fonction déterministe

Au stade du modèle primaire, *i.e.* lorsque les valeurs sont modélisées par une fonction déterministe (ou variable régionalisée), il n'y a aucun sens à parler de la dépendance ou de l'indépendance statistique (ou stochastique), parce qu'aucune source de stochasticité n'intervient.

Les valeurs $\{z(x_i) \mid i = 1, \ldots, n\}$ sont dites spatialement dépendantes si elles dépendent de la position des supports dans D, et spatialement indépendantes dans le cas contraire. La dépendance dont il est question constitue une propriété d'une variable numérique, liée à l'espace, strictement descriptive, et synonyme d'autocorrélation spatiale (Gatrell 1979, Cliff & Ord 1981).

Il est évident que le fait de constituer un échantillon de supports par EAS ne modifie en rien la dépendance spatiale de la population sous-jacente, puisqu'il s'agit ici de la variation spatiale du phénomène lui-même et pas d'une propriété d'une procédure statistique.

5.4.2 Modélisation par une variable aléatoire

La modélisation statistique classique considère que les valeurs observées $\{z_i \mid i = 1, \ldots, n\}$ constituent un échantillon de n réalisations d'une variable aléatoire Z. La variable aléatoire Z est un modèle statistique caractérisé par une loi de probabilité qui rend compte des probabilités des événements $z_\alpha < Z < z_\beta$ (loi continue) ou $Z = z$ (loi discrète), avec z, z_α et z_β des réels. Le nombre d'événements possibles — *i.e.* de probabilité non nulle — est infini dans le cas d'un loi continue, mais éventuellement fini dans le cas d'une loi discrète.

Il est plus facile de comprendre la modélisation des données en examinant d'abord son contre-pied, *i.e.* la simulation de données. Considérons par exemple une variable aléatoire $Z \sim \mathcal{N}(\mu, \sigma^2)$: Z représente une infinité de valeurs possibles, distribuées selon une loi normale, de moyenne μ et de variance σ^2. Il est possible de tirer n valeurs $\{z_i \mid i = 1, \ldots, n\}$ en échantillonnant l'infinité de valeurs possibles que peut prendre Z. Selon la façon dont l'échantillon est généré, les valeurs $\{z_i \mid i = 1, \ldots, n\}$ seront dites indépendantes ou dépendantes. Ainsi, si l'échantillonnage est aléatoire simple, alors les valeurs $\{z_i \mid i = 1, \ldots, n\}$ sont indépendantes. Si les valeurs sont échantillonnées de sorte que la séquence $(z_1, \ldots, z_n)$ présente une structure d'autocorrélation, alors les valeurs sont dites dépendantes. A ce stade, il est licite de parler de dépendance des valeurs, mais il n'y a aucun sens à parler de dépendance spatiale puisque nous n'avons pas encore fait intervenir l'espace. En considérant à présent l'espace, il convient d'examiner comment il est possible d'associer l'ensemble $\{z_j \mid j = 1, \ldots, n\}$ où j décrit l'ordre dans lequel les valeurs ont été générées, à l'ensemble $\{x_i \mid i = 1, \ldots, n\}$ où i décrit l'ordre spatial des supports, pour former les données spatiales $\{z(x_i) \mid i = 1, \ldots, n\}$. Pour simplifier, considérons que les supports sont des points sur une droite orientée D (*e.g.*, la flèche du temps). Il y a deux façons de générer les valeurs issues de $\mathcal{N}(\mu, \sigma^2)$:

- indépendamment les unes des autres, par tirage aléatoire simple,
- de façon non indépendante, par tirage autocorrélé.

En ce qui concerne l'affectation des valeurs $\{z_j \mid j = 1, \ldots, n\}$ aux points sur D $\{x_i \mid i = 1, \ldots, n\}$, nous proposons d'affecter les valeurs les unes à la suite des autres selon la séquence $j = 1, \ldots, n$:

- en choisissant chaque point au hasard et sans remise dans la séquence $i = 1, \ldots, n$,
- en suivant l'ordre des points sur D (séquence $i = 1, \ldots, n$),
- en suivant un certain ordre Θ, défini sur la séquence $i = 1, \ldots, n$ en fonction des valeurs générées, afin de contrecarrer les effets du tirage aléatoire des valeurs, indépendant ou dépendant.

Le Tableau 5.1 montre quel est le résultat attendu, en termes de dépendance et d'indépendance spatiales, lorsque l'on croise les deux types de tirages et les trois types d'affectations. L'affectation au hasard conduit de façon la plus probable à l'indépendance spatiale, du moins si le degré de dépendance des valeurs générées n'est pas extrême. L'affectation dans l'ordre de D implique une structure de dépendance spatiale identique à celle du tirage des valeurs puisque la valeur z_i est affectée au point x_i pour $i = 1, \ldots, n$. Enfin, en choisissant de façon *ad hoc* un certain ordre d'affectation Θ, il est possible d'obtenir la dépendance spatiale à partir de valeurs générées de façon indépendante, et l'indépendance spatiale à partir de valeurs obtenues par des tirages non indépendants.

En revenant au point de vue de la modélisation statistique classique, il apparaît évident d'après ce qui précède que des données spatialement dépendantes seront essentiellement vues comme des données tirées de façon non indépendante d'une VA Z. Autrement dit, on ne peut pas considérer que les valeurs $\{z_i \mid i = 1, \ldots, n\}$ constituent un échantillon aléatoire simple de Z, ce qui a évidemment des conséquences importantes en ce qui concerne le calcul d'intervalle de confiance ou l'application des tests statistiques. Ainsi, la modélisation statistique classique des valeurs observées $\{z(x_i) \mid i = 1, \ldots, n\}$ que nous venons de décrire n'exploite pas le caractère régionalisé de la variable $z(x)$ et l'autocorrélation spatiale est vue avant tout comme un paramètre de nuisance (*e.g.*, Dutilleul *et al.* 1993), *i.e.* un paramètre dont on doit tenir compte pour la validité de la procédure statistique, mais qui n'est pas de premier intérêt pour le scientifique (Lindsay 1985).

	Tirages dans $\mathcal{N}(\mu, \sigma^2)$	
Affectation	indépendants	non indépendants
au hasard	indépendance spatiale	indépendance spatiale
dans le sens de D	indépendance spatiale	dépendance spatiale
dans l'ordre Θ	dépendance spatiale	indépendance spatiale

Tableau 5.1: Résultats, en termes de dépendance et d'indépendance spatiales, de l'affectation des valeurs tirées d'une loi normale $\mathcal{N}(\mu, \sigma^2)$ à des points situés sur une droite orientée D (détails dans le texte).

5.4.3 Modélisation par un ensemble de variables aléatoires

Contrairement à l'approche statistique classique qui consiste à modéliser les valeurs observées $\{z(x_i) \mid i = 1, \ldots, n\}$ comme des réalisations d'une seule VA, le point de vue de l'approche géostatistique consiste à voir chaque valeur $z(x_i)$ comme la réalisation d'une VA $Z_\xi(x_i)$ et à modéliser statistiquement la dépendance spatiale par la covariance entre les VA $\{Z_\xi(x_i), Z_\xi(x_j)\}$ pour $i \neq j = 1, \ldots, n$ (Matérn 1960, Matheron 1965). La dépendance spatiale d'un couple de valeurs $\{z(x_i), z(x_j)\}$ est traduite dans le modèle par $\mathrm{Cov}_\xi[Z_\xi(x_i), Z_\xi(x_j)] \neq 0$, et les VA $Z_\xi(x_i)$ et $Z_\xi(x_j)$ sont statistiquement dépendantes. Dans le cas limite de l'indépendance spatiale (effet de pépite pur), les VA du modèle sont statistiquement indépendantes et $\mathrm{Cov}_\xi[Z_\xi(x_i), Z_\xi(x_j)] = 0$ pour $i \neq j = 1, \ldots, n$. Ainsi, dans le contexte *model-based*, la dépendance spatiale des données est traduite par un modèle statistique composé d'un ensemble de VA statistiquement dépendantes. Dans ce cadre, le fait que le motif d'échantillonnage résulte d'un EAS ne peut pas modifier la dépendance statistique des VA, simplement parce que le mode de sélection des supports n'intervient pas dans leur définition.

En se plaçant dans le cadre *design-based* et en considérant l'EAS, l'échantillon est constitué en tirant de façon indépendante un ensemble de n points $\{x_i\}$ selon la séquence $i = 1, \ldots, n$, ce qui conduit aux valeurs observées $\{z(x_i) \mid \ i = 1, \ldots, n\}$. Cette procédure d'échantillonnage est susceptible d'être répétée de sorte que l'ensemble des valeurs $z(x_i)$ pour i fixé (valeurs associées au $i^{ème}$ point tiré au hasard) peut être modélisé par une variable $Z_p(x_i)$. Comme le tirage du $i^{ème}$ point est un événement indépendant des tirages des autres points, il s'en suit que les variables $\{Z_p(x_i) \mid i = 1, \ldots, n\}$ sont statistiquement indépendantes. Ainsi, dans le contexte *design-based*, la sélection aléatoire des supports garantit l'indépendance statistique des VA induites par la réplication du dispositif, que la

VR soit spatialement autocorrélée ou pas (de Gruijter & ter Braak 1990, Brus & de Gruijter 1993). Ce qui rend l'échantillon aléatoire c'est le processus aléatoire sous-jacent et pas la composition de l'échantillon réellement construit (Hedayat & Sinha 1991). En conséquence, un échantillon obtenu par EAS peut être dit indépendant — par rapport à tous les autres échantillons qui peuvent être construits de cette façon — bien que les valeurs observées $\{z(x_i) \mid \ i = 1, \ldots, n\}$ ne soient pas nécessairement spatialement indépendantes.

Dans le cadre *design-based*, les variables aléatoires $Z_p(x_i)$ n'ont évidemment rien à voir avec celles définies dans l'approche *model-based* : les variables $Z_p(x_i)$ sont aléatoires du fait d'un dispositif d'échantillonnage probabiliste tandis que les variables $Z_\xi(x_i)$ sont aléatoires du fait d'un modèle probabiliste. Ainsi, l'indépendance statistique des $Z_p(x_i)$ n'est en rien contradictoire avec la dépendance statistique des $Z_\xi(x_i)$, simplement parce que ces VA correspondent à deux sources de stochasticité différentes :

- randomisation sur les supports dans le cas *design-based*,
- randomisation sur les valeurs dans le cas *model-based*.

Il est cependant possible de combiner les deux sources de stochasticité, ce qui revient à mixer les approches *design-based* et *model-based*. Dans ce cas, les propriétés des VA obtenues rendent compte à la fois du type de dispositif d'échantillonnage utilisé et du type de variation spatiale du phénomène étudié. Cette approche est surtout utilisée afin de comparer les performances de plusieurs dispositifs d'échantillonnage, en moyenne pour toutes les réalisations d'un modèle de superpopulation (Cochran 1946, Quenouille 1949, Hedayat & Sinha 1991, Brus & de Gruijter 1997).

5.5 Efficacité de l'échantillonnage

L'*efficacité de l'échantillonnage* est définie comme l'inverse de la variance d'échantillonnage, de sorte que l'échantillonnage le plus efficace est celui qui conduit à l'estimation la plus précise (Scherrer 1983). Considérons par exemple l'échantillonnage en vue d'estimer la moyenne globale z_D définie comme l'intégrale d'espace :

$$z_D = \frac{1}{[D]} \int_D z(x)\, dx \tag{5.4}$$

D'un point de vue statistique, il est souhaitable de choisir un dispositif d'échantillonnage le plus efficace possible, *i.e.* celui qui minimise la variance de l'erreur d'estimation de z_D pour un estimateur donné, par exemple la moyenne arithmétique :

$$\bar{z} = \frac{1}{n} \sum_{i=1}^{n} z(x_i) \tag{5.5}$$

Dans la mesure où la structuration spatiale des phénomènes écologiques constitue la règle plutôt que l'exception, l'étude de l'efficacité de l'échantillonnage sur une base strictement géométrique — dans le cadre de la géométrie probabiliste — nous semble inadaptée, parce qu'elle suppose l'absence de structure spatiale (*e.g.*, Hoel 1943). En conséquence, nous faisons référence à la géostatistique afin de tenir compte des caractéristiques spatiales du phénomène étudié.

En pratique, plusieurs situations sont à envisager selon la stabilité du phénomène, la connaissance de la population, et le type de comparaison envisagé. Un premier prérequis est la stabilité de la structure d'autocorrélation spatiale et la connaissance, au moins approximative, du variogramme. La variance d'erreur d'estimation géostatistique classique σ_E^2 n'étant pas conditionnée par les valeurs de $z(\cdot)$, il est possible de calculer l'efficacité statistique de plusieurs motifs d'échantillonnage particuliers en randomisant les populations, *i.e.* en considérant l'infinité des réalisations d'une FA. La comparaison des performances de plusieurs dispositifs d'échantillonnage implique également une randomisation sur les motifs d'échantillonnage.

Une forme plus stricte de la stabilité des structures spatiales est la stabilité de l'implantation des valeurs de $z(\cdot)$ dans D. La prise en compte de cette stabilité dans la comparaison de l'efficacité des motifs d'échantillonnage ou des dispositifs d'échantillonnage implique évidemment l'existence d'un échantillon préalable. L'efficacité doit cette fois tenir compte à la fois du variogramme et des valeurs observées de $z(\cdot)$.

Trois options sont envisageables selon que la randomisation concerne le motif d'échantillonnage, les valeurs associées à la population spatiale, ou les deux à la fois. Lorsque la randomisation concerne la population, il est nécessaire d'utiliser la variance d'erreur d'estimation conditionnée σ_C^2 et de recourir à des simulations conditionnelles (Section 4.6.5). Dans ce qui suit, nous supposons que le variogramme de $z(\cdot)$ est isotrope et de forme connue, et nous utilisons la variance d'erreur d'estimation non conditionnée σ_E^2.

5.5.1 Efficacité des motifs d'échantillonnage

Dans les expériences de terrain il est souvent possible d'appliquer la randomisation des unités expérimentales afin de ne pas avoir à spécifier un modèle d'autocorrélation spatiale réaliste. En revanche, dans les applications de l'échantillonnage du plan, il est souvent nécessaire, pour des raisons pratiques, d'utiliser un échantillonnage systématique (Dalenius *et al.* 1960) et de modéliser la variation spatiale par un modèle stochastique, par exemple une FAST-2 (Matérn 1960).

Lorsque le motif d'échantillonnage spatial est fixé, l'efficacité est calculée en moyenne pour toutes les réalisations du modèle comme $\mathrm{Var}_\xi[Z_D^* - Z_D]$ en utilisant σ_E^2. Dans ce cadre, il est possible de comparer différents motifs d'échantillonnage systématique, formés par les sommets des polygones congruents d'une tessellation régulière dans $\mathbb{R}^2$, *i.e.* par les sommets d'hexagones, de carrés ou de triangles équilatéraux (Section 2.1.2.3, p. 15).

En considérant un modèle d'autocorrélation spatiale circulaire[6], il est possible de formuler le problème du motif d'échantillonnage régulier le plus efficace en termes de couverture du plan par des cercles d'intersections mutuelles minimales, autrement dit, en termes strictement géométriques (Dalenius *et al.* 1960). Pour le modèle circulaire, les auteurs montrent qu'en général le motif à base de triangles équilatéraux n'est pas optimal, mais que les différences avec les autres motifs sont plutôt faibles. Par ailleurs, dans le cas d'un modèle d'autocorrélation spatiale exponentiel, Matérn (1960) indique que le motif à base de triangles équilatéraux conduit à une précision plus grande que les autres motifs réguliers. En conséquence, Dalenius *et al.* (1960) concluent qu'il n'existe pas

[6] La genèse du modèle circulaire est similaire à celle du modèle sphérique mais s'effectue dans un espace bidimensionnel au lieu d'un espace tridimensionnel (*cf.* McBratney & Webster 1986).

de motif régulier qui serait optimal pour toutes les fonctions d'autocorrélation convexes simultanément. Pour un variogramme particulier, il suffit en pratique de calculer σ_E^2 pour tous les motifs d'échantillonnage envisagés afin de retenir le plus efficace.

5.5.2 Efficacité des dispositifs d'échantillonnage

L'utilisation d'un modèle de superpopulation afin de comparer l'efficacité de différents dispositifs d'échantillonnage pour un structure d'autocorrélation donnée est une approche largement utilisée dans le cas d'un espace monodimensionnel (Cochran 1946, Karakostas & Wynn 1989, Karakostas 1990, Papageorgiou & Karakostas 1998), généralisable à un espace bidimensionnel (Quenouille 1949). Cette comparaison implique la randomisation du motif d'échantillonnage (pour une population donnée) et de la population (pour une superpopulation donnée). Dans ce cas, l'efficacité est calculée en moyenne sur tous les échantillons que peut générer un dispositif donné, pour toutes les réalisations d'un modèle, soit formellement, comme $\mathrm{E}_\xi \left[\mathrm{Var}_p \left[Z_D^* - Z_D\right]\right]$.

A noter que si l'on considère les dispositifs d'échantillonnage *non informatifs*[7], alors E_ξ et E_p sont interchangeables (Cassel *et al.* 1977, p. 109, Särndal 1978). Ainsi, il est formellement équivalent d'écrire $\mathrm{E}_p \left[\mathrm{Var}_\xi \left[Z_D^* - Z_D\right]\right]$ ou $\mathrm{E}_\xi \left[\mathrm{Var}_p \left[Z_D^* - Z_D\right]\right]$ puisque les dispositifs que nous considérons sont tous de type non informatif, et puisque $\mathrm{Var}_p \left[Z\right] = \mathrm{E}_p \left[\left(Z - \mathrm{E}_p \left[Z\right]\right)^2\right]$ et $\mathrm{Var}_\xi \left[Z\right] = \mathrm{E}_\xi \left[\left(Z - \mathrm{E}_\xi \left[Z\right]\right)^2\right]$ (Brus & de Gruijter 1997). Par la suite nous utilisons l'une ou l'autre de ces expressions selon le contexte. Nous considérons naturellement que la portée de l'autocorrélation n'est pas négligeable : si tel n'était pas le cas, en moyenne les différents dispositifs seraient tous équivalents (Matérn 1960).

Deux types d'études peuvent être envisagés selon qu'il s'agit de choisir entre plusieurs dispositifs pour une taille d'échantillon n fixée, ou entre plusieurs tailles d'échantillons pour un dispositif donné.

5.5.2.1 Choix du dispositif pour une taille d'échantillon fixée

L'efficacité des dispositifs d'échantillonnage classiques tels que l'échantillonnage aléatoire simple (EAS), l'échantillonnage stratifié (STR), et l'échantillonnage systématique (ES) a été étudiée, notamment par Cochran (1946) et Madow (1946, 1949) dans $\mathbb{R}$, et par Zubrzycki (1958), Matérn (1960), Matheron (1965), Ripley (1981) et Iachan (1985) dans $\mathbb{R}^2$.

Cochran (1946) compare l'efficacité des trois types d'échantillonnage unidimensionnel et conclut que le STR est toujours au moins aussi précis que l'EAS, et que son efficacité relative est une fonction monotone croissante de n, *i.e.* lorsque la taille des strates diminue. En revanche, il n'existe pas de résultat général concernant l'efficacité de l'ES. Ainsi, pour certaines populations, l'ES est plus efficace que le STR pour un certain taux d'échantillonnage, et moins efficace que l'EAS pour un autre taux d'échantillonnage (Cochran 1946). L'efficacité augmente de façon erratique en ce qui concerne l'ES qui peut être plus efficace ou moins efficace que le STR selon l'intensité d'échantillonnage (Madow 1946). Cependant, si la concavité de la fonction de corrélation est tournée vers le haut — si le

[7] Un dispositif d'échantillonnage est dit *non informatif* lorsque les probabilités d'inclusion sont indépendantes des valeurs (Cassel *et al.* 1977).

modèle de variogramme est convexe, comme dans le cas du modèle exponentiel — alors l'ES est en moyenne plus précis que le STR pour n'importe quelle taille d'échantillon (Cochran 1946). En outre, dans les mêmes conditions que précédemment, Madow (1953) a montré que le motif centré de l'ES est plus efficace que l'ES à origine aléatoire. Madow (1953) précise qu'il est facile d'étendre ce résultat à l'échantillonnage bidimensionnel, ce qui s'avère parfaitement cohérent avec l'étude de cas exposée dans Aubry & Debouzie (1999a). Enfin, Madow (1949) explique que dans le cas d'un ES disposé selon une grille, la présence d'une tendance selon les lignes ou les colonnes rend le dispositif assez inefficace (Section 6.3.1.4, p. 158).

Pour simplifier, nous considérons que toutes les strates de l'échantillonnage stratifié sont identiques. Pour une géométrie de strate et un taux d'échantillonnage donnés, l'échantillonnage stratifié à un élément par strate[8] (STR) donne la variance la plus faible (Matérn 1960, Zinger 1964, Ripley 1981). Pour un nombre de points par strate fixé, il est possible de discuter de l'impact de la géométrie de la strate sur l'efficacité du dispositif. En fixant l'aire de la strate, l'efficacité est d'aurant plus grande que la strate est compacte (Ripley 1981). En conséquence, il est possible de classer les géométries par ordre d'efficacité décroissante du dispositif correspondant (Matérn 1960) : cercle, hexagone régulier, carré et triangle équilatéral. En imposant que la stratification dans $\mathbb{R}^2$ soit une tesselation régulière (Section 2.1.2.3, p. 15), l'hexagone est optimal au sens où il minimise la frontière de la strate. Matérn (1960) conjecture que l'hexagone est optimal pour toutes les fonctions de corrélation isotropes non croissantes, *i.e.* pour tous les modèles de variogrammes dont la croissance est monotone. Cependant, bien qu'il soit intéressant sur le plan théorique, le problème de la géométrie optimale de la strate n'a pas de conséquence pratique immédiate parce qu'il n'y a pas de différence significative avec le carré (Matérn 1960, p. 74).

Afin de comparer les trois dispositifs fondamentaux, EAS, STR et ES, Zubrzycki (1958) et Ripley (1981, pp. 22-27) se placent dans le cadre des FAST-2 isotropes et considèrent la corrélation $r(h)$ ou la covariance $C(h)$ tandis que Matheron (1965) se place dans le cadre plus général des FAI-0 isotropes et utilise le variogramme $\gamma(h)$. En fait, comme c'est le cas pour de nombreux résultats concernant les variances en géostatistique linéaire, les formules obtenues avec la covariance (ou la corrélation) peuvent être généralisées au cas des FAI-0 et s'exprimer en termes de variogramme (Journel & Huijbregts 1978, p. 36). Matheron (1965) donne les variances pour l'EAS et le STR, mais pas pour l'ES. En revanche, Zubrzycki (1958) et Ripley (1981) fournissent également la variance d'erreur d'estimation dans le cas de l'ES. Pour un échantillon de taille n prélevé dans le domaine D en suivant le dispositif d, la variance d'erreur d'estimation σ_d^2 s'écrit, respectivement pour l'EAS, le STR de strate v et l'ES :

$$\sigma_{EAS}^2 = \frac{1}{n}\bar{\gamma}(D,D) \tag{5.6}$$

$$\sigma_{STR}^2 = \frac{1}{n}\bar{\gamma}(v,v) \tag{5.7}$$

$$\sigma_{ES}^2 = \bar{\gamma}(D,D) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma(x_i,x_j) \tag{5.8}$$

[8] En écologie, l'échantillonnage stratifié à un élément par strate semble assez rarement utilisé (pour un exemple récent, voir Legendre *et al.* 1997).

avec

$$\bar{\gamma}(b,b) = \frac{1}{[b]^2} \int_b \int_b \gamma(x,y)\, dxdy \tag{5.9}$$

où $\gamma(x,y)$ est le variogramme calculé pour la distance entre deux points $x, y \in \mathbb{R}^2$ décrivant indépendamment le domaine b d'aire $[b]$.

Zubrzycki (1958) et Matheron (1965) montrent ainsi que le STR est toujours au moins aussi efficace que l'EAS puisque $\bar{\gamma}(D,D) \geq \bar{\gamma}(v,v)$, ce qui est conforme aux résultats obtenus dans $\mathbb{R}$ (Cochran 1946). Zubrzycki (1958) compare également le STR et l'ES dans le cas d'une corrélation de forme exponentielle et montre que le STR peut être plus efficace que l'ES selon la forme de la strate et la taille du domaine par rapport à la portée. Dans le cas d'un domaine D circulaire, et de strates également circulaires, il existe des domaines pour lesquels le STR est plus efficace que l'ES, mais si la portée est suffisamment grande, alors l'ES est plus efficace que le STR, quel que soit le domaine circulaire. Dans le cas d'une FAST-2 isotrope, de corrélation exponentielle, et des mailles hexagonales ou carrées, le problème de savoir si l'ES est toujours plus efficace que le STR, reste ouvert (Zubrzycki 1958). D'une façon générale, l'écart entre les variances de l'ES et du STR peut s'écrire, en termes de variogramme :

$$\sigma^2_{ES} - \sigma^2_{STR} = \frac{1}{n^2} \sum_{\substack{i,j \\ i \neq j}} \{\bar{\gamma}(v_i, v_j) - \gamma(x_i, x_j)\} \tag{5.10}$$

avec

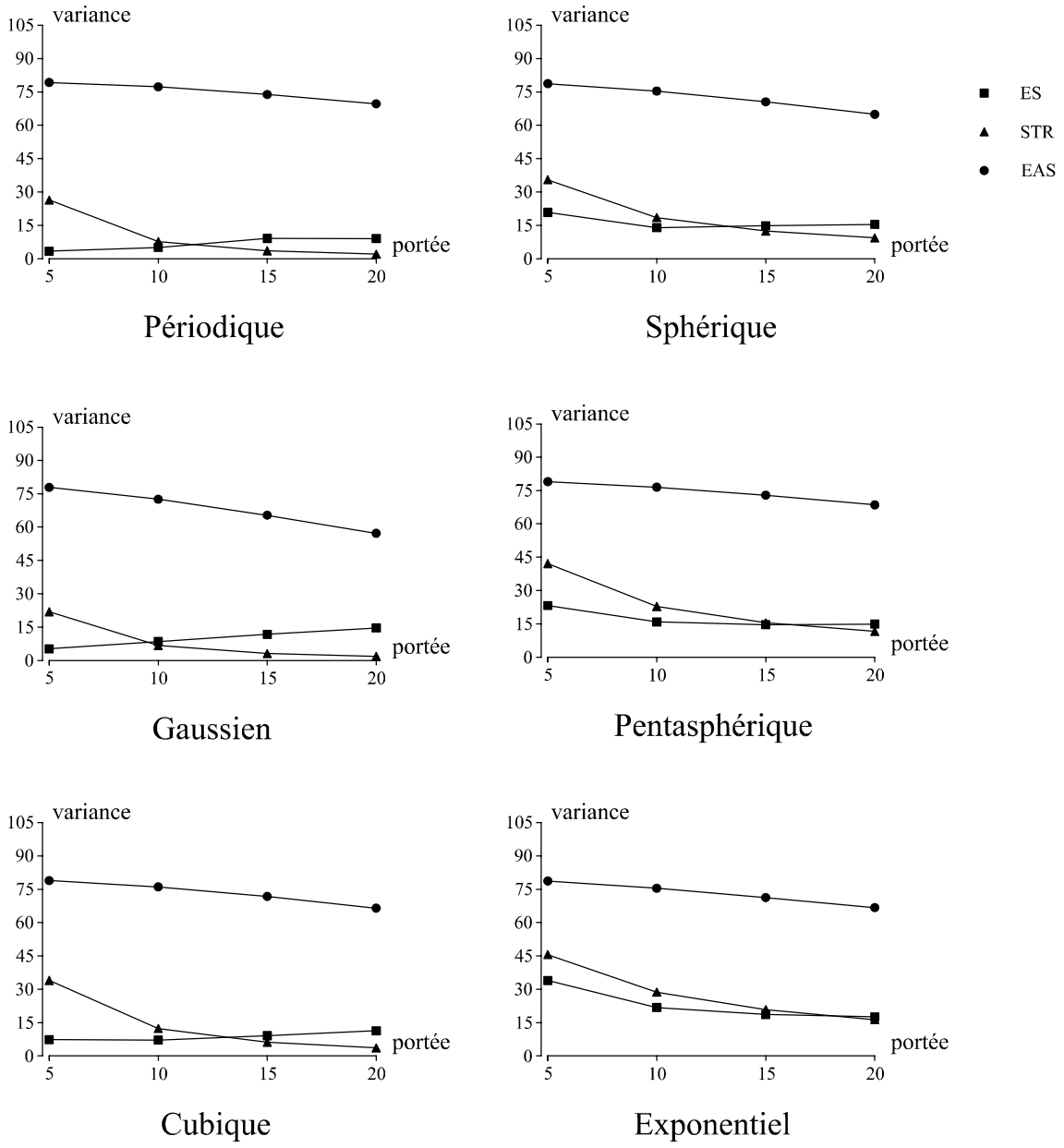
$$\bar{\gamma}(v_i, v_j) = \frac{1}{[v]^2} \int_{v_i} \int_{v_j} \gamma(x,y)\, dxdy \tag{5.11}$$

où $x, y \in \mathbb{R}^2$ sont deux points décrivant indépendamment les strates v_i et v_j de même aire $[v]$, centrées respectivement en x_i et x_j. Ainsi, si $\gamma(x_i, x_j) > \bar{\gamma}(v_i, v_j)$ pour tout $i \neq j$, alors l'ES est plus efficace que le STR (Zubrzycki 1958).

Afin d'illustrer l'efficacité relative des trois dispositifs fondamentaux, nous considérons un domaine D carré de côté $L = 30$ unités, discrétisé par une grille de 10×10 strates carrées, et une VR définie en tout point $x \in D$. L'échantillon spatial est composé de $n = 100$ points prélevés par EAS, ES ou par STR (un élément par strate). Pour juger de l'impact de la forme de l'autocorrélation spatiale, six modèles de variogrammes pour FAST-2 sont considérés : périodique, gaussien, cubique, pentasphérique, sphérique et exponentiel (Annexe F). Les modèles sont tous paramétrés de la même façon avec une pépite $c_0 = 1$, un seuil $c_0 + c = 8000$ et une portée $a \in \{5, 10, 15, 20\}$. Les variances sont calculées pour les trois dispositifs fondamentaux en utilisant les expressions (5.6), (5.8) et (5.7). Le calcul de $\bar{\gamma}(b,b)$ est effectué par intégration de Monte-Carlo en discrétisant b grâce à un échantillon aléatoire stratifié par une grille 55×55, à deux points distincts par strate.

Les résultats montrent que l'EAS est toujours beaucoup moins efficace que l'ES ou le STR (Fig. 5.3). Pour l'EAS ou le STR, l'efficacité croît lorsque la portée augmente, mais cette croissance est plus forte pour le STR que pour l'EAS (Fig. 5.3). En revanche, selon le modèle de variogramme, l'efficacité de l'ES peut éventuellement décroître lorsque la portée augmente de sorte que l'efficacité relative de l'ES et du STR dépend à la fois du type de variogramme et de la portée (Fig. 5.3). Lorsque la portée est faible vis-à-vis du

domaine D, l'ES s'avère toujours plus efficace que le STR, mais le rapport s'inverse lorsque la portée augmente, pour des portées critiques qui dépendent du modèle de variogramme (Fig. 5.3). Dans le cas d'une autocorrélation exponentielle, le STR devient légèrement plus efficace que l'ES lorsque la portée est grande vis-à-vis du domaine D, ce qui constitue un élément de réponse au problème mentionné par Zubrzycki (1958) (p. 122).



Figure˜5.3: Efficacité relative des trois dispositifs d'échantillonnage fondamentaux EAS, STR et ES, en fonction du modèle de variogramme et de la portée (détails dans le texte).

Il peut s'avérer nécessaire d'évaluer l'efficacité d'autres dispositifs que l'EAS, l'ES ou le STR, ce qui nécessite d'établir des expressions analogues à (5.6), (5.8) ou (5.7), ou de recourir à une approche de Monte-Carlo (Domburg *et al.* 1994). En effet, pour un domaine D et un modèle de variogramme isotrope $\gamma(h)$, et quel que soit le dispositif d'échantillonnage probabiliste envisagé, il est toujours possible de calculer la variance

d'erreur d'estimation grâce à une méthode de Monte-Carlo. Si, formellement, il est équivalent de considérer $\mathrm{E}_\xi\left[\mathrm{Var}_p\left[Z_D^* - Z_D\right]\right]$ ou $\mathrm{E}_p\left[\mathrm{Var}_\xi\left[Z_D^* - Z_D\right]\right]$ (Domburg *et al.* 1994), en revanche, dans le cadre de l'approche de Monte-Carlo, chaque expression correspond à une implémentation particulière. En effet, l'expression $\mathrm{E}_\xi\left[\mathrm{Var}_p\left[Z_D^* - Z_D\right]\right]$ consiste, d'un point de vue opératoire, à simuler une réalisation non conditionnelle de la FA caractérisée par $\gamma\left(h\right)$, puis à répliquer le dispositif d'échantillonnage de nombreuses fois afin de tendre asymptotiquement vers la variance d'échantillonnage $\mathrm{Var}_p\left[Z_D^* - Z_D\right]$. Cette opération est ensuite répétée en simulant de nombreuses réalisations afin de tendre asymptotiquement vers l'espérance dans le modèle $\mathrm{E}_\xi\left[\cdot\right]$ (Domburg *et al.* 1994). En revanche, l'expression $\mathrm{E}_p\left[\mathrm{Var}_\xi\left[Z_D^* - Z_D\right]\right]$ consiste, toujours d'un point de vue opératoire, à calculer numériquement la variance dans le modèle $\mathrm{Var}_\xi\left[Z_D^* - Z_D\right] \equiv \sigma_E^2$ pour le motif d'échantillonnage courant, puis à répéter ce calcul pour de nombreux motifs obtenus en appliquant le même dispositif d'échantillonnage, et cela afin de tendre asymptotiquement vers l'espérance d'échantillonnage $\mathrm{E}_p\left[\cdot\right]$.

Il faut s'attendre à ce que ces deux méthodes donnent des résultats numériques légèrement différents en fonction des détails de l'implémentation (nombre et précision des simulations, nombre de réplications, précision du calcul de σ_E^2). La première méthode utilisant la simulation stochastique est cependant plus difficile à mettre en oeuvre que la seconde qui utilise essentiellement une méthode d'intégration de Monte-Carlo. En effet, la première méthode nécessite d'affecter les valeurs d'une même réalisation $z^{(\ell)}\left(\cdot\right)$ à de nombreux semis de points différents, ce qui ne peut pas s'effectuer avec toutes les méthodes de simulation puisque certaines d'entre elles ne peuvent affecter les valeurs d'une réalisation $z^{(\ell)}\left(\cdot\right)$ que sur un seul semis de points, *e.g.* la méthode utilisant la factorisation d'une matrice de covariance (Section 4.3.3). Toutefois, le recours à la méthode des bandes tournantes doit permettre d'implémenter la première approche, en assurant toutefois une précision suffisante des simulations unidimensionnelles (Section 4.3.2). L'implémentation de la seconde méthode de Monte-Carlo ne pose quant à elle aucune difficulté dès lors que le calcul numérique de σ_E^2 est maîtrisé.

Afin d'illustrer l'équivalence du calcul des expressions (5.6), (5.7) ou (5.8), et de la seconde méthode de Monte-Carlo, nous considérons par exemple le modèle de variogramme $\mathrm{Expo}\left(1, 7999, 10\right)$. Pour chaque dispositif fondamental EAS, ES et STR, les résultats de la seconde méthode de Monte-Carlo sont obtenus en calculant la variance σ_E^2 par intégration de Monte-Carlo (Section 4.6.3), pour 100 motifs d'échantillonnage générés aléatoirement. Les résultats du Tableau 5.2 permettent de vérifier l'équivalence des deux procedures, les différences observées étant dues à la nature numérique des calculs.

	σ_d^2	$\mathrm{E}_p\left[\sigma_E^2\right]$
EAS	75.40	73.70
STR	28.73	28.69
ES	21.83	21.99

Tableau 5.2: Equivalence du calcul de la variance d'erreur d'estimation moyenne au moyen des expressions théoriques (σ_d^2) et de la seconde méthode de Monte-Carlo ($\mathrm{E}_p\left[\sigma_E^2\right]$), pour les dispositifs EAS, STR et ES, un domaine carré et un variogramme exponentiel.

L'efficacité de dispositifs obtenus en combinant les trois dispositifs fondamentaux a été étudiée notamment par Quenouille (1949), Bellhouse (1977) et Koop (1990). Quenouille (1949) montre que l'alignement des points d'échantillonnage tend à diminuer l'efficacité dans le cas des dispositifs de type EAS et STR. Cependant, dans le cas d'une autocorrélation négative, l'alignement peut augmenter l'efficacité dans le cas des dispositifs de type STR (Koop 1990). Dans le cas de l'ES, aucun résultat général ne peut être dégagé, et l'efficacité dépend de la fonction de covariance spatiale (Quenouille 1949, Koop 1990). En faisant référence à l'étude de Quenouille (1949), et en utilisant un modèle de corrélation spatiale exponentiel, Bellhouse (1977) considère trois classes de dispositifs selon que les unités sont :

1. alignées selon les deux axes de coordonnées,
2. alignées selon un axe et indépendantes selon l'autre,
3. indépendantes selon les deux axes.

Bellhouse (1977) montre que dans chacune des classes, le dispositif optimal est de type systématique. En outre, tout dispositif de la classe 1 est moins efficace que celui qui lui est associé dans la classe 2 et par conséquent, le dispositif optimal de la classe 1 est moins efficace que celui de la classe 2. De plus, le dispositif de type systématique de la classe 3 est plus efficace que celui de la classe 1. Conformément aux résultats de Quenouille (1949), l'échantillonnage de type systématique serait plus efficace en considérant indépendamment les abscisses et les ordonnées, tout du moins dans le cas d'un variogramme exponentiel. Par exemple, en cas d'autocorrélation spatiale positive, le motif d'échantillonnage systématique devrait être en quinconce afin de maximiser les distances entre points et donc minimiser la redondance d'information (Matérn 1960, *op. cit.* Scherrer 1983, p. 78).

5.5.2.2 Choix de la taille d'échantillon pour un dispositif fixé

Dans le cas de l'échantillonnage aléatoire simple, il est bien connu que la précision augmente avec la taille d'échantillon. Il faut se garder de généraliser ce résultat à tous les dispositifs car l'EAS constitue un cas très spécial parmi les dispositifs d'échantillonnage. Ainsi, contrairement à l'EAS, l'augmentation de la taille de l'échantillon systématique n'augmente pas nécessairement la précision (Hedayat & Sinha 1991). Dans $\mathbb{R}$, Cochran (1946) précise que le graphe de la variance de la moyenne pour l'ES ne présente pas la même régularité que celui de l'EAS.

Pour un dispositif quelconque d, il est toujours possible d'obtenir le graphe $\sigma_d^2(n)$ de la variance d'erreur d'estimation de la moyenne en fonction de la taille d'échantillon n en utilisant la seconde méthode de Monte-Carlo décrite dans la Section 5.5.2.1 (p. 124). En fixant une précision limite, il suffit ensuite de déterminer l'effectif correspondant sur le graphe $\sigma_d^2(n)$.

5.5.3 Interprétation des efficacités

L'efficacité de l'échantillonnage spatial est une propriété statistique définie pour un estimateur et concerne un type de motif ou de dispositif d'échantillonnage, et pas un échantillon en particulier. En pratique, n'importe quelle valeur estimée à partir d'un échantillon

sera probablement différente de la valeur réelle, et l'erreur d'estimation est évidemment imprévisible. C'est précisément la fonction de la statistique d'indiquer quelle peut être l'amplitude de cette erreur d'estimation (Madow & Madow 1944).

Considérons par exemple l'estimation de la moyenne globale z_D par un échantillonnage de type non informatif. Trois types d'efficacité peuvent être définis selon que la source de stochasticité concerne le motif d'échantillonnage, répliqué de façon aléatoire dans le domaine D ($\mathrm{Var}_p\,[Z_D^* - Z_D]$), la population, tirée aléatoirement d'un modèle de superpopulation ξ ($\mathrm{Var}_\xi\,[Z_D^* - Z_D]$), ou les deux à la fois ($\mathrm{E}_\xi\,[\mathrm{Var}_p\,[Z_D^* - Z_D]] \equiv \mathrm{E}_p\,[\mathrm{Var}_\xi\,[Z_D^* - Z_D]]$).

D'un point de vue pratique, on peut se demander comment faire le lien entre ces efficacités, calculées en moyenne pour plusieurs motifs d'échantillonnage et/ou pour une infinité de populations, et le résultat obtenu à partir d'un échantillon donné, *i.e.* dans le cas d'une population donnée échantillonnée selon un motif particulier. Pour simplifier, nous considérons le cas d'une VR spatialement continue définie sur des supports ponctuels — ce qui conduit à une population infinie — autocorrélée positivement, et modélisable par une FA à dérive constante.

5.5.3.1 Randomisation du motif d'échantillonnage

L'efficacité la plus simple à comprendre est celle qui est définie par randomisation du motif d'échantillonnage ($\mathrm{Var}_p\,[Z_D^* - Z_D]$). Intuitivement, pour une population présentant une variabilité spatiale qui n'est pas négligeable, mais ne présentant pas de tendance marquée, il faut s'attendre à ce que la variance d'échantillonnage soit d'autant plus grande que la randomisation du motif est importante. En outre, dans le cas général d'une autocorrélation spatiale positive, il est souhaitable que les points ne soient pas proches les uns des autres afin de minimiser la redondance d'information. Le cas extrême est évidemment celui de l'échantillonnage aléatoire simple dans lequel tous les points sont choisis aléatoirement. Par hasard, un motif obtenu par EAS peut être fortement agrégé, auquel cas l'erreur d'estimation risque d'être assez élevée en valeur absolue, ou au contraire couvrir assez régulièrement D, et l'estimation peut s'avérer excellente. Mais en moyenne, il ne faut pas s'attendre à une estimation précise, parce que la couverture de l'espace est inégale avec des zones sous-échantillonnées et d'autres suréchantillonnées (Maling 1989).

Un moyen d'éviter les agrégats de points dans D est de stratifier l'échantillonnage aléatoire par une grille, en tirant au hasard un seul point par strate. Intuitivement, on comprend que cette stratification *a priori* — qui ne nécessite aucune information concernant la population, ni la connaissance d'une variable auxiliaire qui serait corrélée à la variable étudiée — conduit à une estimation qui a des chances d'être souvent meilleure que dans le cas de l'EAS.

Enfin, il est encore possible de réduire davantage la randomisation en échantillonnant les points selon une grille dont seule l'origine sera choisie au hasard. Là encore, on peut espérer obtenir souvent une meilleure estimation que dans le cas de l'EAS, parce qu'en présence d'autocorrélation spatiale positive, l'ES impose une distance minimale entre les points échantillonnés et évite ainsi la collecte d'informations redondantes (Scherrer 1983, p. 78). La stratification et l'échantillonnage par grille peuvent donc être vus comme des techniques de réduction de la variance d'échantillonnage (de Gruijter & ter Braak 1992).

5.5.3.2 Randomisation de la population

Lorsque le motif d'échantillonnage est fixé, et en particulier s'il ne peut pas être répliqué pour une même population (cas de l'échantillonnage préférentiel), seule une efficacité par randomisation de la population elle-même peut être définie ($\mathrm{Var}_\xi\,[Z_D^* - Z_D]$).

Selon la stabilité du phénomène, il faut s'attendre à ce qu'un motif qui a conduit à une erreur d'estimation d'une certaine amplitude pour une population donnée, continue à se comporter ainsi pour d'autres populations similaires, à quelques fluctuations près. C'est précisément le rôle du modèle que de produire des populations qui fluctuent, tout en conservant des propriétés spatiales plus au moins précises. Dans le cas d'une structure très stable, les fluctuations doivent nécessairement être moindres que dans le cas d'une structure moins stable. Ceci justifie alors le recours à une variance conditionnelle du type σ_C^2 plutôt qu'à une variance non conditionnée par les valeurs telle que σ_E^2 (Section 4.6).

Le risque essentiel attaché à cette approche est qu'un motif peut être optimal sous un modèle précis et se révéler désastreux si le modèle ne s'applique plus : la question est de savoir à quel point la perte d'efficacité est grande lorsqu'on s'écarte des conditions décrites par le modèle. L'idéal serait d'obtenir des motifs efficaces et robustes pour une grande classe de modèles (Iachan 1984).

5.5.3.3 Double randomisation

La double randomisation permet d'étudier l'efficacité pour un certain type de dispositif d'échantillonnage et pour un certain type de fluctuation de la population. L'interprétation nécessite donc de se référer aux deux sections précédentes.

Il faut cependant noter que la précision $\mathrm{Var}_p\,[Z_D^* - Z_D]$ obtenue pour une population finie diffère nécessairement de celle obtenue en moyenne dans le cadre d'une superpopulation de modèle ξ ($\mathrm{E}_\xi\,[\mathrm{Var}_p\,[Z_D^* - Z_D]]$), parce que la structure d'autocorrélation spatiale de la population diffère de son espérance dans le modèle.

Si la taille du domaine d'étude devient grande par rapport à la portée de l'autocorrélation spatiale, les résultats pour des populations données vont coïncider avec les résultats en moyenne, du fait de l'ergodicité (Section 4.2.1.2).

5.6 Stratégies d'échantillonnage

Une *stratégie d'échantillonnage* est classiquement définie comme le couple formé par un dispositif d'échantillonnage (S_d, P_d) et un estimateur $\widehat{\theta}$ du paramètre populationnel θ (Cassel *et al.* 1977, Smith 1976, Särndal 1978).

Dans le contexte de l'échantillonnage spatial, nous désignons par *stratégie d'échantillonnage* le couple formé par le motif d'échantillonnage et l'estimateur, sans faire nécessairement intervenir la notion de dispositif d'échantillonnage, cette définition plus large permettant de considérer aussi bien l'échantillonnage probabiliste que l'échantillonnage préférentiel.

L'inférence d'un paramètre θ d'une population finie à partir d'un échantillon amène à distinguer classiquement trois catégories de problèmes (Cassel *et al.* 1977, p. 35*)* :

a. le choix d'un estimateur $\widehat{\theta}$ pour un dispositif donné (S_d, P_d),
b. le choix d'un dispositif (S_d, P_d) pour un estimateur donné $\widehat{\theta}$,
c. le choix d'une stratégie, *i.e.* le choix simultané d'un dispositif (S_d, P_d) et d'un estimateur $\widehat{\theta}$.

Quel que soit le type de choix considéré, il est évident que le but est de minimiser une variance d'erreur estimation $\mathrm{Var}\left[\widehat{\theta} - \theta\right]$ (Cassel *et al.* 1977). Dans ce contexte, le problème du choix d'une stratégie d'échantillonnage est d'abord un problème de choix d'un type de variance. En effet, la variance peut être reliée à différentes sources de stochasticité (de Gruijter & ter Braak 1990) :

1. le dispositif d'échantillonnage (randomisation du motif d'échantillonnage),
2. le modèle de variation spatiale (randomisation de la population),
3. les deux sources précédentes (double randomisation).

Dans le cas 1, la variance est $\mathrm{Var}_p\left[\widehat{\theta} - \theta\right]$, dans le cas 2 il s'agit de $\mathrm{Var}_\xi\left[\widehat{\theta} - \theta\right]$ et dans le cas 3 de $\mathrm{E}_\xi\left[\mathrm{Var}_p\left[\widehat{\theta} - \theta\right]\right]$.

Afin de minimiser $\mathrm{Var}_p\left[\widehat{\theta} - \theta\right]$, il est possible d'introduire de l'information dans le dispositif par stratification au moyen d'une variable auxiliaire (Cochran 1977, Särndal 1978, Hedayat & Sinha 1991). En fait, l'augmentation de la précision par stratification est une idée ancienne. En effet, la théorie de l'échantillonnage aléatoire et indépendant a été développée par Bernouilli il y a plus de 250 ans et la théorie qui mesure le gain d'efficacité en introduisant de l'information auxiliaire par stratification a été indiquée par Poisson un siècle plus tard (Hansen & Hurwitz 1943). Les développements récents dans le domaine de l'échantillonnage concernent le choix d'une stratégie guidée par une information auxiliaire incorporée dans un modèle plutôt que dans le dispositif, ce qui revient à considérer $\mathrm{Var}_\xi\left[\widehat{\theta} - \theta\right]$ plutôt que $\mathrm{Var}_p\left[\widehat{\theta} - \theta\right]$. Le recours à un modèle présente en effet l'intérêt de permettre de déterminer à la fois l'échantillonnage optimal et l'estimateur optimal (Hansen *et al.* 1983), et par conséquent, une stratégie d'échantillonnage optimale.

Les théoriciens de l'échantillonnage se répartissent habituellement dans deux camps selon qu'ils plaident en faveur d'une inférence *design-based* ou *model-based* (Royall 1970, Hansen *et al.* 1983, Royall 1983). Cependant, cette coupure semble graduellement tendre vers un compromis où l'on fait un bon usage des deux approches (Iachan 1984, Urquhart 1997). Dans ce contexte, le choix d'une stratégie d'échantillonnage peut être qualifié de *model-assisted* (Särndal *et al.* 1992, *op. cit.* Brus & de Gruijter 1992). L'approche *model-assisted* fait référence à la double randomisation, autrement dit, la variance à minimiser est $\mathrm{E}_\xi\left[\mathrm{Var}_p\left[\widehat{\theta} - \theta\right]\right]$.

Les considérations statistiques qui précèdent sont importantes parce que la définition d'une stratégie d'échantillonnage sur des bases scientifiques est une étape préliminaire difficile mais essentielle pour la recherche écologique (Legendre *et al.* 1989).

La difficulté rencontrée pour définir une stratégie d'échantillonnage spatial en écologie se traduit par la diversité des avis exprimés sur ce sujet dans la littérature. Par exemple, certains auteurs considèrent que l'échantillonnage doit maximiser la dépendance spatiale des données lorsque l'objectif est de décrire la variabilité spatiale, mais qu'elle doit au contraire la minimiser lorsqu'il s'agit d'estimer la densité moyenne (*e.g.*, Cardina *et al.* 1997). Il nous semble plutôt évident qu'il faut minimiser l'erreur d'estimation du paramètre envisagé, ce qui ne va pas nécessairement de pair avec la maximisation ou la minimisation de l'autocorrélation spatiale dans les données, mais requiert un échantillon le plus représentatif possible de la population sous-jacente. D'autres auteurs considèrent tout simplement qu'il n'existe pas de théorie de l'échantillonnage écologique (*e.g.*, Chessel 1992, p. 5).

Il convient de reconnaître qu'en écologie il n'existe pas de stratégie optimale dans l'absolu, *e.g.* indépendante du type de structure spatiale échantillonnée (Legendre *et al.* 1989). L'intérêt du recours aux modèles de superpopulations est précisément de permettre de comparer l'efficacité de différentes stratégies d'échantillonnage en fonction du type de structure spatiale. Les spécialistes en statistique mathématique effectuent ce type de comparaison de façon théorique, mais il est beaucoup plus facile de recourir à une approche de Monte-Carlo. En effet, avec la simulation numérique de modèles de superpopulations tels que les fonctions aléatoires (Section 4.3), couplée éventuellement à la réplication du dispositif d'échantillonnage (approche *model-assisted*), il est possible d'étudier les performances de n'importe quelle stratégie d'échantillonnage, pour des géométries de domaine et des structures d'autocorrélation spatiale particulières.

Ainsi, s'il n'existe pas de théorie de l'échantillonnage écologique (Chessel 1992), il existe néanmoins des outils qui autorisent des études au cas par cas, au moins dans une perspective univariée[9].

[9]Le développement d'une théorie de l'échantillonnage dans le cas d'une multivariable constitue un problème ouvert (Cassel *et al.* 1977), et *a fortiori* dans le cas d'une multivariable régionalisée.

Chapitre 6

Estimation spatiale

> *"I am unaware that there have been many attempts to formalize the notion of map accuracy, to quantify it, and to measure it* [...] *the peruser of a map is in a position similar to one who is given a point estimate without any notion of its standard error"* (Switzer 1971)
>
> *"Competent scientists do not believe their own models or theories, but rather treat them as convenient fictions."* (Vardeman 1987)

Une variable régionalisée $z(\cdot)$ définie sur un domaine D est généralement connue de façon partielle à travers un échantillon de supports $s = \{s_i \mid i = 1, \ldots, n\}$ répartis dans D. L'*estimation spatiale* consiste à pallier un manque d'information concernant $z(\cdot)$ du fait de l'échantillonnage. L'estimation spatiale concerne donc des valeurs objectives, *i.e.* qui existent en dehors de nos choix méthodologiques (Chauvet 1994, p. 61), par exemple l'intégrale d'espace :

$$z_D = \frac{1}{[D]} \int_D z(x)\, dx \tag{6.1}$$

avec x un point décrivant le domaine D d'aire $[D] = \int_D dx$. Dans ce sens, le terme *estimation spatiale* ne doit pas être confondu avec *estimation d'un paramètre statistique* (*e.g.*, l'espérance). En faisant référence à la géostatistique *model-based*, il convient souvent de parler de *prédiction* d'une variable aléatoire définie sur un certain support (éventuellement D lui-même).

Selon le point de vue adopté, l'estimation spatiale peut être une estimation statistique ou une prédiction. Par exemple, dans le cas où la VR $z(\cdot)$ est vue comme une population fixée définie sur un domaine D borné, échantillonnée au hasard (par exemple), l'estimation spatiale de z_D se confond avec l'estimation de l'espérance des moyennes d'échantillons $\mathrm{E}_p\left[\bar{Z}\right]$. En revanche, dans le cas où la VR $z(\cdot)$ est vue comme une réalisation quelconque d'une superpopulation $Z(\cdot)$, alors z_D est une réalisation de la VA Z_D, et l'estimation spatiale peut être vue comme la prédiction de Z_D à partir des données et du modèle $Z(\cdot)$, ou même, éventuellement, comme l'estimation de $\mathrm{E}_\xi[Z_D]$.

A partir d'un échantillon, le problème de l'*estimation spatiale* d'une VR $z(\cdot)$ peut se poser à deux échelles spatiales différentes :

- à l'échelle globale : il s'agit d'estimer la valeur moyenne de $z(\cdot)$ sur D,
- à l'échelle locale : il s'agit de cartographier la variation spatiale de $z(\cdot)$ sur D.

En écologie, on a autant besoin d'estimations locales que d'estimations globales (Houllier 1986). En fait, il se trouve que ces deux problèmes peuvent s'exprimer en termes similaires, et par conséquent faire intervenir les mêmes types de méthodes. En effet, l'estimation locale vise à obtenir des valeurs représentatives de $z(\cdot)$ pour des sous-régions de D tandis que l'estimation globale vise à obtenir une valeur représentative de $z(\cdot)$ pour D tout entier. Ainsi, l'estimation globale peut être vue comme le cas limite de l'estimation locale, *i.e.* lorsqu'il n'y a qu'une région, confondue avec D. Mais il apparaît immédiatement que l'incertitude associée à l'estimation locale sera plus grande que celle associée à l'estimation globale.

Dans ce qui suit, nous examinons successivement les principes généraux de l'estimation spatiale, puis la méthode d'estimation proposée par la géostatistique, et enfin le problème de la précision de ces estimations.

6.1 Méthodologie de l'estimation spatiale

Considérons que le domaine $D \subset \mathbb{R}^2$ sur lequel est définie la variable régionalisée $z(\cdot)$ est discrétisé par un ensemble de supports $\mathcal{U}$. En théorie, si les supports sont des points mathématiques, alors l'ensemble $\mathcal{U}$ est infini. En pratique, le support d'une mesure n'est pas un point mathématique mais plutôt une surface plus ou moins étendue (*e.g.*, placette, quadrat), de sorte que $\mathcal{U}$ est toujours fini. Les supports peuvent être de géométries et/ou d'aires différentes, et éventuellement se chevaucher partiellement. Pour simplifier, et aussi parce qu'en pratique il n'y a pas d'intérêt à ce que les supports des mesures se chevauchent, nous considérons que $\mathcal{U}$ forme une partition de D, autrement dit, une tessellation (Section 2.1.2.3). Si l'aire de chaque support est négligée par rapport à l'aire de D, alors les supports sont dits *quasi-ponctuels* et traités comme s'il s'agissait de points mathématiques (Ripley 1981, Maling 1989).

Dans ce qui suit, nous considérons uniquement des variables additives, *i.e.* des variables dont une valeur moyenne peut être calculée comme la moyenne arithmétique d'un ensemble de valeurs élémentaires. Pour que des variables écologiques classiques telles que les densités (*e.g.*, densités d'animaux, de végétaux, ou de propagules) puissent être manipulées comme des variables additives, il faut que l'aire des supports soit constante (Journel & Huijbregt 1978). En outre, il est plus aisé de traiter des supports ayant tous la même géométrie. En conséquence, nous considérons par la suite que les supports sont de même géométrie, donc de même aire : en pratique c'est généralement le cas dans l'échantillonnage écologique (*e.g.*, méthode des quadrats).

A partir d'un ensemble de supports $s = \{s_i \mid i = 1, \ldots, n\}$, l'estimation spatiale consiste à estimer une valeur $z(u_0)$ localisée en un support $u_0 \notin s$ pour lequel aucune mesure ou observation n'a été réalisée. Deux situations peuvent être distinguées selon que le support u_0 est entouré de supports de s ou bien situé dans une position périphérique par rapport à s, voire même dans un autre domaine que D : dans le premier cas il s'agit d'une *interpolation* et dans le second cas, d'une *extrapolation*. Il est évident que l'extrapolation est une opération beaucoup plus incertaine que l'interpolation. Interpolation et extrapolation sont deux aspects du problème général de l'*inférence spatiale* qui vise à reconstituer au mieux $z(\cdot)$ à partir d'une connaissance imparfaite (fragmentaire, incertaine, floue). Dans la situation idéale, l'inférence spatiale doit s'effectuer en exploitant

le plus possible l'information disponible (Stein 1994). Pour s'approcher de cet idéal, de nombreuses méthodes ont été proposées, notamment :

- les méthodes d'interpolation (revues dans Lam 1983, Myers 1991a, 1994a, Liu & Rossini 1996, Tab. 1),
- les réseaux de neurones (Pariente 1994a, 1994b),
- la combinaison des réseaux de neurones et de l'interpolation (Rizzo & Dougherty 1994),
- la combinaison de la classification (dure ou floue), et de l'interpolation (Gascuel-Odoux *et al.* 1993, Voltz *et al.* 1997, Hendricks-Franssen *et al.* 1997, Burrough *et al.* 1997, Boucneau *et al.* 1998, Ahn *et al.* 1999),
- les systèmes à base de connaissances (Lagacherie 1992, Dimitrakopoulos 1993, Ledreux *et al.* 1994, Zhu *et al.* 1996).

Cependant, si nous nous limitons à l'interpolation, la plupart des méthodes couramment employées estiment $z(u_0)$ comme la combinaison linéaire :

$$z^*(u_0) = \sum_{i=1}^{k} \lambda_i z(s_i) \tag{6.2}$$

avec $\sum \lambda_i = 1$. L'utilisation d'un estimateur linéaire de la forme (6.2) nécessite de choisir le voisinage de u_0, *i.e.* un ensemble de k supports dans s, et de définir le vecteur des pondérateurs $\boldsymbol{\lambda} = (\lambda_i \mid i = 1, \ldots, k)^{\mathrm{T}}$.

6.1.1 Définition du voisinage

L'estimation s'effectue en *voisinage unique* ou *voisinage global* lorsque toutes les données de l'échantillon sont utilisées ($k = n$), et d'estimation en *voisinage glissant* lorsque l'estimateur fait intervenir une partie des données ($k < n$). Il existe de nombreuses façons de définir la géométrie et la taille d'un voisinage glissant, revues notamment par Davis (1986, pp. 370-374), Isaaks & Srivastava (1989, pp. 338-350), Deutsch & Journel (1992, pp. 31-34), Goovaerts (1997, pp. 178-179), et Myers (1997, pp. 387-389).

Le plus simple est d'utiliser un voisinage centré en u_0, circulaire dans le cas d'une variation spatiale isotrope, et elliptique en cas d'une variation anisotrope. L'utilisation d'une ellipse afin de tenir compte de l'anisotropie permet d'obtenir une meilleure estimation spatiale (Goovaerts 1997, p. 178). Par exemple, à l'aide de simulations de patchs de poissons (ou de plancton), et de campagnes acoustiques, Kalikhman & Ostrovsky (1997) montrent que la reconstruction des distributions spatiales est d'autant meilleure que l'orientation de l'ellipse définissant le voisinage glissant correspond à la direction d'élongation des patchs. Ces auteurs concluent que l'utilisation d'un voisinage isotrope peut entraîner des distorsions considérables dans le cas d'une structure spatiale anisotrope.

Nous considérons par la suite que l'estimation s'effectue avec un voisinage non précisé v comportant k données.

6.1.2 Définition des pondérateurs

De même que pour définir le voisinage de u_0, il existe de nombreuses façons de déterminer les pondérateurs de $\boldsymbol{\lambda}^{\mathrm{T}}$. Si tous les pondérateurs sont égaux à $1/k$, la valeur estimée $z^*(u_0)$ n'est autre que la moyenne arithmétique des données du voisinage v :

$$z^*(u_0) = \frac{1}{k} \sum_{i=1}^{k} z(s_i) \tag{6.3}$$

Pour une VR spatialement autocorrélée, il est évident que l'estimation définie par (6.3) n'exploite pas complètement l'information disponible. Dans le cas le plus général, l'autocorrélation spatiale est positive, et il semble raisonnable de considérer qu'une valeur $z(s_i)$ estime d'autant mieux $z(u_0)$ que le support s_i est proche de u_0 : il s'agit du concept d'extension (David 1977).

Il convient en premier lieu de définir la proximité spatiale. Si le domaine D est convexe et ne comporte pas d'obstacles, le plus simple est d'utiliser la distance euclidienne. En revanche, si D est concave et/ou comporte des obstacles, il peut être nécessaire d'imposer que les distances soient calculées sans franchir la frontière ∂D, et/ou en contournant les obstacles (*e.g.*, Little *et al.* 1997), ce qui revient à calculer la longueur de géodésiques (Section 2.2.1.2, p. 21). Pour simplifier, nous considérons par la suite la distance euclidienne $d(u_0, s_i)$.

Soit $\mathrm{ppv}(u_0)$ le plus proche voisin de u_0 dans le voisinage v. La méthode la plus frustre consiste à estimer $z(u_0)$ selon :

$$z^*(u_0) = z(\mathrm{ppv}(u_0)) \tag{6.4}$$

ce qui consiste à affecter un poids $\lambda = 1$ à la valeur $z(\mathrm{ppv}(u_0))$ et un poids $\lambda = 0$ à toutes les autres (Isaaks & Srivastava 1989). D'un point de vue géométrique, cette pondération revient à définir des régions d'influence pour chaque support de l'échantillon, et à considérer une valeur uniforme pour chaque région d'influence (Switzer 1967). Si $\mathbb{R}^2$ est muni de la distance euclidienne, ces zones d'influence sont définies par la tessellation de Voronoï (Section 2.2.2.4) associée au semis s (Isaaks & Srivastava 1989). Comme les discontinuités entre différents polygones de Voronoï sont certainement sans aucun rapport avec la variation spatiale de la VR, cette approche ne devrait pas être utilisée dans le cadre de la cartographie (estimation locale), mais uniquement réservée à l'estimation globale. Dans le contexte de l'estimation globale, la moyenne de $z(\cdot)$ sur D est définie comme l'intégrale d'espace (6.1). Soit $v(s_i)$ le polygone de Voronoï associé au support s_i, la moyenne globale z_D peut être estimée selon :

$$z_D^* = \frac{1}{[D]} \sum_{i=1}^{k} [v(s_i)] \cdot z(s_i) \tag{6.5}$$

avec $[v]$ l'aire du polygone v, ce qui revient à calculer l'estimateur (6.2) avec $\lambda_i = [v(s_i)] / [D]$.

Dans le contexte de l'estimation locale, il convient d'éviter d'utiliser l'estimateur (6.4) qui ne fait intervenir qu'un seul voisin et lui préférer des estimateurs pour lesquels $\lambda_i = f(\delta_i)$ où δ_i exprime la proximité spatiale de u_0 par rapport à s_i. Par défaut, nous

considérons que δ_i est la distance euclidienne dans $\mathbb{R}^2$ $d_i = d(u_0, s_i)$. La pondération $f(d_i)$ la plus simple est celle de la méthode de l'inverse de la distance ou IWD (*Inverse Weighted Distance*) :

$$\lambda_i = d_i^{-p} / \sum_{i=1}^{k} d_i^{-p} \tag{6.6}$$

avec un exposant $p \geq 0$. Afin d'obtenir un interpolateur exact, *i.e.* tel que $z^*(s_i) = z(s_i)$ lorsque $z(s_i)$ fait partie des données, il suffit de définir l'estimateur comme :

$$z^*(u_0) = \begin{cases} z(s_i) & \text{si } d_i = 0 \\ \left[\sum_{i=1}^{k} d_i^{-p}\right]^{-1} \sum_{i=1}^{k} d_i^{-p} z(s_i) & \text{sinon} \end{cases} \tag{6.7}$$

L'estimateur (6.6) s'avère très général puisque, si $p \to 0$, les pondérateurs deviennent similaires et l'estimateur tend vers la moyenne arithmétique (6.3), et si $p \to \infty$, pratiquement tout le poids est affecté à la donnée la plus proche et l'estimateur tend vers (6.4) (Isaaks & Srivastava 1989). Implicitement, le choix de p correspond à une hypothèse concernant la structure d'autocorrélation spatiale de $z(\cdot)$. Lorsqu'il n'y a pas d'autocorrélation spatiale (effet de pépite pur), le choix qui convient est $p = 0$, autrement dit, l'estimateur de la moyenne arithmétique (6.3).

La méthode IWD a le mérite de la simplicité, mais elle présente l'inconvénient de fixer *a priori* un type de variation spatiale, le plus souvent en d^{-2}. Lorsque la structure d'autocorrélation de $z(\cdot)$ est proche de ce schéma de variation spatiale fixé *a priori*, la méthode IWD peut donner de bons résultats. En revanche, un désaccord flagrant entre la structure d'autocorrélation réelle et celle imposée lors de l'interpolation peut conduire à de très mauvais résultats. Une solution à ce problème consiste à définir des pondérateurs qui tiennent compte explicitement de la structure d'autocorrélation spatiale de $z(\cdot)$ comme le propose la géostatistique avec la méthode du krigeage.

6.2 Estimation par krigeage

La géostatistique désigne sous le nom de *krigeage*[1] davantage une méthode de construction d'un prédicteur plutôt qu'un prédicteur particulier. Ainsi, il convient de distinguer le krigeage en tant que méthode et les nombreux prédicteurs qui résultent de son application, principalement :

- en géostatistique linéaire stationnaire : le krigeage simple (KS) et le krigeage ordinaire (KO),
- en géostatistique linéaire non stationnaire : le krigeage universel (KU) ou krigeage avec un modèle de tendance (KT),
- en géostatistique non linéaire : le krigeage lognormal (KL), le krigeage disjonctif (KD), le krigeage d'indicatrices (KI), etc.

[1] Le terme de *krigeage* a été formé au CEA par Pierre Carlier à la fin des années 1950 en hommage à Daniel Krige, ingénieur des mines ayant développé le premier cette procédure pour l'estimation des gisements d'or et d'uranium, en Afrique du Sud (Cressie 1990, Olea 1992).

Pour une revue des différents prédicteurs de krigeage, on peut notamment consulter Journel & Huijbregts (1978), Isaaks & Srivastava (1989), Deustch & Journel (1992), Goovaerts (1997), et en particulier pour le krigeage lognormal, Rivoirard (1990). Nous nous contentons ici :

- de préciser la nature du krigeage, du point de vue classique, puis du point de vue de la théorie de la régression aux moindres carrés,
- d'exposer les systèmes du KS, du KO, et d'un KS modifié que nous utilisons par la suite,
- de discuter de l'intérêt et des limites du krigeage.

6.2.1 Le krigeage

Considérons pour simplifier que les unités échantillonnées sont quasi-ponctuelles, *i.e.* $s = \{x_i \mid i = 1, \ldots, n\}$, avec x un point de $\mathbb{R}^2$. Modélisons la variable régionalisée $z(\cdot)$ comme une réalisation d'une FAI-0 $Z(\cdot)$. Dans ce cadre, il est possible de dériver de nombreux prédicteurs de $Z(x_0)$ exploitant l'information disponible, *i.e.* les données $\{z(x_i) \mid i = 1, \ldots, n\}$ ainsi que la structure d'autocorrélation spatiale entre les VA $Z(x)$, décrite sous la forme du variogramme $\gamma(\cdot)$, ou de la covariance $C(\cdot)$ si $Z(\cdot)$ est une FAST-2.

La forme des prédicteurs dépend des spécifications de la FA ainsi que du respect d'un certain nombre de critères. Ainsi, en géostatistique linéaire les prédicteurs de krigeage classiques sont construits selon la démarche LAUO composée des étapes de Linéarité, d'Autorisation, d'Universalité, et d'Optimalité. Ces quatre étapes sont hiérarchisées et ne peuvent pas être considérées dans n'importe quel ordre (Chauvet 1994, pp. 70-75).

6.2.1.1 Linéarité

La contrepartie des hypothèses peu contraignantes qui sont faites lors de la modélisation est que l'espérance $\mathrm{E}_\xi[\cdot]$ et la variance $\mathrm{Var}_\xi[\cdot]$ ne peuvent opérer que sur des formes linéaires de la FA (Chauvet 1994). Nous considérons donc toujours des combinaisons linéaires de la forme (6.2) :

$$Z^*(x_0) = \sum_{i=1}^{k} \lambda_i Z(x_i) \tag{6.8}$$

ce qui peut s'écrire plus simplement

$$Z_0^* = \sum_{i=1}^{k} \lambda_i Z_i \tag{6.9}$$

6.2.1.2 Autorisation

Selon le type de FA retenu — FAST-2, FAI-0, ou FAI-k ($k > 0$) — toutes les combinaisons linéaires de VA (CL) ne sont pas utilisables. Une CL est dite *combinaison linéaire autorisée* (CLA) si son espérance et sa variance peuvent être définies, *i.e.* s'il est possible d'écrire $\mathrm{E}_\xi[\sum_i \lambda_i Z_i]$ et $\mathrm{Var}_\xi[\sum_i \lambda_i Z_i]$. Dans le modèle FAST-2, toute CL est une CLA mais il faut noter que le passage aux incréments — modèle intrinsèque — introduit une restriction de

l'ensemble des CL autorisées : cette restriction est une contrepartie de l'affaiblissement de l'hypothèse de stationnarité et du gain en généralité des modèles FAI-k. L'ensemble des CLA est ainsi de plus en plus restreint à mesure que l'ordre des FAI-k augmente. Dans le modèle intrinsèque classique (FAI-0), une CL est une CLA si et seulement si le poids total de

$$\sum_{i=1}^{k} \lambda_i Z_i - Z_0 \tag{6.10}$$

est nul, d'où la contrainte sur les inconnues λ_i :

$$\sum_{i=1}^{k} \lambda_i - 1 = 0 \tag{6.11}$$

6.2.1.3 Universalité

Il est souhaitable que le prédicteur obtenu soit sans biais dans le modèle. L'absence de ξ-biais est connue en géostatistique sous le nom d'*universalité*. La contrainte d'universalité impose que l'espérance de l'erreur de prédiction $Z_0^* - Z_0$ soit nulle en moyenne dans le modèle, soit $\mathrm{E}_\xi [Z_0^* - Z_0] = 0$. Cette condition introduit éventuellement une contrainte sur les inconnues λ_i. Dans la littérature, il y a souvent confusion entre la contrainte d'autorisation et celle d'universalité (*e.g.*, Todini & Ferraresi 1996). La modélisation par une FAST-2 à dérive inconnue ne conduit à aucune contrainte d'autorisation mais à la contrainte d'universalité :

$$\sum_{i=1}^{k} \lambda_i = 1 \tag{6.12}$$

En revanche, la modélisation par une FAI-0 à dérive nulle conduit à la contrainte d'autorisation qui s'écrit comme la contrainte (6.12), mais il n'y a pas de contrainte d'universalité (l'espérance d'une CLA dans ce modèle est nulle). Les deux contraintes sont donc formellement identiques, mais leurs signification et importance respectives sont différentes : il est possible de ne pas respecter la contrainte d'universalité mais formellement impossible de ne pas respecter la contrainte d'autorisation (Chauvet 1994, p. 88). Dans la littérature, la contrainte (6.12) est pratiquement toujours mentionnée comme une contrainte d'universalité (contrainte de non biais), ce qui ne change évidemment rien aux calculs.

6.2.1.4 Optimalité

L'*optimalité* consiste à sélectionner le meilleur prédicteur linéaire sans biais ou BLUP (*Best Linear Unbiased Predictor*), au sens d'un certain critère optimisé dans le cadre du modèle. Généralement, l'optimalité impose que $\mathrm{Var}_\xi [Z_0^* - Z_0]$ soit minimale, cette variance jouant le rôle d'un critère de qualité de la prédiction, mais d'autres critères peuvent être utilisés (*e.g.*, Baczkowski & Mardia 1990, Cressie 1993). La variance d'erreur de prédiction (EP) s'écrit en faisant intervenir les fonctions structurales $c_{ij} \equiv C(x_i, x_j)$ ou $\gamma_{ij} \equiv \gamma(x_i, x_j)$, ce qui donne, pour le modèle FAST-2 à dérive inconnue :

$$\mathrm{Var}_\xi [Z_0^* - Z_0] = c_{00} - 2\sum_{i=1}^{k} \lambda_i c_{i0} + \sum_{i=1}^{k}\sum_{j=1}^{k} \lambda_i \lambda_j c_{ij} \tag{6.13}$$

et pour le modèle FAI-0 à dérive nulle :

$$\mathrm{Var}_\xi \left[Z_0^* - Z_0 \right] = 2 \sum_{i=1}^{k} \lambda_i \gamma_{i0} - \sum_{i=1}^{k} \sum_{j=1}^{k} \lambda_i \lambda_j \gamma_{ij} \tag{6.14}$$

La minimisation sous contrainte de la variance d'EP — notée habituellement σ_E^2 — conduit à un système de krigeage dont la résolution fournit les pondérateurs $\boldsymbol{\lambda} = (\lambda_i \mid i = 1, \ldots, k)^{\mathrm{T}}$ et permet de calculer la variance de krigeage σ_K^2 — forme minimisée de σ_E^2 — pour chaque prédicteur Z_0^*. Par exemple, dans le modèle FAI-0 à dérive nulle, la variance de krigeage s'écrit (Journel & Huijbregts 1978, p. 306, Chauvet 1994, p. 74) :

$$\sigma_K^2 = \sum_{i=1}^{k} \lambda_i \gamma_{i0} + \mu \tag{6.15}$$

où μ est un multiplicateur de Lagrange introduit par la contrainte d'autorisation.

Il convient de rappeler que le critère σ_E^2 dérive de l'utilisation du variogramme ou de la covariance, et ignore la forme de la loi spatiale de la FA. Cette approche est parfaitement compatible avec l'hypothèse d'une distribution gaussienne multivariée complètement spécifiée par la covariance, ou par le variogramme si l'on opère sur les incréments d'ordre 1. En géostatistique linéaire, il existe donc un besoin implicite de normalité pour opérer avec le maximum d'efficacité, et les performances se dégradent à mesure que l'on s'écarte de cette hypothèse (Chauvet 1985).

Toutefois, en formulant le krigeage en termes de projection (Journel & Huijbregts 1978, pp. 557-573), il n'est pas nécessaire de recourir à un modèle probabiliste mais uniquement de faire référence à l'ajustement d'une surface à travers les données disponibles, dans un espace muni d'une certaine métrique, autrement dit, à un certain type de régression. Dans ce contexte, il devient possible de définir une famille de distances comprenant le variogramme comme un cas particulier, ce qui permet d'éviter la connotation gaussienne liée à son utilisation, et fournit des outils plus robustes vis-à-vis des valeurs extrêmes (Journel 1988). Le système de krigeage est obtenu en minimisant la norme de l'EP dans l'espace muni de cette nouvelle distance, mais il devient difficile d'interpréter clairement le résultat minimisé. En effet, la valeur minimale de la norme de l'EP ne peut plus être interprétée comme une variance, *i.e.* comme un paramètre de la distribution de l'EP (Journel 1988).

6.2.1.5 Krigeage et régression aux moindres carrés

Le nom *krigeage* témoigne davantage de la rencontre d'outils probabilistes et d'un champ d'application particulier qu'il ne désigne une nouvelle méthode car, à tout prendre, le krigeage n'est pas autre chose qu'une régression linéaire multiple de variance minimale, à partir de données autocorrélées, ou autrement dit, une régression linéaire multiple selon les moindres carrés généralisés (Journel 1986a, Olea 1992, Chauvet 1994, p. 67, Goovaerts 1999).

Considérons une VR $z(\cdot)$ définie sur des supports ponctuels $u = (x, y)$, modélisée par une FA $Z(\cdot)$, et le *krigeage universel* (KU) dont le but et de prendre en compte

une dérive déterministe[2] quelconque $\mathrm{E}_\xi [Z(u)] = m(u)$ grâce à la dichotomie (Matheron 1969) :

$$Z(u) = m(u) + R(u) \tag{6.16}$$

avec $R(\cdot)$ une FAST-2 ou une FAI-0 stricte modélisant la VR résiduelle $r(\cdot) = z(\cdot) - m(\cdot)$. Le système du KU peut être établi, soit par la démarche LAUO (*e.g.*, Chauvet 1994, pp. 83-91), soit dans le cadre de la théorie de la régression linéaire selon les moindres carrés (*e.g.*, Corsten 1985, 1989, Stein et Corsten 1991).

Notons $\mathbf{Z}$ le vecteur aléatoire modélisant le vecteur $\mathbf{z}$ des valeurs observées sur un ensemble fini de supports Ω tel que $\mathrm{Card}(\Omega) = n$. Modélisons la fonction de dérive $m(\cdot)$ par un polynôme de degré k en x et y, soit :

$$m(x, y) = \sum_{i+j=0}^{k} a_{ij} x^i y^j \tag{6.17}$$

comportant $p = \binom{k+2}{2} = (k+1)(k+2)/2$ coefficients. Par exemple, pour $k = 3$ on obtient un polynôme $m(x, y)$ composé de $p = 10$ monômes de coefficients respectifs β_ℓ $(\ell = 0, 1, \ldots, p-1)$:

$$\beta_0 1 + \beta_1 x + \beta_2 y + \beta_3 xy + \beta_4 x^2 + \beta_5 y^2 + \beta_6 x^2 y + \beta_7 xy^2 + \beta_8 x^3 + \beta_9 y^3 \tag{6.18}$$

Le modèle de régression linéaire pour $\mathbf{Z}$ peut s'écrire de façon classique :

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{R} \tag{6.19}$$

avec $\mathbf{X}$ une matrice $n \times p$ de monômes calculés pour les supports $u \in \Omega$, $\boldsymbol{\beta}$ le vecteur des p coefficients de régression, et $\mathbf{R}$ un vecteur aléatoire comportant n variables, soit $\mathbf{R} = (R(u) \mid u \in \Omega)^{\mathrm{T}}$.

Dans la régression linéaire aux moindres carrés ordinaires, les résidus sont non corrélés, d'espérance $\mathrm{E}_\xi [R(u)] = 0$ et de variance $\mathrm{Var}_\xi [R(u)] = \sigma^2$ pour $u \in \Omega$. Soit $\mathbf{x}_0$ le vecteur des valeurs des monômes de $m(x_0, y_0)$ pour un point u_0. Dans le modèle (6.19) la VA en u_0 s'écrit :

$$Z_0 = \mathbf{x}_0^{\mathrm{T}} \boldsymbol{\beta} + R_0 \tag{6.20}$$

Le meilleur estimateur sans biais ou BLUE (*Best Linear Unbiased Estimator*) de la dérive $m(x_0, y_0) = \mathbf{x}_0^{\mathrm{T}} \boldsymbol{\beta}$ est obtenu comme (Corsten 1985) :

$$\widehat{m}(x_0, y_0) = \mathbf{x}_0^{\mathrm{T}} \widehat{\boldsymbol{\beta}} \tag{6.21}$$

avec l'estimateur de $\boldsymbol{\beta}$ selon les moindres carrés ordinaires ou *estimateur OLS* (*Ordinary Least Squares*) :

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}} \mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{z} \tag{6.22}$$

Le BLUP de Z_0 est :

$$Z_0^* = \mathbf{x}_0^{\mathrm{T}} \widehat{\boldsymbol{\beta}} \tag{6.23}$$

[2] On considère classiquement une dérive déterministe, mais une dérive stochastique modélisée par une FA de grande portée est également envisageable (Ripley 1988a, p. 7, Chauvet 1994, p. 91).

puisque $\mathrm{E}_\xi[R_0] = 0$. Le BLUP de Z_0 coïncide avec le BLUE de la dérive (6.21). Cependant, la variance d'estimation de la dérive vaut (Stein & Corsten 1991) :

$$\sigma^2 \left[\mathbf{x}_0^{\mathrm{T}} \left(\mathbf{X}^{\mathrm{T}} \mathbf{X} \right)^{-1} \mathbf{x}_0 \right] \tag{6.24}$$

et diffère de la variance d'EP $Z_0^* - Z_0$ qui vaut (Corsten 1985, 1989, Stein & Corsten 1991) :

$$\sigma^2 \left[\mathbf{x}_0^{\mathrm{T}} \left(\mathbf{X}^{\mathrm{T}} \mathbf{X} \right)^{-1} \mathbf{x}_0 + 1 \right] \tag{6.25}$$

Les estimateur et prédicteur issus du modèle de régression linéaire selon les moindres carrés ordinaires sont ceux de l'*analyse par surface de tendance* (*trend surface analysis*) (*e.g.*, Marcus & Vandermeer 1966, Gittins 1968, Davis 1986, pp. 405-430, Dessaint & Caussanel 1994, Legendre *et al.* 1997). Lorsqu'on ne cherche pas à prédire les données $\mathbf{z}$ elles-mêmes, l'interpolation par surface de tendance est formellement équivalente au KU avec un variogramme à effet de pépite pur. Par ailleurs, on sait que le système du KU a la même structure que celui du krigeage en FAI-k (Chauvet 1986, Christensen 1990). Ainsi, l'interpolation par surface de tendance est également équivalente au krigeage en FAI-k avec une covariance généralisée à effet de pépite pur[3] (Marcotte & David 1988).

Dans le cas général, $z(\cdot)$ présente une structure d'autocorrélation spatiale et le modèle d'erreur de l'interpolation par surface de tendance est inadéquat. Considérons à présent la structure de corrélation du modèle d'erreur exprimée sous la forme d'une matrice de covariance $n+1 \times n+1$ symétrique, définie-positive, et partitionnée comme (Stein & Corsten 1991) :

$$\mathbf{C}^* = \begin{bmatrix} \mathbf{C} & \mathbf{c}_0 \\ \mathbf{c}_0^{\mathrm{T}} & c_{00} \end{bmatrix} \tag{6.26}$$

où $\mathbf{C}$ est la matrice de covariance $n \times n$ des éléments de $\mathbf{R}$, $\mathbf{c}_0$ le vecteur des covariances entre les éléments de $\mathbf{R}$ et R_0, et c_{00} la variance de R_0. Si $\boldsymbol{\beta}$ est connu, alors Z_0 est prédit linéairement comme (Corsten 1989, Stein & Corsten 1991) :

$$Z_0 = \mathbf{x}_0^{\mathrm{T}} \boldsymbol{\beta} + \mathbf{c}_0^{\mathrm{T}} \mathbf{C}^{-1} \left(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta} \right) \tag{6.27}$$

parce que la meilleure approximation linéaire de R_0 par les résidus $\mathbf{R} = \mathbf{Z} - \mathbf{X}\boldsymbol{\beta}$ au sens de la minimisation de la variance d'EP est $\mathbf{c}_0^{\mathrm{T}} \mathbf{C}^{-1} \mathbf{R}$ (Corsten 1985, 1989). Le critère optimisé par (6.27) est le même que celui considéré généralement pour l'optimalité du krigeage (Corsten 1989). En pratique, $\boldsymbol{\beta}$ n'est pas connu mais estimé à partir des données $\mathbf{z}$ par son estimateur des moindres carrés généralisés ou *estimateur GLS* (*Generalized Least Squares*) :

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}} \mathbf{C}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{C}^{-1} \mathbf{z} \tag{6.28}$$

ce qui donne le BLUE de la dérive comme en (6.21) et le BLUP de Z_0 :

$$Z_0^* = \mathbf{x}_0^{\mathrm{T}} \widehat{\boldsymbol{\beta}} + \mathbf{c}_0^{\mathrm{T}} \mathbf{C}^{-1} \left(\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\beta}} \right) \tag{6.29}$$

et en remplaçant $\mathbf{Z}$ par sa réalisation $\mathbf{z}$, la prédiction :

$$z_0^* = \mathbf{x}_0^{\mathrm{T}} \widehat{\boldsymbol{\beta}} + \mathbf{c}_0^{\mathrm{T}} \mathbf{C}^{-1} \left(\mathbf{z} - \mathbf{X}\widehat{\boldsymbol{\beta}} \right) \tag{6.30}$$

[3] Mis à part le fait qu'une surface de tendance ne constitue pas un interpolateur exact puisqu'on peut avoir $z^*(s_i) \neq z(s_i)$, avec $z(s_i)$ une valeur figurant dans les données.

Contrairement au modèle d'erreurs non corrélées, le BLUE de la dérive (6.21) ne coïncide plus avec le BLUP de Z_0 (6.29), ce qui se traduit en géostatistique par des systèmes de krigeage spécifiquement dédiés à l'estimation optimale de la dérive elle-même (*e.g.*, Castelier & Laurenge 1993, Chauvet 1994, pp. 91-98). Enfin, la variance d'EP $Z_0^* - Z_0$ vaut (Corsten 1989, Stein & Corsten 1991) :

$$c_{00} - \mathbf{c}_0^{\mathrm{T}} \mathbf{C}^{-1} \mathbf{c}_0 + \mathbf{x}_a^{\mathrm{T}} \left(\mathbf{X}^{\mathrm{T}} \mathbf{C}^{-1} \mathbf{X} \right)^{-1} \mathbf{x}_a \tag{6.31}$$

avec $\mathbf{x}_a = \mathbf{x}_0 - \mathbf{X}^{\mathrm{T}} \mathbf{C}^{-1} \mathbf{c}_0$.

Dans le contexte de la régression, si la covariance de $R(\cdot)$ n'est pas définie (hypothèse FAI-0 stricte, ou plus généralement, hypothèse FAI-k de la géostatistique) il convient de considérer les contrastes (incréments d'ordre 1 de la géostatistique), voire les contrastes généralisés (incréments généralisés de la géostatistique). Il s'avère donc en principe possible d'exprimer dans les termes de la théorie de la régression tout système de krigeage obtenu dans le cadre de la démarche LAUO de la géostatistique linéaire.

Stein & Corsten (1991) estiment que les multiplicateurs de Lagrange introduits par la minimisation sous contraintes de la démarche LAUO rend la notation géostatistique inutilement complexe et obscure. Effectivement, l'utilisation de la théorie de la régression rend inutile les multiplicateurs de Lagrange, bien que certains auteurs utilisent à la fois la régression et les multiplicateurs de Lagrange (*e.g.*, Lefèbvre *et al.* 1996). Sans entrer dans le débat assez subjectif de la lisibilité des notations et de la facilité de compréhension des différentes approches, nous utilisons par la suite la forme géostatistique des systèmes de krigeage parce qu'elle est la plus largement répandue et admet une formulation duale très utile.

6.2.2 Systèmes de krigeage

En fonction de la problématique, du type de VR, du modèle de FA, et du critère optimisé, l'application de la démarche LAUO conduit à un estimateur de krigeage particulier. Il ne s'agit pas pour nous de tous les décrire, mais d'introduire le krigeage ordinaire (KO) et le krigeage simple (KS) qui constituent deux systèmes de krigeage fondamentaux, ainsi qu'un krigeage modifié (KM) que nous utilisons par la suite.

Tout système de krigeage admet deux formulations. La formulation primale, classique, exprime le prédicteur comme une combinaison linéaire des données, tandis que la formulation duale l'exprime comme une combinaison linéaire de covariances ou de variogrammes (Dubrule 1983a, Journel & Rossi 1989). La formulation duale est particulièrement utile en pratique, notamment pour calculer des cartes isoplèthes (Trochu 1993).

Dans ce qui suit, nous considérons que l'estimation est effectuée en un point x_0 à partir des données mesurées ou observées en n supports ponctuels $\{x_i \mid i = 1, \ldots, n\}$. Les supports sont supposés distincts afin d'assurer que les matrices de krigeage ne sont jamais singulières (O'Dowd 1991). Dans les systèmes de krigeage, en forme primale et duale, nous substituons automatiquement $z(\cdot)$ à $Z(\cdot)$.

6.2.2.1 Krigeage ordinaire

Le krigeage ordinaire est établi le plus souvent en considérant une FAI-0 de dérive nulle, autrement dit, une FA qui satisfait aux hypothèses :

$$\mathrm{E}_\xi \left[Z(x) - Z\left(x+h\right)\right] = m(h) = 0 \tag{6.32}$$

$$\mathrm{Var}_\xi \left[Z(x) - Z\left(x+h\right)\right] = 2\gamma\left(h\right) \tag{6.33}$$

Dans ce modèle, il n'y a pas de contrainte d'universalité mais une contrainte d'autorisation $\sum_\beta \lambda_\beta = 1$. Il est également possible de considérer le krigeage ordinaire dans un modèle FAST-2 à moyenne inconnue (*e.g.*, Goovaerts 1999). Dans ce cas, il n'y a pas de contrainte d'autorisation mais une contrainte d'universalité $\sum_\beta \lambda_\beta = 1$. L'application de la démarche LAUO à ces modèles donne un système de krigeage de la même forme mais dans lequel il faut substituer le variogramme à la covariance si la FA n'est pas FAST-2 mais uniquement FAI-0 stricte (*cf.* Aubry 1996a, pp. 35-40). Le système de krigeage ordinaire en forme primale s'écrit :

$$\begin{cases} \sum\limits_{\beta=1}^{n} \lambda_\beta \gamma\left(x_\alpha, x_\beta\right) + \mu = \gamma\left(x_\alpha, x_0\right) \\ \sum\limits_{\beta=1}^{n} \lambda_\beta = 1 \end{cases} \tag{6.34}$$

pour $\alpha = 1, \ldots, n$, soit sous forme matricielle $\mathbf{\Gamma\lambda} = \boldsymbol{\gamma}$:

$$\begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1n} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma_{n1} & \cdots & \gamma_{nn} & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} . \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma_{10} \\ \vdots \\ \gamma_{n0} \\ 1 \end{bmatrix} \tag{6.35}$$

L'estimation s'effectue point par point selon :

$$z_0^* = \sum_{\alpha=1}^{n} \lambda_\alpha z_\alpha \tag{6.36}$$

Le système de krigeage ordinaire peut s'écrire sous forme duale (*cf.* Aubry 1996b, p. 44) :

$$\begin{cases} \sum\limits_{\beta=1}^{n} \nu_\beta \gamma\left(x_\alpha, x_\beta\right) + \phi = z\left(x_\alpha\right) \\ \sum\limits_{\beta=1}^{n} \nu_\beta = 0 \end{cases} \tag{6.37}$$

pour $\alpha = 1, \ldots, n$, soit sous forme matricielle $\mathbf{\Gamma\nu} = \mathbf{z}$:

$$\begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1n} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma_{n1} & \cdots & \gamma_{nn} & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} . \begin{bmatrix} \nu_1 \\ \vdots \\ \nu_n \\ \phi \end{bmatrix} = \begin{bmatrix} z_1 \\ \vdots \\ z_n \\ 0 \end{bmatrix} \tag{6.38}$$

L'estimation s'effectue selon la fonction d'interpolation :

$$z_0^* = \phi + \sum_{\alpha=1}^{n} \nu_\alpha \gamma\left(x_\alpha, x_0\right) \tag{6.39}$$

6.2.2.2 Krigeage simple

Le krigeage simple est établi en considérant une FAST-2 de dérive connue — ce qui est plus contraignant que dans le cas du KO — autrement dit une FA qui satisfait aux hypothèses :

$$\mathrm{E}_\xi\left[Z(x)\right] = m(h) = m \tag{6.40}$$

$$\mathrm{Cov}_\xi\left[Z(x), Z\left(x+h\right)\right] = C\left(h\right) \tag{6.41}$$

Dans ce modèle, il n'y a pas de contrainte d'universalité ni de contrainte d'autorisation. L'application de la démarche LAUO donne le système de krigeage simple en forme primale (*cf.* Aubry 1996b, pp. 45-47) :

$$\sum_{\beta=1}^{n} \lambda_\beta C\left(x_\alpha, x_\beta\right) = C\left(x_\alpha, x_0\right) \tag{6.42}$$

pour $\alpha = 1, \ldots, n$, soit sous forme matricielle $\mathbf{C}\boldsymbol{\lambda} = \mathbf{c}$:

$$\begin{bmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nn} \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} c_{10} \\ \vdots \\ c_{n0} \end{bmatrix} \tag{6.43}$$

L'estimation s'effectue point par point selon :

$$z_0^* = m + \sum_{\alpha=1}^{n} \lambda_\alpha \left[z_\alpha - m\right] \tag{6.44}$$

Le système de krigeage simple peut s'écrire sous forme duale :

$$\sum_{\beta=1}^{n} \nu_\beta C\left(x_\alpha, x_\beta\right) = \left[z\left(x_\alpha\right) - m\right] \tag{6.45}$$

pour $\alpha = 1, \ldots, n$, soit sous forme matricielle $\mathbf{C}\boldsymbol{\nu} = \mathbf{z}$:

$$\begin{bmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nn} \end{bmatrix} \cdot \begin{bmatrix} \nu_1 \\ \vdots \\ \nu_n \end{bmatrix} = \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix} \tag{6.46}$$

L'estimation s'effectue selon la fonction d'interpolation :

$$Z_0^* = m + \sum_{\alpha=1}^{n} \nu_\alpha C\left(x_\alpha, x_0\right) \tag{6.47}$$

6.2.2.3 Krigeage modifié

La minimisation de la variance d'EP $\mathrm{Var}_\xi\left[Z_0^* - Z_0\right] = \mathrm{E}_\xi\left[\left(\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{Z} - Z_0\right)^2\right]$ a l'inconvénient de produire des valeurs estimées moins variables que les valeurs réelles, les fortes valeurs ayant tendance à être sous-estimées et les faibles valeurs à être surestimées (David 1977,

de Fouquet 1993, Gunnarsson *et al.* 1998, Goovaerts 1999). Ce *phénomène de lissage* est en fait une propriété de la plupart des méthodes d'interpolation (Switzer 1979, 1983). Pour pallier le lissage, il est possible de procéder à des corrections *ad hoc* (*e.g.*, Pan 1994, Olea & Pawlowsky 1996), ou bien de minimiser un critère tel que celui issu des travaux de Cressie (1993) (Hosseini *et al.* 1994) :

$$\mathrm{E}_\xi\left[\left(\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{Z}-Z_0\right)^2\right]+2m_1\left(\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{1}-1\right)+m_2\left(\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{C}\boldsymbol{\lambda}-\sigma^2\right) \tag{6.48}$$

de sorte que $\mathrm{Var}_\xi\left[Z_0^*\right]=\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{C}\boldsymbol{\lambda}=\sigma^2$, avec $\sigma^2\equiv C\left(0\right)$ (Hosseini *et al.* 1994). La minimisation de (6.48) conduit à un système de krigeage simple modifié qui peut s'écrire sous forme primale (Hosseini *et al.* 1994) :

$$\sum_{\beta=1}^{n}\lambda_\beta C\left(x_\alpha,x_\beta\right)=\left[C\left(x_\alpha,x_0\right)-m_1\right]/m_2 \tag{6.49}$$

pour $\alpha=1,\ldots,n$, avec :

$$m_1=\frac{\mathbf{c}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{1}-m_2}{\mathbf{1}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{1}} \tag{6.50}$$

et

$$m_2=\frac{\left[\mathbf{c}^{\mathrm{T}}\mathbf{C}^{-1}\left(\mathbf{c}\mathbf{1}^{\mathrm{T}}-\mathbf{1}\mathbf{c}^{\mathrm{T}}\right)\mathbf{C}^{-1}\mathbf{1}\right]^{1/2}}{\left[\sigma^2\mathbf{1}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{1}-1\right]^{1/2}} \tag{6.51}$$

avec $\mathbf{1}$ un vecteur $n\times 1$ rempli de 1. L'estimation s'effectue point par point selon :

$$z_0^*=m+\sum_{\alpha=1}^{n}\lambda_\alpha\left[z_\alpha-m\right] \tag{6.52}$$

Le système de krigeage modifié peut s'écrire sous forme duale :

$$\sum_{\beta=1}^{n}\nu_\beta C\left(x_\alpha,x_\beta\right)=\left[z\left(x_\alpha\right)-m\right] \tag{6.53}$$

pour $\alpha=1,\ldots,n$, et l'estimation s'effectue selon la fonction d'interpolation :

$$z_0^*=m+\sum_{\alpha=1}^{n}\nu_\alpha\left[C\left(x_\alpha,x_0\right)-m_1\right]/m_2 \tag{6.54}$$

6.2.3 Intérêts et limites

Les prédicteurs du krigeage ont été largement comparés entre eux et à d'autres méthodes d'estimation spatiale tant sur le plan des principes mis en jeu qu'en ce qui concerne les performances réelles établies d'après des études de cas (Lam 1983, Journel & Rossi 1989, Boufassa & Armstrong 1989, Hass 1990, Laslett & McBratney 1990, Englund 1990, Weber & Englund 1992, 1994, Carr & Mao 1993, Laslett 1994, Hosseini *et al.* 1994, Pariente 1994a, Gotway *et al.* 1996, Weisz *et al.* 1995b, Brus *et al.* 1996, Borga & Vizzaccaro 1997, Ashraf *et al.* 1997, Goovaerts 1997, etc.).

Cependant ce type de considération se révèle souvent de peu d'enseignement parce que :

- la nature des méthodes comparées n'est pas toujours bien comprise,
- la comparaison est limitée à un jeu de données et est menée selon une méthodologie trop faible et/ou par différents praticiens.

Dans ces conditions il est difficile de faire la part entre ce qui est dû à la méthode proprement dite et ce qui provient de l'utilisateur.

L'objet de cette section n'est donc pas de chercher à évaluer les performances d'un certain type de krigeage, pour un ensemble fini de cas particuliers, mais plutôt d'examiner ses prétendus avantages et inconvénients. Nous abordons successivement les questions :

- de l'intérêt pratique du krigeage,
- de la robustesse du krigeage,
- des pondérateurs négatifs du krigeage.

6.2.3.1 Intérêt du krigeage

Les nombreuses applications du krigeage (Annexe G) sont motivées essentiellement par trois caractéristiques de la méthode, qui sont autant de sources potentielles de malentendus :

- l'absence de "biais",
- le caractère "optimal" du krigeage,
- la nature de la variance de krigeage.

Les malentendus sont dus à une compréhension superficielle qui s'accompagne d'un exposé très incomplet de la théorie du krigeage lors du transfert de la technique dans les différents domaines d'application. Néanmoins, une fois que les malentendus éventuels ont été identifiés, il est possible de reconnaître l'intérêt pratique **réel** du krigeage et de comparer différents prédicteurs entre eux.

Absence de biais L'absence de biais ou, dans le vocabulaire géostatistique, l'universalité du krigeage, se définit formellement comme :

$$\mathrm{E}_\xi \left[Z_0^* - Z_0 \right] = 0 \tag{6.55}$$

L'intérêt d'un formalisme précis est de permettre une définition sans ambiguïté de la notion d'absence de biais. Ainsi, l'expression (6.55) montre clairement que l'absence de biais du krigeage (ξ-biais) est définie dans le cadre d'un modèle de superpopulation. L'absence de ξ-biais, parfaitement légitime lorsqu'il s'agit de dériver un BLUP tel que le prédicteur du krigeage Z_0^*, n'a cependant pas de contrepartie directe dans le cadre de l'estimation locale de la valeur z_0 d'une population fixée. En pratique, l'absence de ξ-biais se traduit généralement par une compensation globale des surestimations et des sous-estimations locales, ce qui confère au krigeage d'intéressantes propriétés pour l'estimation globale. En revanche, cette propriété n'offre aucune garantie en ce qui concerne la précision des estimations **locales**.

Optimalité de l'interpolation par krigeage A nouveau, il convient d'insister sur le fait que l'optimalité dont il est question est définie en moyenne pour une infinité de réalisations d'une fonction aléatoire $Z(\cdot)$ et ne concerne pas une réalisation particulière, *i.e.* la VR étudiée $z(\cdot)$. En pratique, l'interpolation effectuée par krigeage n'est pas nécessairement meilleure — au sens d'un ou plusieurs critères à préciser — que celle obtenue par la méthode IWD (Weisz *et al.* 1995b, Brus *et al.* 1996, Dodson & Marks 1997), par les splines (Deutsch & Journel 1992, Hosseini *et al.* 1994, Gotway *et al.* 1996, Brus *et al.* 1996), voire même par ajustement de surfaces multiquadratiques[4] (Borga & Vizzaccaro 1997).

Les résultats de Weber & Englund (1992) semblent montrer que la méthode IWD peut donner de meilleurs résultats que le KS ou le KO. Néanmoins le krigeage n'est pas appliqué dans les règles de l'art — les variogrammes sont ajustés à la main et le voisinage glissant est défini *a priori*, sans validation croisée (Section 7.2.2) — ce qui affaiblit considérablement le crédit de cette étude. Le problème essentiel posé par la méthode IWD est qu'il n'existe pas de puissance qui donne de bons résultats quel que soit le jeu de données, la méthode étant notamment sensible à la forme de la distribution statistique (Weber & Englund 1994, Gotway *et al.* 1996). Ainsi, une puissance correcte devrait être déterminée par validation croisée (*e.g.*, Varekamp *et al.* 1996), ce qui revient en fait à proposer un modèle de variation spatiale en adéquation avec les données.

Les splines peuvent donner des prédictions dont la précision moyenne est semblable à celle du krigeage (Hosseini *et al.* 1994). En fait, en utilisant la forme duale du krigeage, on montre que les splines sont équivalentes au krigeage en FAI-k pour une covariance généralisée d'une certaine forme (Dubrule 1983a, 1984, Chauvet 1994, pp. 165-168). Ainsi, le recours aux splines revient à kriger en utilisant une structure de covariance qui n'est pas estimée d'après les données mais fixée *a priori*.

En général, les interpolateurs qui lissent assez fortement donnent d'assez bons résultats, ce qui explique le succès relatif des splines et du krigeage, et les mauvais résultats du krigeage modifié (Hosseini *et al.* 1994). Les conditions requises pour une interpolation satisfaisante sont exposées de façon assez limpide dans Isaaks & Srivastava (1989, pp. 49-50).

Nature de la variance de krigeage La variance de krigeage σ_K^2 (forme minimisée de σ_E^2) est généralement vue comme une mesure de précision de l'estimation qui permettrait de construire un "intervalle de confiance" (Lam 1983, Robertson 1987, Gotway *et al.* 1996). Il faut à nouveau insister sur le fait que, dans le cas d'une VR particulière, la variance de krigeage ne constitue pas à proprement parler une mesure de précision locale (Journel 1986a, Journel & Rossi 1989), et qu'elle ne permet pas de construire des estimations par intervalle fiables, *i.e.* en rapport avec la précision réelle des valeurs estimées.

Pour s'en convaincre, il suffit de kriger chaque valeur $z(x_j)$ à partir des données restantes $\{z(x_i) \mid i = 1, \ldots, n \; ; \; i \neq j\}$ et de former les erreurs d'estimation réelles $\varepsilon_j = z^*(x_j) - z(x_j)$ pour $j = 1, \ldots, n$. On constate qu'il n'y a généralement aucune corrélation entre les $|\varepsilon_j|$ et les variances de krigeage associées (Journel & Rossi 1989). En conséquence, il est le plus souvent illusoire de recourir à la variance de krigeage pour calculer des estimations par intervalle **locales**.

[4] Borga & Vizzaccaro (1997) montrent que l'interpolation par surface multiquadratique conique est formellement équivalente au krigeage avec variogramme linéaire.

Intérêt pratique réel du krigeage Les deux sections précédentes ne doivent pas laisser penser que la minimisation de σ_E^2 et les propriétés du prédicteur qui en résulte (prédicteur du krigeage) sont sans véritable intérêt en pratique. La minimisation de σ_E^2 est une démarche tout à fait raisonnable qui a pour conséquence essentielle une interpolation assez lisse des données, ce qui n'est pas autre chose qu'une régression multiple aux moindres carrés avec erreurs autocorrélées (Section 6.2.1.5).

Le krigeage possède en outre plusieurs propriétés intéressantes en pratique. Le krigeage est tout d'abord un interpolateur exact, mais cette propriété n'est pas propre à ce type de procédure et peut être aisément assurée pour n'importe quel algorithme d'interpolation. L'intérêt primordial du krigeage est évidemment la prise en compte explicite de la forme de la dépendance spatiale des données, de l'implantation et de la géométrie des supports des données ou de ceux des valeurs à estimer. En particulier, les pondérateurs du krigeage dépendent non seulement des distances $d(x_i, x_0)$ entre les supports $\{x_i \mid i = 1, \ldots, n\}$ et le support x_0 de la valeur à estimer, mais également des distances mutuelles $d(x_i, x_j)$ pour $i, j = 1, \ldots, n$. Il en découle deux propriétés particulièrement intéressantes (Oliver & Webster 1991, Olea 1992) :

- les valeurs dont les supports sont proches les uns des autres reçoivent collectivement le même poids qu'une seule valeur dont le support serait proche du centroïde de l'agrégat (*declustering*),
- les données séparées du support à estimer par des données en position intermédiaire ont une influence réduite (effet d'écran ou *screen effect*).

S'il est certainement abusif de considérer que les propriétés du krigeage en font **la** meilleure méthode, tant du point de vue théorique que pratique (*e.g.*, Olea 1992), il convient de reconnaître que la prise en compte de nombreuses caractéristiques spatiales des données par le krigeage est évidemment préférable à l'utilisation d'une procédure universelle dont les paramètres sont fixés *a priori*. C'est en ce sens que nous pouvons recommander l'utilisation du krigeage en écologie pour l'estimation locale (cartographie).

Evidemment, le krigeage ne présente d'intérêt que dans le cas de l'interpolation d'une VR spatialement autocorrélée. Cependant, la procédure reste valide dans le cas d'un effet de pépite pur, et l'estimation par krigeage s'avère simplement équivalente à la moyenne arithmétique (Myers 1991a). En général, le krigeage n'est utile que si la pépite relative $c_0/(c_0 + c)$ est modérée (Myers 1991a).

Nous recommandons également d'utiliser le krigeage pour l'estimation globale, du moins lorsque le motif d'échantillonnage est irrégulier. En effet, dans le cas de l'estimation globale et d'une répartition homogène des supports dans D (*e.g.*, échantillonnage systématique), le krigeage ne présente pratiquement aucun intérêt par rapport à la moyenne arithmétique (Matheron 1965, Petitgas 1991, Papritz & Webster 1995b, Aubry 1996b).

Comparaison des prédicteurs du krigeage En ce qui concerne les avantages comparés des différents types de prédicteurs obtenus dans le cadre de la méthode du krigeage, on s'accorde généralement sur les points suivants :

- le KS peut être utilisé lorsque l'hypothèse de stationnarité (globale et pas seulement locale) est raisonnable, mais il faut lui substituer le KO dans les autres cas (Boufassa & Armstrong 1989),
- la dichotomie tendance/résidu du KU est fondamentalement indéterminée (Armstrong 1984a, de Kwaadsteniet 1990, de Fouquet 1993), et il serait préférable de recourir au formalisme des FAI-k (Armstrong 1984a, Chauvet 1986),
- en présence d'une tendance et lorsque l'effet de pépite est faible, le KU ne conduit pas nécessairement à de meilleurs prédictions que le KO en voisinage glissant, et la différence se manifeste essentiellement en situation d'extrapolation (Journel & Rossi 1989, Posa & Marcotte 1992),
- en présence d'une distribution statistique asymétrique, les méthodes de krigeage linéaire (KS mais surtout KO) et non linéaires (*e.g.*, KL, KD, KI) peuvent donner des résultats similaires (Boufassa & Armstrong 1989, Englund 1990).

Ainsi, il ne semble pas toujours justifié en pratique d'investir beaucoup d'efforts dans des méthodes complexes, l'algorithme essentiel restant celui du krigeage ordinaire (Deutsch & Journel 1992).

En écologie, l'apport pratique des méthodes de krigeage non stationnaire ou non linéaire reste à évaluer. En particulier, l'intérêt du krigeage lognormal (KL) est loin d'être établi. Simard *et al.* (1992, 1993) déconseillent l'utilisation du KL parce que les résultats sont très sensibles aux écarts des données par rapport à la distribution lognormale stricte. Goovaerts (1999) fait remarquer que le résultat final est également très sensible au modèle de variogramme qui contrôle la variance de krigeage, parce que la transformation inverse non-biaisée qui intervient dans le KL fait intervenir l'exponentiation à la fois de l'estimation et de la variance de krigeage. En conséquence, le manque de robustesse du KL nous conduit à y renoncer.

6.2.3.2 Robustesse du krigeage

L'étude de la robustesse d'un prédicteur de krigeage (*e.g.*, KO ou KU) en tant que procédure d'interpolation nécessite en principe de considérer la perturbation :

- du modèle de variogramme à travers le choix du type de modèle et/ou de ses paramètres,
- de la position des supports des données,
- des valeurs des données.

Les résultats géostatistiques sont en général assez robustes vis-à-vis du choix du modèle de variogramme, pour autant que les paramètres aient été déterminés dans les règles de l'art (Journel & Huijbregts 1978, pp. 167, 233).

Les prédictions du krigeage sont surtout sensibles au comportement à l'origine du variogramme (effet de pépite, allure linéaire ou parabolique) qui traduit la régularité

spatiale de la VR, ainsi qu'à la portée qui intervient dans le calcul des pondérateurs du krigeage. En revanche, le seuil du variogramme a un impact uniquement sur les variances de krigeage et pas sur les estimations. En effet, il est possible de modifier l'échelle du variogramme, par exemple en le standardisant, sans pour autant que cela influence les pondérateurs du krigeage (*e.g.*, Goovaerts 1997, pp. 105-106).

Cependant, il convient d'envisager la robustesse du krigeage du point de vue numérique en particulier, ce qui nécessite de distinguer le système de krigeage en tant que tel et le krigeage en tant qu'algorithme numérique implémenté sur un ordinateur donné. En effet, il ne faut pas confondre le *conditionnement* du problème d'analyse numérique associé au krigeage, et la *stabilité numérique* des algorithmes destinés à résoudre ce problème (O'Dowd 1991).

Conditionnement La solution d'un système d'équations linéaires de la forme $\mathbf{Ax} = \mathbf{b}$ est obtenue par inversion ou par factorisation de la matrice $\mathbf{A}$ (*e.g.*, Press *et al.* 1989, pp. 27-46). Le système $\mathbf{Ax} = \mathbf{b}$ est *a priori* d'autant plus sensible à une petite perturbation que le nombre de conditionnement (*condition number*) de $\mathbf{A}$ est élevé, le nombre de conditionnement étant défini comme (Golub & van Loan 1983, p. 25, Stewart 1985) :

$$\kappa(\mathbf{A}) = \|\mathbf{A}\| \ \left\|\mathbf{A}^{-1}\right\| \tag{6.56}$$

avec $\|\cdot\|$ une norme[5] satisfaisant $\|\mathbf{Ax}\| \leq \|\mathbf{A}\| \, \|\mathbf{x}\|$. Le nombre de conditionnement[6] prend les valeurs extrêmes $\kappa(\mathbf{A}) = 1$ pour une matrice identité et $\kappa(\mathbf{A}) = \infty$ pour une matrice singulière, *i.e.* une matrice pour laquelle il n'existe pas de matrice inverse $\mathbf{A}^{-1}$ et, en conséquence, pas de solution unique[7] pour le système $\mathbf{Ax} = \mathbf{b}$. Dans le krigeage, $\kappa(\mathbf{A})$ est déterminé à la fois par le variogramme et par la configuration spatiale des supports (Diamond & Armstrong 1984, Ababou *et al.* 1994).

Soit $\mathbf{C}$ une matrice de covariance stationnaire, *i.e.* correspondant à un modèle de variogramme de transition. La sensibilité aux perturbations des valeurs des données est liée au nombre de conditionnement $\kappa(\mathbf{C})$ (Dietrich 1990). Si le modèle de covariance est seulement semi-défini positif et non pas strictement défini positif, il est possible que $\mathbf{C}$ soit singulière pour certaines configurations d'échantillonnage (Myers & Journel 1990). Lorsque $\mathbf{C}$ n'est pas singulière, $\kappa(\mathbf{C})$ augmente avec la densité de l'échantillonnage (supports de plus en plus proches les uns des autres), la taille du système (très nombreuses données), et la portée (O'Dowd 1991, Ababou *et al.* 1994). Ainsi, quel que soit le type de perturbation envisagé, le problème de la robustesse du krigeage est étroitement lié au nombre de conditionnement.

Dans le cas d'un modèle de variogramme parabolique à l'origine tel que le modèle gaussien, la matrice de covariance $\mathbf{C}$ peut s'avérer très mal conditionnée (Dietrich 1990, O'Dowd 1991, Ababou *et al.* 1994). Une matrice de covariance $\mathbf{C}$ mal conditionnée implique que la matrice de krigeage $\mathbf{A}$ sera également mal conditionnée, la réciproque n'étant pas nécessairement vraie (O'Dowd 1991) — sauf bien entendu dans le cas du KS pour lequel $\mathbf{A} \equiv \mathbf{C}$. Il en découle que le modèle gaussien peut conduire à un système de krigeage

[5] Différentes définitions de la norme d'une matrice sont revues dans Golub & van Loan (1983, pp. 11-16) et Posa (1989).

[6] Le problème de l'estimation de $\kappa(\mathsf{A})$ est traité notamment par Cline *et al.* (1979).

[7] A moins d'utiliser l'*inverse généralisée* de Moore-Penrose (Wackernagel 1993, pp. 9-10).

beaucoup plus instable que dans le cas d'un modèle sphérique ou exponentiel (Diamond & Armstrong 1984, Warnes 1986, Bárdossy 1988, Posa 1989).

Parmi les modèles de transition classiques, le modèle exponentiel est le plus stable (Posa & Marcotte 1992). En effet, le modèle exponentiel est encore plus stable que le modèle sphérique (Davis & Morris 1997). Bien qu'il soit également parabolique à l'origine, le modèle cubique s'avère plus robuste que le modèle gaussien (Ababou *et al.* 1994). Au-delà du comportement parabolique à l'origine, c'est le caractère infiniment différentiable du modèle gaussien qui est la cause de son instabilité (Dietrich 1990, Ababou *et al.* 1994). Bien que le modèle périodique n'ait pas été étudié jusqu'à présent, il faut s'attendre à observer une instabilité au moins aussi importante qu'avec le modèle gaussien. Ainsi, l'instabilité des modèles de variogramme est fonction du degré de régularité spatiale sous-jacent.

D'après ce qui précède, et en anticipant certains résultats exposés plus loin (Sections 11.2.4 & 11.3.3), nous faisons la proposition suivante :

Proposition 2 *Soit* $\mathbf{C}$ *une matrice de covariance correspondant à un modèle de variogramme de transition exponentiel (*$\mathbf{C}_1$*), pentasphérique (*$\mathbf{C}_2$*), sphérique (*$\mathbf{C}_3$*), cubique (*$\mathbf{C}_4$*), gaussien (*$\mathbf{C}_5$*), ou périodique (*$\mathbf{C}_6$*). A paramétrage constant* $c_0 = 0$*,* $c > 0$*, et* $a > 0$*, les nombres de conditionnement* $\kappa(\mathbf{C})$ *sont ordonnés selon :* $\kappa(\mathbf{C}_1) < \kappa(\mathbf{C}_2) < \kappa(\mathbf{C}_3) < \kappa(\mathbf{C}_4) < \kappa(\mathbf{C}_5) < \kappa(\mathbf{C}_6)$.

Dans tous les cas, l'introduction d'une pépite $c_0 > 0$ entraîne une diminution du nombre de conditionnement (O'Dowd 1991, Ababou *et al.* 1994), ce qui constitue en pratique un bon moyen pour augmenter la stabilité du système de krigeage (Posa 1989, Dietrich 1990).

Par ailleurs, dans le cadre du KU, Diamond & Armstrong (1984) montrent que l'instabilité du système de krigeage croît à mesure que le nombre de termes de la dérive augmente. Par conséquent, une matrice de KU établie d'après un variogramme gaussien (ou périodique), sans pépite, et avec une tendance polynomiale de degré élevé, sera particulièrement mal conditionnée.

Sukhatme (1989) traite le calcul de l'effet d'une perturbation du variogramme sur les pondérateurs du krigeage dans le cadre général de la théorie de la perturbation, sans toutefois en dériver des recommandations pratiques. Stein & Handcock (1989) étudient les propriétés asymptotiques du krigeage (*i.e.*, pour $n \to \infty$) lorsque la covariance est mal spécifiée, sujet qui intéresse surtout les statisticiens, et qui est généralement éloigné des préoccupations très concrètes des utilisateurs.

Stabilité numérique Quel que soit le conditionnement du système de krigeage, il est évidemment recommandé d'utiliser un algorithme de résolution numériquement stable et d'effectuer les calculs avec des flottants codés sur un grand nombre d'octets. Différents algorithmes sont revus dans Press *et al.* (1989) et McCarn & Carr (1992).

En exprimant le krigeage en termes de projection, Kacewicz (1991) propose d'utiliser l'orthogonalisation de Gram-Schmidt (Golub & van Loan 1983, pp. 150-152, Press *et al.* 1989, pp. 376-378) parce que cette procédure est numériquement stable tout en étant d'une complexité algorithmique comparable à celle de la décomposition de Cholesky.

En considérant le krigeage sous sa forme classique, nous recommandons d'utiliser la méthode de Crout avec pivot partiel décrite dans Press *et al.* (1989, pp. 39-46), et d'effectuer les calculs avec des flottants codés sur 10 octets.

6.2.3.3 Pondérateurs négatifs

La géostatistique linéaire n'impose aucune contrainte sur le signe des pondérateurs $\boldsymbol{\lambda} = (\lambda_i \mid i = 1, \ldots, n)^{\mathrm{T}}$ qui peuvent éventuellement être négatifs (Chauvet 1988, Myers 1991a). Parmi les facteurs susceptibles de produire des pondérateurs négatifs $\lambda_i < 0$, Chauvet (1988) mentionne notamment :

- les agrégats de points, et en particulier les doublets de points,
- une portée élevée par rapport au domaine, ce qui conduit le krigeage à se rapprocher d'un ajustement polynomial et produit une alternance de pondérateurs de signes opposés,
- la régularité spatiale de la VR, un variogramme de type gaussien ayant tendance à produire des pondérateurs fortement négatifs.

Les agrégats de points entraînent des pondérateurs négatifs du fait d'un effet d'écran (Deutsch 1996). Lorsque la configuration n'est pas de type agrégé (*e.g.*, grille régulière), les pondérateurs négatifs révèlent des difficultés pour le krigeage à respecter l'hypothèse globale de stationnarité et/ou la régularité locale, ce qui se traduit par une instabilité de l'interpolation (Chauvet 1988). Les pondérateurs négatifs ont pour conséquences (Chauvet 1988) :

- de produire des estimations négatives lorsque les $\lambda_i < 0$ coïncident avec des valeurs élevées,
- de produire des estimations supérieures à la valeur maximale des données,
- de modifier de façon importante l'estimation lorsque la localisation des données est légèrement perturbée, cette sensibilité traduisant une instabilité de la procédure de krigeage.

Pour certains auteurs, la non-convexité du krigeage — capacité à produire des estimations en dehors de l'intervalle de variation des données — peut constituer un avantage (Journel 1986b). Cependant, les estimations négatives n'ont pas de sens lorsque la VR est définie dans $\mathbb{R}^+$, ce qui constitue le cas général en écologie (température, concentration, densité, etc.). En outre, les estimations doivent parfois appartenir à un intervalle précis, par exemple $[0, 1]$ s'il s'agit de l'estimation du degré d'appartenance à une classe floue, dans le cadre d'une cartographie continue des sols (de Gruijter *et al.* 1997, McBratney & Odeh 1997).

Le cas traité dans Aubry (1996b, pp. 25-27) constitue un exemple de situation propice à la production de pondérateurs négatifs. La VR considérée est la densité d'amandes du châtaignier *Castanea sativa* définie pour des quadrats de 0.5 m^2 installés sous un arbre. La population spatiale est définie par une grille de 18×18 quadrats discrétisant un domaine carré D (Aubry 1996b, Fig. 5.1a). L'échantillon est initialement constitué par une grille de 6×6 quadrats (Aubry 1996b, Fig. 5.1b). Afin de permettre l'estimation du

variogramme pour les faibles distances, nous avons ajouté 18 quadrats formant des doublets de quadrats voisins, répartis en quinconce (Aubry 1996b, Fig. 5.1c). L'échantillon comporte donc au total $n = 54$ quadrats, ce qui correspond à une intensité d'échantillonnage $f = 1/6$. Nous n'avons pas assimilé les quadrats à des supports quasi-ponctuels. En conséquence, le variogramme empirique a été dérégularisé (Section 4.4, p. 90), et utilisé pour estimer les densités des quadrats par krigeage ordinaire de blocs, en voisinage unique et en voisinage glissant (Aubry 1996b).

L'estimation locale produit des densités négatives extrêmes de -15.8 pour le krigeage en voisinage unique, et -12.9 pour le krigeage en voisinage glissant (Aubry 1996b, Tab. 5.1). Bien plus que les doublets de quadrats, c'est surtout la configuration très particulière de la répartition des amandes sous le châtaignier qui génère des pondérateurs négatifs. En effet, la présence de valeurs nulles au centre — correspondant au tronc — et aux bords de D perturbe la procédure de KO qui tente d'introduire une dérive absente du modèle et réagit en produisant des pondérateurs négatifs. Il en découle des valeurs négatives qui correspondent généralement à des densités nulles mal estimées. Ainsi, dans cette étude de cas, l'utilisation d'un modèle à dérive constante ne constitue pas la meilleure option. Le recours à un voisinage glissant ne fait que diminuer légèrement l'amplitude des valeurs nulles et ne résoud pas le problème.

Cela étant, notre objectif était celui de l'estimation globale, et la présence d'estimations locales négatives, bien que sémantiquement incorrecte, n'a pas de conséquence sur la qualité de l'estimation globale (Aubry 1996b, pp. 26-27). En effet, si la configuration des données est homogène (*e.g.*, résultant d'un échantillonnage systématique), l'estimation est globalement non biaisée (Chauvet 1988). Ainsi, si certaines estimations locales sont négatives (sous-estimation) cela signifie qu'en d'autres zones du domaine on peut avoir des surestimations, et finalement un rééquilibrage global (Chauvet 1988). Une mise à zéro des prédictions négatives, outre qu'elle escamotte le problème et évite une discussion critique riche en enseignements, biaiserait l'estimation globale (Chauvet 1988). En revanche, dans un contexte de cartographie, il convient de remédier aux estimations locales négatives.

Il existe plusieurs approches possibles pour empêcher les pondérateurs négatifs. D'après Chauvet (1988), une approche judicieuse consiste à modifier le modèle en introduisant une composante non stationnaire — soit implicitement avec les FAI-k, soit explicitement avec le modèle de KU — ce qui évite de perdre des propriétés intéressantes du krigeage.

Dans le cas d'un effet d'écran à l'origine des pondérateurs négatifs, il serait plus judicieux de modifier la définition du voisinage (Chauvet 1998) plutôt que de corriger *a posteriori* les pondérateurs (Deutsch 1996). En effet, bien qu'elle ait le mérite d'être simple, une telle correction *ad hoc* se traduit pas une estimation globale biaisée et une augmentation de la variance de krigeage.

D'autres auteurs proposent d'introduire *a priori* des contraintes de positivité sur les pondérateurs (Barnes & Johnson 1984, Limic & Mikelic 1984, *op. cit.* Dubrule & Kostov 1986), ce qui est une condition suffisante pour éviter les estimations négatives, mais pas nécessaire. Cette approche très pragmatique résulte d'une incompréhension du problème (Journel 1986b), et conduit à réduire la classe des prédicteurs autorisés (Chauvet 1988).

La modification du krigeage qui semble la plus fondée consiste à faire porter la contrainte sur les estimations elles-même plutôt que sur les pondérateurs. Ainsi, la régression

multiple sous contraintes linéaires (Mallet 1980) pourraît être appliquée au krigeage dans la mesure où le krigeage est une régression multiple de données autocorrélées (Dubrule & Kostov 1986, Journel 1986b). Comme la théorie des splines fournit la solution de l'interpolation sous contrainte d'égalité ou d'inégalité et qu'il existe un lien étroit entre les splines et le krigeage sous forme duale, Dubrule & Kostov (1986) proposent d'étendre cette solution au krigeage. Cependant, la méthode exposée par Kostov & Dubrule (1986) nécessite de recourir à la *programmation quadratique*[8], ce qui complique singulièrement la procédure de krigeage.

Enfin, Barnes & You (1992) exposent une procédure *ad hoc* forçant chaque valeur interpolée à respecter des bornes inférieure et supérieure. Cependant, la contrainte d'intervalle qui est introduite a un effet strictement local — *i.e.* n'influence pas l'interpolation en d'autres points — ce qui risque d'introduire un biais global.

6.3 Précision des estimations spatiales

L'étape d'estimation spatiale est souvent un préalable obligé dans une prise de décision d'enjeu scientifique, environnemental, ou économique. Dans ces conditions, associer une mesure de précision à une estimation spatiale constitue une étape critique (Olsen 1994). Ainsi, il convient de substituer à l'*estimation ponctuelle* — au sens de *estimation par une valeur unique* — une estimation par intervalle, la largeur de cet intervalle correspondant à la précision de l'estimation.

La définition opératoire de l'intervalle d'estimation dépend du cadre inférentiel utilisé (Robinson 1982). Dans l'approche statistique classique, l'estimation porte sur un paramètre θ d'une population infinie modélisée par une variable aléatoire Z. Un estimateur $\widehat{\theta}$ est construit conditionnellement aux valeurs d'un échantillon de taille n, considéré comme un échantillon aléatoire simple, ce qui conduit à la définition d'un intervalle de confiance à $100(1-\alpha)$ %, autrement dit, un intervalle $[a,b]$ tel que sa probabilité de contenir θ s'écrit :

$$\Pr\left(a < \theta < b\right) = 1 - \alpha \tag{6.57}$$

avec α la probabilité que l'intervalle ne contienne pas θ, ou risque d'erreur.

A partir du moment où l'on substitue l'estimation par intervalle à l'estimation ponctuelle, la question de savoir si l'estimation ponctuelle s'effectue avec ou sans biais est relativement secondaire, et il s'avère surtout intéressant d'obtenir une estimation la plus précise possible, ce qui se traduit, pour α constant, par un intervalle de confiance le plus étroit possible.

Le calcul d'un intervalle de confiance nécessite de connaître $f(\widehat{\theta} \mid \theta)$, *i.e.* la fonction densité de probabilité de l'estimateur $\widehat{\theta}$ conditionnellement à la vraie valeur θ. Du fait de l'échantillonnage aléatoire simple et du théorème de la limite centrale, la loi normale $\mathcal{N}(0,1)$ est généralement utilisée comme loi de probabilité pour construire un intervalle de confiance $[a,b]$ à $100(1-\alpha)$ % selon :

$$\left[\widehat{\theta} - \sigma_{\widehat{\theta}} \cdot G_{1-\frac{\alpha}{2}}, \widehat{\theta} + \sigma_{\widehat{\theta}} \cdot G_{1-\frac{\alpha}{2}}\right] \tag{6.58}$$

[8] La *programmation quadratique* est une branche de la *programmation mathématique* dédiée à la minimisation de fonctions quadratiques sous contraintes linéaires d'égalité ou d'inégalité (Karmanov 1995).

avec $\sigma_{\widehat{\theta}}$ l'écart-type de $\widehat{\theta}$ et $G_{1-\frac{\alpha}{2}}$ la valeur critique de la loi normale centrée-réduite pour la probabilite $1-\frac{\alpha}{2}$. Lorsque n est petit, la loi de Student est généralement substituée à la loi normale, bien qu'en toute rigueur cette pratique ne soit justifiée que si la VA Z est elle-même distribuée selon une loi normale.

De façon similaire à l'approche statistique classique, le problème du calcul d'un intervalle pour une estimation spatiale peut en général se décomposer en :

- un calcul de variance de l'estimateur,
- une hypothèse distributionnelle concernant l'estimateur.

Dans ce qui suit, en considérant une variable régionalisée quantitative, nous examinons essentiellement le problème de l'estimation par intervalle de la moyenne globale, mais abordons également succinctement la question de la précision des estimations locales.

6.3.1 Estimation globale

Soit $z(\cdot)$ une variable régionalisée quantitative définie sur un domaine $D \subset \mathbb{R}^2$ fini. Considérons pour simplifier le cas de supports quasi-ponctuels $x \in D$. A partir d'un échantillon spatial $s = \{x_i \mid i = 1, \ldots, n\}$, la moyenne globale z_D définie comme l'intégrale d'espace (6.1) est estimée par la combinaison linéaire :

$$z_D^* = \sum_{i=1}^{n} \lambda_i z(x_i) \tag{6.59}$$

avec $\sum \lambda_i = 1$. Pour simplifier la présentation, nous considérons uniquement la moyenne arithmétique $\bar{z}$, qui correspond au cas particulier $\lambda_i = n^{-1}$ pour $i = 1, \ldots, n$. L'estimation par intervalle de z_D nécessite de considérer simultanément plusieurs facteurs :

- le type d'échantillonnage ayant conduit à s,
- la structure d'autocorrélation de $z(\cdot)$,
- le cadre inférentiel choisi.

Dans ce qui suit, nous examinons les relations entre ces trois facteurs en considérant successivement plusieurs méthodes pour l'estimation par intervalle de la moyenne, puis nous présentons des études de cas à partir de données synthétiques et de données réelles relevant de l'écologie végétale.

6.3.1.1 Type d'échantillonnage

Parmi les différents dispositifs d'échantillonnage spatial, nous avons retenu trois dispositifs de base : l'échantillonnage aléatoire simple (EAS), l'échantillonnage systématique selon une grille, à départ aléatoire (ES), et l'échantillonnage aléatoire stratifié par une grille, à un élément par strate (STR), qui représente un compromis entre les deux précédents (Section 5.1.3.3). Il faut ajouter à ces dispositifs l'échantillonnage non probabiliste (ou préférentiel) qui ne fait pas appel à une règle de sélection mettant en jeu un tirage aléatoire, susceptible d'être appliquée dans D plusieurs fois et indépendamment.

6.3.1.2 Structure d'autocorrélation

Les variables régionalisées qui justifient d'un traitement spécifique sont bien évidemment celles qui présentent une structure d'autocorrélation spatiale. Néanmoins, pour que notre discussion soit complète, nous considérons également le cas limite des VR ne présentant par de structure d'autocorrélation, désignées dans la théorie de l'échantillonnage comme des *populations en ordre aléatoire* (*populations in random order*).

6.3.1.3 Cadre inférentiel

Le choix d'un cadre inférentiel est étroitement lié au type d'échantillonnage et à la structure d'autocorrélation spatiale. Pour des VR spatialement autocorrélées, le Tableau 6.1 indique l'applicabilité de l'approche *design-based*, de l'approche *model-based*, et d'une approche intermédaire entre les deux précédentes, en fonction du type d'échantillonnage. L'approche *design-based* est inapplicable si un échantillonnage probabiliste est inapplicable et l'approche *model-based* est inapplicable si la construction d'un modèle est impossible par manque de données (de Gruijter & ter Braak 1990). Compte tenu des types d'inférence envisagés ici, il convient de distinguer l'estimation :

- par intervalle de confiance de $\bar{z}$,
- par intervalle de prédiction de z_D.

Nous parlerons d'*intervalle de confiance* dans le cas de l'inférence de type *design-based* et d'*intervalle de prédiction* dans le cas de l'approche *model-based*. La réplication du dispositif d'échantillonnage — qui constitue la source de stochasticité dans l'inférence *design-based* — a pour conséquence que l'intervalle de confiance est lui-même une variable aléatoire. En revanche, dans le cas de l'inférence *model-based*, le modèle et le motif d'échantillonnage sont fixés, de sorte que l'intervalle de prédiction est unique. Dans le cas de l'approche intermédiaire entre l'inférence *design-based* et *model-based* que nous décrivons plus loin, nous utiliserons le terme d'*intervalle d'estimation* qui est dépourvu de connotation quant au type d'inférence.

	Echantillonnage		
	probabiliste		non probabiliste
Approche	EAS	ES ou STR	préférentiel
design-based	immédiate	problématique	inapplicable
intermédiaire	adaptée	adaptée	inapplicable
model-based	adaptée	adaptée	adaptée

Tableau 6.1: Applicabilité de trois types d'inférences statistiques en fonction du type d'échantillonnage, dans le cas d'une variable régionalisée spatialement autocorrélée.

6.3.1.4 Approche design-based

Dans l'approche *design-based*, la population est fixée de sorte que la moyenne globale z_D constitue un *paramètre populationnel*. Du fait de la réplication du dispositif d'échantillonnage $d = (S_d, P_d)$, une valeur estimée $\widehat{z}_D$ calculée à partir d'un échantillon particulier est une réalisation d'un estimateur $\widehat{Z}_D$ d'espérance $\mathrm{E}_p\left[\widehat{Z}_D\right]$ et de variance $\mathrm{Var}_p\left[\widehat{Z}_D\right]$. Toujours dans le cadre de la réplication du dispositif, le p-biais de l'estimateur $\widehat{Z}_D$ est défini comme l'erreur d'estimation moyenne :

$$\mathrm{B}_p\left[z_D\right] = \mathrm{E}_p\left[\widehat{Z}_D\right] - z_D = \mathrm{E}_p\left[\widehat{Z}_D - z_D\right] \tag{6.60}$$

et la précision de l'estimateur est mesurée au moyen de l'erreur quadratique moyenne ou p-MSE (*Mean Squared Error*) :

$$\mathrm{MSE}_p\left[\widehat{Z}_D\right] = \mathrm{E}_p\left[\left(\widehat{Z}_D - z_D\right)^2\right] = \mathrm{Var}_p\left[\widehat{Z}_D\right] + \left\{\mathrm{B}_p\left[z_D\right]\right\}^2 \tag{6.61}$$

L'estimateur $\widehat{Z}_D$ est dit sans p-biais si $\mathrm{E}_p\left[\widehat{Z}_D\right] = z_D$, et dans ce cas, la p-MSE (6.61) est égale à la variance de l'estimateur $\mathrm{Var}_p\left[\widehat{Z}_D\right]$. La distribution d'échantillonnage associée à $\widehat{Z}_D$, ou p-distribution (Smith 1976, Brus & de Gruijter 1997), est caractérisée notamment par sa p-espérance $\mathrm{E}_p\left[\widehat{Z}_D\right]$ et sa p-variance $\mathrm{Var}_p\left[\widehat{Z}_D\right]$.

Dans l'approche *design-based*, les probabilités d'inclusion (Section 5.1.1, p. 103) des unités d'échantillonnage dans l'échantillon permettent de construire des estimateurs dits de Horvitz-Thompson (EHT) (Horvitz & Thompson 1952, Hedayat & Sinha 1991). Dans ce cadre, la moyenne globale z_D est estimée sans p-biais à partir d'un échantillon par :

$$\widehat{z}_D = \frac{1}{N}\sum_{i=1}^{n}\frac{z_i}{\pi_i} \tag{6.62}$$

et l'estimateur sans p-biais de la p-variance de $\widehat{Z}_D$ s'écrit (Hedayat & Sinha 1991) :

$$\widehat{\mathrm{Var}_p}\left[\widehat{Z}_D\right] = \sum_i z_i^2\left[\frac{1}{\pi_i} - 1\right] + \sum_{i\neq j} z_i z_j\left[\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right] \tag{6.63}$$

$$= \sum_{i<j}\left(\pi_i\pi_j - \pi_{ij}\right)\left[\frac{z_i}{\pi_i} - \frac{z_j}{\pi_j}\right]^2 \tag{6.64}$$

avec $\pi_{ij} > 0$ pour tout couple d'unités (u_i, u_j). L'EHT inclut à la fois des caractéristiques du dispositif d'échantillonnage à travers les probabilités d'inclusion du premier ordre (π_i) et du second ordre (π_{ij}), et des propriétés de l'échantillon lui-même à travers les valeurs $\{z_i \mid i = 1, \ldots, n\}$. Ainsi, l'EHT ne dépend pas des caractéristiques spatiales telles que la géométrie des supports, leurs positions relatives, leur implantation dans D, ou la structure d'autocorrélation spatiale de $z(\cdot)$.

Considérons que $\widehat{z}_D$ est la moyenne arithmétique de l'échantillon, notée $\bar{z}$. Avant même d'examiner le calcul de l'intervalle de confiance de $\bar{z}$ d'un point de vue théorique, il faut s'attendre à ce que l'application de dispositifs différents conduise à différentes p-distributions. Afin d'illustrer ce point, nous choisissons trois variables régionalisées $z(x, y)$ définies dans un carré unité 1×1 dont les côtés sont parallèles aux axes des coordonnées, avec x l'abscisse et y l'ordonnée d'un support ponctuel. La variation spatiale des trois VR est décrite par trois surfaces analytiques :

- une fonction isotrope désignée abusivement par la suite *demi-sphère* parce qu'elle fait intervenir une demi-sphère de diamètre $2r = 1$ centrée dans D (Fig. 6.1.a) :

$$z(x, y) = \begin{cases} 0 & \text{si } d \leq 0 \\ \sqrt{d} & \text{sinon} \end{cases} \tag{6.65}$$

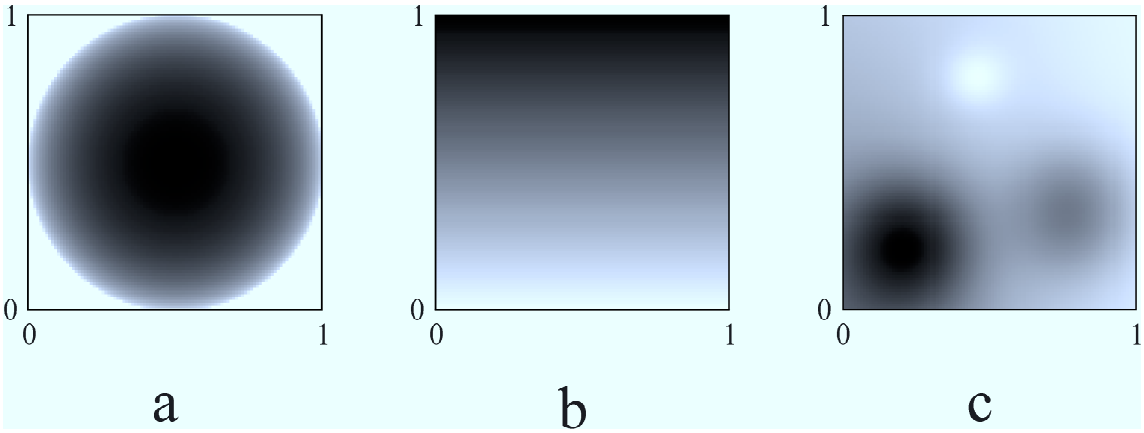
  avec $d = r^2 - (x - r)^2 - (y - r)^2$,

- un gradient linéaire (Fig. 6.1.b) :

$$z(x, y) = y \tag{6.66}$$

- une somme d'exponentielles simulant un paysage vallonné (Fig. 6.1.c) :

$$\begin{aligned} z(x, y) \quad = \quad & \frac{3}{4} e^{-\left[(9x-2)^2+(9y-2)^2\right]/4} + \frac{3}{4} e^{-(9x+1)^2/49-(9y+1)/10} + \\ & \frac{1}{2} e^{-\left[(9x-7)^2+(9y-3)^2\right]/4} - \frac{1}{5} e^{-(9x-4)^2-(9y-7)^2} \end{aligned} \tag{6.67}$$



Figure˜6.1: Images de trois variables régionalisées décrites par trois surfaces analytiques dans un domaine 1×1. (a) Demi-sphère. (b) Gradient linéaire. (c) Somme d'exponentielles.

Pour chaque variable régionalisée $z(x, y)$ définie sur un domaine D carré de côté $L = 1$, la moyenne globale z_D se calcule comme la double intégrale :

$$z_D = \int_0^1 \int_0^1 z(x, y)\, dxdy \tag{6.68}$$

ce qui donne :

- $z_D = \frac{\pi}{12} \simeq 0.2618$ pour la demi-sphère (6.65),
- $z_D = \frac{1}{2}$ pour le gradient linéaire (6.66),
- $z_D \simeq 0.4069$ pour la somme d'exponentielles (6.67).

L'échantillonnage est ponctuel, ce qui implique que la population est infinie. Trois échantillons sont construits pour chaque surface :

- échantillon aléatoire simple de $n = 100$ points (EAS),
- échantillon systématique selon une grille régulière 10×10, à départ aléatoire (ES),
- échantillon aléatoire stratifié par une grille régulière 10×10 centrée dans D, à un élément par strate (STR).

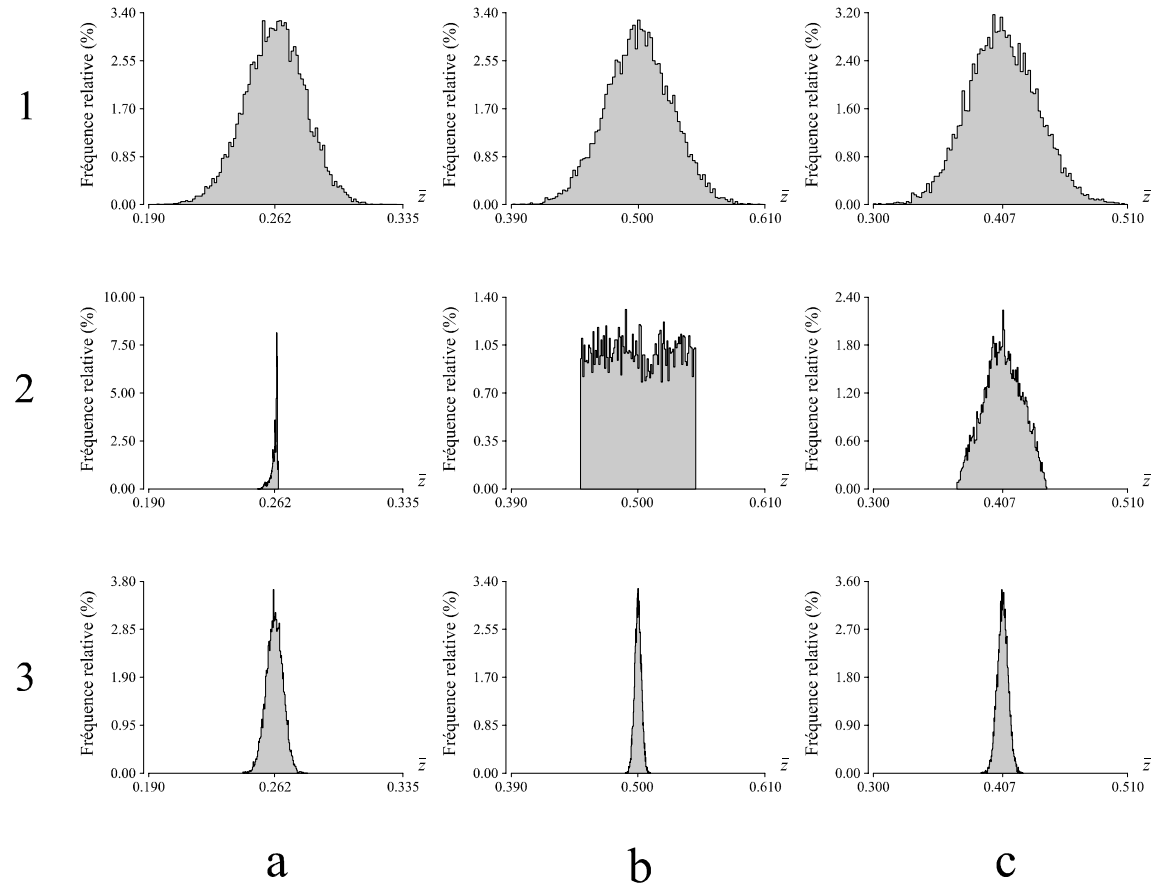
Pour chaque point (x, y), nous calculons $z(\cdot)$ en fonction de chaque modèle analytique, et pour chaque ensemble de 100 valeurs, nous calculons la moyenne associée $\bar{z}$. Chaque dispositif d'échantillonnage est répliqué 10^4 fois sur chaque surface, *i.e.* en tirant 10^4 EAS indépendants, en randomisant 10^4 fois l'origine de la grille de l'ES, et en tirant 10^4 STR indépendants. La p-distribution de $\bar{Z}$ est approximée par la distribution des 10^4 moyennes d'échantillons $\bar{z}$ (Fig. 6.2).

Dans le cas de l'EAS et du STR, l'asymétrie et l'aplatissement des p-distributions témoignent d'une distribution asymptotique gaussienne[9], indépendamment de la VR (Tab. 6.2). En revanche, dans le cas de l'ES, la p-distribution n'est pas gaussienne et dépend de la structure spatiale de la VR. Ainsi, la p-distribution est fortement asymétrique pour la demi-sphère, rectangulaire pour le gradient, et triangulaire pour la somme d'exponentielles (Fig. 6.2). Par ailleurs, dans le cas d'une population finie, Madow & Madow (1944) précisent que le nombre d'échantillons distincts que l'on peut construire selon un dispositif systématique est d'ordinaire trop faible pour que l'on puisse supposer la normalité. Ainsi, dans le cas des populations finies spatialement autocorrélées échantillonnées par ES, il n'est jamais possible de faire référence à la distribution gaussienne.

Dans tous les cas, les trois dispositifs implémentés sont sans p-biais (Tab. 6.2). La précision du STR s'avère toujours plus grande que celle de l'EAS (Tab. 6.2). En revanche, les performances de l'ES par rapport à l'EAS et au STR dépendent de la structure spatiale de la VR. L'ES est extrêmement précis dans le cas de la demi-sphère, et plus précis que le STR, mais il est moins précis que le STR dans le cas de la somme d'exponentielles, et il s'avère même moins précis que l'EAS dans le cas du gradient (Fig. 6.2, Tab. 6.2).

Cette expérience de Monte-Carlo montre clairement que la p-distribution dépend du dispositif d'échantillonnage et de la structure spatiale de la VR et qu'en conséquence, il ne peut pas exister de procédure *design-based* universelle pour calculer l'intervalle de confiance de $\bar{z}$.

[9] La forme d'une distribution Gaussienne est caractérisée notamment par une asymétrie $\beta_1 = 0$ et un aplatissement $\beta_2 = 3$.



Figure~6.2: Approximation de la p-distribution de $\bar{Z}$ par la distribution de 10^4 moyennes d'échantillons $\bar{z}$. (1) Dispositif EAS. (2) Dispositif ES. (3) Dispositif STR. (a) Demi-sphère. (b) Gradient linéaire. (c) Somme d'exponentielles.

	Demi-sphère			Gradient linéaire			Somme d'exponentielles		
	EAS	ES	STR	EAS	ES	STR	EAS	ES	STR
E_p	0.2619	0.2618	0.2618	0.5004	0.5000	0.5000	0.4068	0.4069	0.4069
Var_p	2.9	0.0	0.2	8.2	8.3	0.1	8.2	2.4	0.2
β_1	-0.1	-1.6	-0.1	-0.0	0.0	-0.0	0.1	-0.0	-0.0
β_2	3.0	5.5	3.0	2.9	1.8	2.9	3.0	2.4	3.1

Tableau 6.2: Moyennes, variances, asymétries (β_1) et aplatissements (β_2) des p-distributions des échantillonnages EAS, ES et STR, pour la demi-sphère, le gradient linéaire et la somme d'exponentielles. Les variances doivent être multipliées par un facteur 10^{-4}.

Echantillonnage aléatoire simple Les probabilités d'inclusion de l'EAS s'écrivent (Hedayat & Sinha 1991, p. 70) :

$$\pi_i = \frac{n}{N} \tag{6.69}$$

$$\pi_{ij} = \frac{n\,(n-1)}{N\,(N-1)} \tag{6.70}$$

la moyenne de la population est estimée sans p-biais par $\bar{z}$, et la p-variance associée peut être estimée sans p-biais par (Cochran 1977, Hedayat & Sinha 1991) :

$$\widehat{\mathrm{Var}_p}\left[\bar{Z}\right] = \left[\frac{1}{n} - \frac{1}{N}\right] s^2 \tag{6.71}$$

$$= \frac{1}{n} s^2 \left[1 - \frac{n}{N}\right] \tag{6.72}$$

avec s^2 l'estimateur non biaisé de la variance de la population et $(1 - n/N)$ la *correction pour population finie*, négligeable si n/N est petit (*e.g.*, $n/N < 0.05$).

La nature de l'EAS permet de faire intervenir le théorème de la limite centrale (Särndal 1978), de sorte que la p-distribution associée à $\bar{Z}$ suit la loi normale[10] et qu'un intervalle de confiance peut être construit comme en (6.58). Cette procédure est valide :

- indépendamment de la distribution statistique des valeurs de $z(\cdot)$, en vertu du théorème de la limite centrale,
- indépendamment de l'autocorrélation spatiale (de Gruijter & ter Braak 1990, Brus & de Gruijter 1993, 1997), parce que la p-distribution ne dépend pas de la structure de la population sous-jacente (Smith 1976, Cochran 1977, Hansen *et al.* 1983).

Le caractère universel de l'EAS est très intéressant sur le plan théorique, mais il se traduit généralement par des estimations peu précises parce que :

- aucune information concernant la structure de la population n'est exploitée,
- les points générés par l'EAS sont souvent agrégés, ce qui entraîne une redondance d'information d'autant plus grande que la VR est spatialement régulière.

Echantillonnage systématique Les probabilités de sélection de premier ordre π_i sont les mêmes que dans le cas de l'EAS, et par conséquent, la moyenne de la population est estimée sans p-biais par $\bar{z}$ (Cochran 1977, Hedayat & Sinha 1991). Dans le cas d'une grille, seule l'origine est susceptible d'être sélectionnée au hasard, toutes les autres unités figurant dans l'échantillon étant espacées régulièrement à partir de l'origine, *i.e.* de façon non indépendante. D'une façon générale, Koop (1990) signale que l'on ne peut plus obtenir d'estimateur de la variance qui soit non biaisé dès qu'il y a alignement des unités ou un caractère systématique.

Dans le cadre *design-based*, un seul échantillon ne permet pas d'estimer sans biais la variance parce que le dispositif n'assure pas l'inclusion de toutes les combinaisons

[10] En toute rigueur, la p-distribution est une loi normale si la population est infinie. Pour des populations finies, nous utilisons la loi normale comme approximation asymptotique parce qu'en pratique $\binom{\mathsf{N}}{\mathsf{n}}$ est assimilable à l'infini (Madow & Madow 1944).

possibles de paires d'unités (Scherrer 1983, Hedayat & Sinha 1991). En effet, le dispositif systématique a pour conséquence que beaucoup de probabilités d'inclusion de second ordre sont nulles (Cassel *et al.* 1977, Hedayat & Sinha 1991). Or il n'est pas possible d'obtenir un estimateur sans biais pour la variance de $\bar{Z}$ si $\pi_{ij} = 0$ pour certains couples d'unités (u_i, u_j) (Hedayat & Sinha 1991, p. 51).

Du point de vue des statisticiens, les probabilités $\pi_{ij} = 0$ interdisent d'utiliser l'estimateur de l'EAS (6.71) (Horvitz & Thompson 1952). Cette pratique conduirait en effet à un *biais de sélection* en ce que les probabilités π_{ij} utilisées pour construire l'estimateur ne correspondent pas à celles du dispositif (Kotz & Johnson 1982), puisque les probabilités jointes sont $\pi_{ij} = n\,(n-1)\,/N\,(N-1) \simeq \pi_i \pi_j$ pour l'EAS, tandis que pour l'ES on a souvent $\pi_{ij} = 0$. Cependant, les écologistes utilisent l'estimateur (6.71) par défaut, notamment en halieutique où l'échantillonnage systématique par transects parallèles est largement utilisé (Murray 1996, Malinen & Peltonen 1996). Cette pratique est du reste encouragée par certains textes insuffisamment précis sur la question (*e.g.*, Scherrer 1983), ou trop optimistes (*e.g.*, Ripley 1981, p. 27).

Dans le cas où $z\,(\cdot)$ n'est pas spatialement autocorrélée, autrement dit dans le cas d'une population en ordre aléatoire, le biais s'avère négligeable (Wolter 1984), et l'ES s'avère en moyenne équivalent à l'EAS (Madow & Madow 1944, Cochran 1977, p. 213, Hedayat & Sinha 1991, p. 248). En pratique, $z\,(\cdot)$ est spatialement autocorrélée, et la dépendance des supports de l'ES interagit avec la dépendance spatiale des valeurs de $z\,(\cdot)$. Dans ce contexte, l'utilisation de l'estimateur de l'EAS (6.71) peut conduire à un biais assez élevé (Osborne 1942, Dunn & Harrison 1993, Aubry & Debouzie 1999a). Ainsi, traiter un échantillon systématique comme s'il s'agissait d'un échantillon aléatoire (Milne 1959) n'est acceptable que si l'autocorrélation spatiale de la population peut être négligée.

En écologie, Greig-Smith juge que l'impossibilité de déterminer la variance d'échantillonnage de la moyenne est une objection fatale à l'utilisation de l'ES et en conséquence, Greig-Smith préfère utiliser l'échantillonnage aléatoire stratifié (*op. cit.* Maling 1989). Une autre attitude, très pragmatique, considère que les avantages pratiques de l'ES contrebalancent largement les difficultés d'ordre statistique (Maling 1989). Pour les statisticiens, le problème posé par l'ES n'est pas insurmontable puisque l'estimation de la variance associée à $\bar{Z}$ peut s'effectuer dans un cadre *design-based*, en utilisant notamment :

- la méthode de composition, *i.e.* en divisant l'échantillon en plusieurs sous-échantillons (Koop 1971, Wolter 1984, Hedayat & Sinha 1991),
- la méthode des départs aléatoires multiples suggérée par Madow & Madow (1944), *i.e.* en construisant plusieurs échantillons à partir d'origines tirées au hasard (Gautschi 1957, Hedayat & Sinha 1991). Cette seconde méthode est toutefois difficile à mettre en pratique (Quenouille 1949).

Enfin, en présence d'autocorrélation spatiale, et en admettant même qu'un estimateur satisfaisant de la variance $\mathrm{Var}_p\left[\bar{Z}\right]$ ait été construit, la question de la distribution de référence à employer pour calculer un intervalle de confiance reste posée puisque l'utilisation de la loi normale n'est supportée par aucun argument théorique, du moins dans le cadre de l'échantillonnage des populations finies[11].

[11] En utilisant un modèle de superpopulation unidimensionnel, Iachan (1983) démontre la normalité asymptotique de la distribution de la moyenne, *i.e.* pour $N, n \to \infty$.

Echantillonnage aléatoire stratifié Considérons le cas général où D est partitionné en L strates D_h comptant $\{N_h \mid h = 1, \ldots, L\}$ unités, un échantillon aléatoire stratifié est constitué en tirant au hasard n_h unités parmi les N_h de chaque strate D_h. Les probabilités d'inclusion s'écrivent alors (Hedayat & Sinha 1991, p. 261) :

$$\pi_i = \frac{n_h}{N_h} \tag{6.73}$$

$$\pi_{ij} = \begin{cases} n_h (n_h - 1) / N_h (N_h - 1) & \text{si } (u_i, u_j) \in D_h \\ n_h n_s / N_h N_s & \text{si } u_i \in D_h \text{ et } u_j \in D_s \end{cases} \tag{6.74}$$

En notant $W_h = N_h/N$, la moyenne de la population est estimée par (Cochran 1977, Hedayat & Sinha 1991) :

$$\widehat{z}_D = \sum_{h=1}^{L} W_h \bar{z}_h \tag{6.75}$$

avec $\bar{z}_h$ la moyenne des n_h valeurs échantillonnées dans la strate D_h. La p-variance de l'estimateur (6.75) s'écrit (Cochran 1977, Hedayat & Sinha 1991) :

$$\widehat{\mathrm{Var}_p}\left[\widehat{Z}_D\right] = \sum_{h=1}^{L} W_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) s_h^2 \tag{6.76}$$

avec s_h^2 l'estimateur de la variance dans la strate D_h :

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (z_{hi} - \bar{z}_h) \tag{6.77}$$

défini si $n_h \geq 2$ pour tout h.

Lorsque D est stratifié *a priori* par n strates comptant chacune une seule unité échantillonnée, l'estimateur (6.75) est la moyenne arithmétique $\bar{z}$, mais s_h^2 n'est pas défini, et par conséquent l'estimateur (6.76) n'est pas non plus défini. En outre, comme $\pi_{ij} = 0$ lorsque $(u_i, u_j) \in D_h$, il n'est pas possible de construire un estimateur de la variance qui soit non biaisé. L'échantillonnage aléatoire stratifié par une grille, à un élément par strate (STR), pose donc des difficultés similaires à l'échantillonnage systématique[12] (Matérn 1960, Ripley 1981). En revanche, contrairement à l'ES, le recours à la loi normale pour calculer l'intervalle de confiance de $\bar{z}$ est théoriquement fondé, en vertu du théorème de la limite centrale.

Echantillonnage non probabiliste Afin qu'il n'y ait pas d'erreur d'estimation, le dispositif probabiliste idéal devrait assurer que les probabilités d'inclusion soient proportionnelles aux valeurs de $z(\cdot)$, ce qui impliquerait paradoxalement de connaître la population *a priori* (Horvitz & Thompson 1952, Hedayat & Sinha 1991). La stratification utilisant une variable auxiliaire liée à la variable étudiée tente de s'approcher de cet idéal, tout en restant dans le cadre de l'échantillonnage probabiliste (Cochran 1977, Hedayat & Sinha 1991). A l'autre extrême, autrement dit en ignorant tout de la population,

[12] C'est peut être pour cette raison que Legendre & McArdle (1997) considèrent, de façon abusive, que l'échantillonnage stratifié à un élément par strate fait partie des dispositifs systématiques.

l'EAS donne la même probabilité d'inclusion à toutes les unités d'échantillonnage, et par conséquent, la même probabilité de sélection à tous les échantillons possibles.

Lorsqu'un écologiste met en place un motif d'échantillonnage préférentiel, c'est en exploitant au mieux sa connaissance du phénomène étudié et son expérience du terrain, de façon heuristique. L'exploitation de cette expertise, bien qu'étant rarement formalisée, tend vers l'idéal de l'échantillonnage probabiliste de sorte que les estimations qui en résultent peuvent s'avérer très précises. Mais dans le cadre *design-based*, la précision des estimations ne peut pas être calculée selon la procédure classique définie pour l'EAS (Voltz *et al.* 1997), et la pratique qui consiste à traiter des échantillons non aléatoires comme s'ils étaient aléatoires n'est pas excusable (Cochran *et al.* 1954). Ceci nous conduit à considérer un mode d'inférence valide lorsque le motif d'échantillonnage n'est pas aléatoire : l'approche *model-based* (Cox *et al.* 1997).

6.3.1.5 Approche model-based

En présence d'autocorrélation spatiale, l'inférence *design-based* s'avère problématique en ce qui concerne les dispositifs ES et STR, et inapplicable dans le cas de l'échantillonnage préférentiel. Comme l'inférence *model-based* au sens strict ne tient pas compte d'un éventuel dispositif d'échantillonnage mais uniquement du motif d'échantillonnage, toute procédure construite dans ce cadre s'applique de façon universelle, que l'échantillon ait été obtenu par EAS, ES, STR, ou par échantillonnage préférentiel.

Dans l'approche *model-based*, l'estimation par intervalle de z_D nécessite de construire un modèle qui prenne en compte autant de caractéristiques de la population qu'il est possible. Dans le cas de l'ES unidimensionnel, plusieurs estimateurs de type *model-based* ont été proposés (*e.g.*, Williams 1956, Heilbron 1978, Bellhouse & Sutradhar 1988), certains faisant appel à des modèles autorégressifs. Ces approches peuvent se généraliser à l'espace bidimensionnel, mais dans ce qui suit, nous considérons uniquement le modèle superpopulationnel proposé par la géostatistique, autrement dit une fonction aléatoire (FA). Evidemment, rien n'interdit d'envisager d'autres modèles que les FA (*cf.* Martin 1979, Haining 1988, Ord 1988), le principe étant d'acquérir le plus d'information possible à propos de la population, puis d'utiliser cette information pour construire des modèles plausibles de la population (Wolter 1984).

Dans la géostatistique *model-based*, la moyenne globale z_D est vue comme une réalisation d'une variable aléatoire Z_D de ξ-espérance $\mu_D = \mathrm{E}_\xi[Z_D]$ et de ξ-variance $\mathrm{Var}_\xi[Z_D]$. De la même façon, la valeur prédite z_D^* est vue comme la réalisation d'un prédicteur Z_D^* d'espérance $\mathrm{E}_\xi[Z_D^*]$ et de variance $\mathrm{Var}_\xi[Z_D^*]$. Toujours dans le cadre du modèle, le ξ-biais est défini comme l'erreur moyenne :

$$\mathrm{B}_\xi[\mu_D] = \mathrm{E}_\xi[Z_D^*] - \mu_D = \mathrm{E}_\xi[Z_D^*] - \mathrm{E}_\xi[Z_D] = \mathrm{E}_\xi[Z_D^* - Z_D] \tag{6.78}$$

et l'erreur quadratique moyenne ou ξ-MSE s'écrit :

$$\mathrm{MSE}_\xi[Z_D^*] = \mathrm{E}_\xi\left[(Z_D^* - \mu_D)^2\right] = \mathrm{Var}_\xi[Z_D^*] + \{\mathrm{B}_\xi[\mu_D]\}^2 \tag{6.79}$$

En pratique, nous ne cherchons pas à estimer l'espérance $\mu_D = \mathrm{E}_\xi[Z_D]$ parce que nous considérons que le modèle n'est pas immanent mais joue un rôle strictement opératoire[13]. C'est donc uniquement l'estimation spatiale de z_D qui fait sens, ce qui revient

[13] Le point de vue opposé est illustré notamment par Griffith *et al.* (1994).

à considérer l'erreur d'estimation $z_D^* - z_D$, *ipso facto* probabilisée par $Z_D^* - Z_D$. En faisant référence au modèle, la précision de l'estimation spatiale est évaluée par la variance $\mathrm{Var}_\xi [Z_D^* - Z_D]$ et non par la ξ-MSE (6.79). Si l'estimateur spatial[14] Z_D^* est sans ξ-biais ($\mathrm{E}_\xi [Z_D^* - Z_D] = 0$), alors $\mathrm{Var}_\xi [Z_D^* - Z_D] = \mathrm{E}_\xi \left[(Z_D^* - Z_D)^2\right]$ mais il ne s'agit pas pour autant d'une MSE au sens des définitions (6.61) ou (6.79), parce que Z_D n'est pas un paramètre mais une VA.

La population étant vue comme un échantillon aléatoire simple de la superpopulation (Smith 1976), la distribution associée à Z_D^* dans le modèle, ou ξ-distribution, peut être assimilée à une loi de Gauss caractérisée par l'espérance $\mathrm{E}_\xi [Z_D^*]$ et la variance $\mathrm{Var}_\xi [Z_D^*]$.

Dans le cadre de l'estimation par intervalle de la moyenne globale z_D de la VR étudiée, il convient d'identifier la VR en tant que réalisation particulière parmi toutes les réalisations que peut générer le modèle. Ainsi, nous notons $z^{(0)}(\cdot)$ la VR étudiée et $z_D^{(0)}$ sa moyenne globale. La fonction aléatoire $Z(\cdot)$ étant vue comme un outil mathématique sans équivalent dans le monde réel sous étude, l'espérance $\mathrm{E}_\xi [Z_D]$ n'est pas immanente et nous ne traitons pas de son estimation. Nous nous intéressons uniquement à la prédiction par intervalle de la valeur réelle $z_D^{(0)}$, aussi nous décidons de construire le modèle en identifiant l'espérance de Z_D et $z_D^{(0)}$, soit formellement :

$$z_D^{(0)} \equiv \mathrm{E}_\xi [Z_D] \tag{6.80}$$

Avec une argumentation similaire, nous identifions le variogramme théorique de $Z(\cdot)$ et le variogramme local de $z^{(0)}(\cdot)$, soit formellement :

$$\gamma(h) = \mathrm{E}_\xi [\gamma_D(h)] \equiv \gamma_D^{(0)}(h) \tag{6.81}$$

Les identifications (6.80) et (6.81) justifient le calcul de l'intervalle de prédiction de $z_D^{(0)}$ d'après la ξ-distribution de $Z_D^* - Z_D$, autrement dit, en exploitant la fluctuation de $Z_D^* - Z_D$ autour de $\mathrm{E}_\xi [Z_D^* - Z_D]$. Une conséquence directe des identifications (6.80) et (6.81) est que nous n'avons pas besoin de faire intervenir la notion d'ergodicité : ce qui est uniquement requis pour la FA c'est un type de stationnarité compatible avec le type de variation spatiale de la VR (Journel 1985).

En pratique, le calcul de l'intervalle de prédiction s'effectue comme en (6.58) grâce à la variance $\mathrm{Var}_\xi [Z_D^* - Z_D]$ et à la loi de Gauss. La variance $\mathrm{Var}_\xi [Z_D^* - Z_D]$ est généralement non conditionnelle et correspond à σ_E^2. Afin de calculer l'intervalle de prédiction de $z_D^{(0)}$, il est souhaitable que la variance d'erreur d'estimation soit conditionnée par les valeurs de l'échantillon, ce qui revient à préférer la variance conditionnelle σ_C^2 à la variance d'erreur d'estimation classique σ_E^2 (Section 4.6.5). Dans le cas de la variance conditionnelle σ_C^2, l'échantillon $\left\{z^{(0)}(x_i) \mid i = 1, \ldots, n\right\}$ est un invariant, ce qui signifie que Z_D^* est une constante telle que $Z_D^* = z_D^{*(0)}$, d'où $\mathrm{Var}_\xi [Z_D^*] = 0$, ce qui conduit à l'égalité :

$$\mathrm{Var}_\xi \left[Z_D^* - Z_D \mid Z^{(0)}(x_i) = z^{(0)}(x_i) \;;\; i = 1, \ldots, n\right] = \mathrm{Var}_\xi [Z_D] \tag{6.82}$$

Dans le cadre des réalisations $\left\{z^{(\ell)}(\cdot) \mid \ell = 1, \ldots, L\right\}$ générées par simulation conditionnelle, l'égalité (6.82) est obtenue lorsque L est suffisamment grand (*e.g.*, $L = 10^4$).

[14] Si le modèle est utilisé pour déterminer Z_{D}^*, alors il conviendrait plutôt de parler de *prédicteur*.

6.3.1.6 Approche intermédiaire

Les approches *design-based* et *model-based* font appel à des sources de stochasticité différentes, mais il est possible de les combiner en un mode d'inférence mixte reposant sur une randomisation complète portant à la fois sur la population et sur l'échantillon. Par exemple, dans le cas où Z_D^* est la moyenne arithmétique et où l'échantillon est prélevé par EAS, on a (Matheron 1965, Aubry & Debouzie 1999a) :

$$\mathrm{E}_p\left[\mathrm{Var}_\xi\left[\bar{Z} - Z_D\right]\right] = \frac{1}{n}\,\mathrm{E}_\xi\left[\sigma^2\right] \tag{6.83}$$

avec $\mathrm{E}_\xi[\sigma^2]$ l'espérance dans le modèle de la variance de population σ^2 définie pour une réalisation $z(\cdot)$. En géostatistique, la ξ-espérance $\mathrm{E}_\xi[\sigma^2]$ est connue comme la *variance de dispersion* de $z(\cdot)$ dans D, notée $\sigma^2(0 \mid D)$ (Matheron 1965, Chauvet 1994). Dans le cas limite où la VR ne présente pas d'autocorrélation spatiale, alors $\mathrm{E}_\xi[\sigma^2] = \sigma^2$ (Matheron 1965), et il n'est pas utile d'utiliser une FA comme modèle probabiliste. Comme l'EAS est un dispositif non informatif (Section 5.5.2, p. 120), il est possible d'écrire :

$$\mathrm{E}_\xi\left[\mathrm{Var}_p\left[\bar{Z} - Z_D\right]\right] = \mathrm{E}_p\left[\mathrm{Var}_\xi\left[\bar{Z} - Z_D\right]\right] = \frac{1}{n}\sigma^2 \tag{6.84}$$

ce qui correspond à la p-variance de la moyenne arithmétique.

L'inférence basée sur la réplication du dispositif d'échantillonnage sur des réalisations conditionnelles constitue une voie de recherche intéressante (Journel 1994b). En effet, cette approche permettrait de bénéficier à la fois des avantages de l'inférence *design-based* et de ceux de l'inférence *model-based* (Iachan 1984, Urquhart 1997). Cependant, Journel (1994b) a recours à seulement $\eta_0 = 10$ réplications du dispositif d'échantillonnage sur $\eta = 100$ réalisations conditionnelles. En accord avec Urquhart (1997), $\eta = 100$ nous semble constituer un très petit nombre de réalisations. Nous estimons qu'il faut au moins $\eta_0 = \eta = 10^3$ et de préférence $\eta_0 = \eta = 10^4$ pour être confiant dans le résultat de la procédure, ce qui est actuellement impraticable. Aussi, nous avons préféré étudier une approche intermédiaire entre les inférences *design-based* et *model-based* au sens strict, afin de traiter le cas des dispositifs tels que l'ES et le STR pour lesquels le calcul de l'intervalle de confiance de $\bar{z}$ est problématique.

En considérant l'expérience de Monte-Carlo qui illustre la nature de la p-distribution, il apparaît évident qu'une façon d'approximer cette p-distribution peut s'effectuer :

- en modélisant de façon déterministe la VR aussi bien que possible à partir de l'échantillon,
- en répliquant le dispositif et en calculant les valeurs de la VR grâce au modèle.

Dans le cas des surfaces analytiques que nous avons utilisées dans l'expérience de Monte-Carlo, la modélisation de la VR peut s'effectuer par ajustement d'une surface de tendance polynomiale d'ordre k, et les valeurs de la VR sont calculées selon :

$$z^*(x, y) = \sum_{i+j=0}^{k} a_{ij} x^i y^j \tag{6.85}$$

avec a_{ij} les coefficients de la tendance.

Evaluons cette approche en considérant 10^4 réplications des dispositifs ES et STR, pour des échantillons de taille $n = 100$, et un modèle de surface de tendance de degré $k = 4$ ajusté à l'échantillon initial. Le cas du gradient linéaire n'est pas considéré parce que le modèle de surface est égal à $z(x, y) = y$. La comparaison entre la p-distribution de référence et l'approximation obtenue montre que les caractéristiques essentielles sont respectées et que les intervalles d'estimation sont très satisfaisants (Tab. 6.3, Fig. 6.3 & 6.4).

	Demi-sphère				Somme d'exponentielles			
	ES	ES*	STR	STR*	ES	ES*	STR	STR*
E_p	0.2618	0.2615	0.2618	0.2607	0.4069	0.4036	0.4069	0.4082
Var_p	3.4	2.3	22.1	15.8	241.5	174.9	19.6	23.4
a	0.2563	0.2582	0.2522	0.2536	0.3769	0.3760	0.3964	0.3956
b	0.2638	0.2638	0.2707	0.2692	0.4359	0.4230	0.4137	0.4145
Δ	0.0075	0.0056	0.0184	0.0156	0.0590	0.0470	0.0174	0.0189

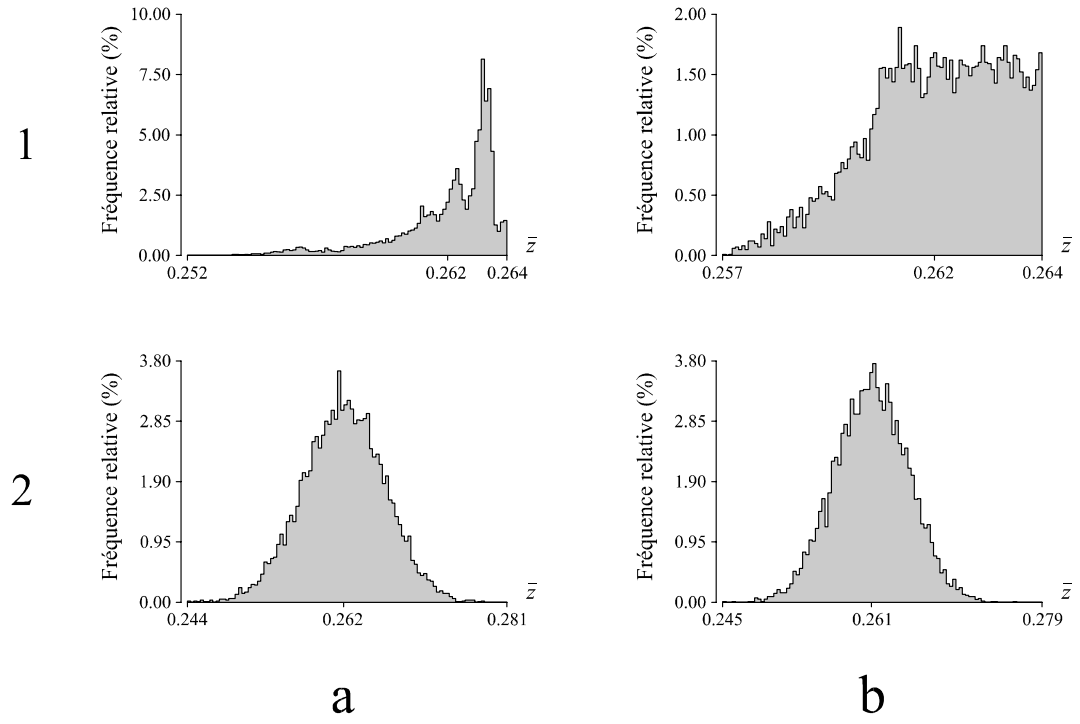
Tableau 6.3: Moyennes et variances des p-distributions de $\bar{Z}$ dans le cas de l'ES et du STR, pour la demi-sphère et la somme d'exponentielles. a, b, Δ : bornes et amplitude de l'intervalle de confiance. ES, STR : références. ES*, STR* : approximations. Les variances doivent être multipliées par un facteur 10^{-6}.

Bien évidemment, dans le cas d'une VR réelle, la variabilité spatiale ne permet pas de recourir uniquement à un modèle de surface de tendance, et par conséquent, un autre type de modèle doit être utilisé.

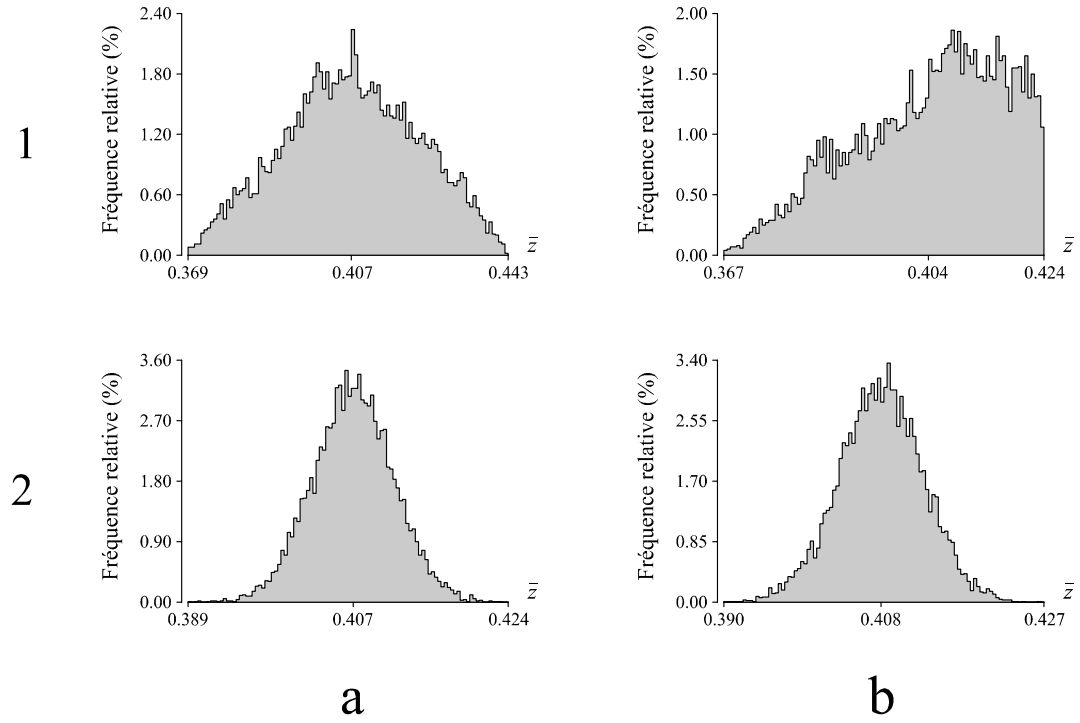
En géostatistique, le krigeage présente l'intérêt de tenir compte de la structure d'autocorrélation spatiale et de conditionner l'interpolation par les valeurs observées. En revanche, l'effet de lissage du krigeage ordinaire (KO) ou du krigeage simple (KS) constitue un inconvénient majeur puisque nous cherchons à obtenir un modèle qui reproduise au mieux la variabilité spatiale de la VR originelle. Dans ce contexte, il est préférable de substituer le krigeage simple modifié (KM) au KO ou au KS. Lorsque les supports sont quasi-ponctuels et que la population est infinie, il convient d'utiliser la forme duale du KM afin d'obtenir une fonction d'interpolation $z^*(x, y)$. Enfin, si la VR ne permet pas de justifier d'un modèle de FA du type FAST-2, il est possible de modéliser la dérive par une surface de tendance afin d'obtenir un résidu qui puisse être correctement modélisé par une FAST-2.

La méthode d'inférence de type Monte-Carlo que nous proposons ici est fondamentalement du type *design-based* puisque l'estimation par intervalle s'effectue grâce à un intervalle de confiance qui fait référence à une p-distribution. Cependant, contrairement à l'inférence *design-based* au sens strict, l'approche intermédiaire utilise une fonction d'interpolation déterministe obtenue en minimisant un critère dans le cadre d'un modèle géostatistique. Ainsi, l'intérêt potentiel de l'approche intermédiaire est qu'elle est :

- plus générale que l'approche *design-based* parce qu'elle évite les problèmes posés par le recours aux probabilités d'inclusion pour estimer la variance d'échantillonnage,
- moins subjective que l'approche *model-based* parce qu'elle exploite la réplication du dispositif d'échantillonnage.



Figure˜6.3: Résultats de l'approche intermédiaire appliquée à la demi-sphère. (1) Dispositif ES. (2) Dispositif STR. (a) p-distribution de référence. (b) p-distribution approximée (détails dans le texte).



Figure˜6.4: Résultats de l'approche intermédiaire appliquée à la somme d'exponentielles. (1) Dispositif ES. (2) Dispositif EAS. (a) p-distribution de référence. (b) p-distribution approximée (détails dans le texte).

En revanche, sa faiblesse vient de ce que l'inférence dépend entièrement de la capacité du modèle à reproduire la population, et notamment sa variabilité. Considérons la robustesse de cette approche en termes de limites de l'intervalle de confiance. D'une façon générale, le calcul de l'intervalle peut s'effectuer de façon non paramétrique à partir de l'approximation de la p-distribution. Si le nombre de réplications m est élevé ($m \gg 100$), les limites $\bar{z}_j$ et $\bar{z}_k$ de l'intervalle de confiance à $100(1-\alpha)$ % peuvent être déterminées à partir de l'ensemble ordonné des m moyennes d'échantillons $\{\bar{z}_1 \leq \bar{z}_2 \leq \ldots \leq \bar{z}_m\}$, en utilisant la méthode des pourcentages (*e.g.*, Buckland 1984) :

$$\begin{array}{l} j = (m+1)\frac{\alpha}{2} \\ k = (m+1)\left(1-\frac{\alpha}{2}\right) \end{array} \tag{6.86}$$

Deux facteurs peuvent être pris en considération en termes de robustesse des limites de l'intervalle de confiance :

- la variabilité introduite par le recours à une méthode de Monte-Carlo,
- la qualité du modèle de surface.

Variabilité introduite par la méthode de Monte-Carlo En invoquant une éventuelle symétrie de la p-distribution d'espérance μ, formellement :

$$f(y \mid \mu = x) = f(x \mid \mu = y) \quad \forall x, y \tag{6.87}$$

le niveau de confiance réel b de l'intervalle suit une loi bêta $\mathcal{B}(k-j, m-k+j+1)$ dont l'écart-type fournit une mesure de la variabilité introduite par la méthode de Monte-Carlo elle-même (Buckland 1984). Pour $m = 10^3$ et $\alpha = 0.05$, le niveau de confiance est $b \in [0.936, 0.964]$, et pour $m = 10^5$, le niveau de confiance est $b \in [0.946, 0.954]$ (Buckland 1984). Par la suite, m est jugé suffisamment grand pour que nous puissions négliger la variabilité introduite par la méthode de Monte-Carlo.

Qualité du modèle de surface Sans avoir besoin d'une étude de sensibilité, il est évident que les limites de l'intervalle calculées comme en (6.86) sont assez sensibles à la qualité du modèle de surface, ce qui constitue la principale faiblesse de l'approche intermédiaire que nous proposons.

Dans le cas du STR, la robustesse peut être augmentée en utilisant la p-distribution uniquement pour estimer la variance d'échantillonnage $\mathrm{Var}_p\left[\bar{Z}\right]$ et en calculant l'intervalle comme en (6.58). L'utilisation de la loi de Gauss n'est cependant utile que si le nombre de réplications n'est pas très élevé ($m \ll 10^4$). Dans le domaine du bootstrap, d'autres méthodes que celle des pourcentages ont été proposées afin de produire de meilleurs intervalles de confiance lorsque la distribution est asymptotiquement normale, notamment la méthode *bias-corrected and accelerated* (Efron & Tibshirani 1993, pp. 178-188).

Dans le cas de l'ES d'une population infinie, le calcul de l'intervalle s'effectue de façon non paramétrique par la méthode des pourcentages (6.86). Si la population est finie, $\mathrm{Card}(S_d)$ est généralement petit, et il n'y a pas de sens à utiliser la méthode des pourcentages. Dans ce cas, il suffit de prendre comme limites de l'intervalle d'estimation $[a, b]$ les valeurs extrêmes :

$$a = \min_{s \in S_d} \bar{z}(s), \quad b = \max_{s \in S_d} \bar{z}(s) \tag{6.88}$$

Toutes choses égales par ailleurs, il semble donc *a priori* que le cas le plus délicat soit celui de l'ES. Néanmoins, la qualité du modèle dépend en grande partie de la représentativité de l'échantillon (Cassel *et al.* 1977, p. 111), cette représentativité étant généralement assurée par une répartition homogène des supports dans le domaine, ce qui est le cas avec l'ES. Ainsi, si la taille d'échantillon est suffisante pour pouvoir construire un modèle raisonnablement fiable (*e.g.*, $n = 100$), ce qui est perdu en robustesse du fait de la forme de la p-distribution de l'ES peut être compensé par une bonne représentativité de l'échantillon, et par conséquent, par un bon modèle de surface. Il est cependant difficile de concevoir une étude de sensibilité de portée suffisamment générale pour être utile, à cause des nombreux facteurs qui devraient être pris en considération : non-homogénéité spatiale, anisotropie, taille de l'échantillon, portée de l'autocorrélation, type de modèle de variogramme, géométrie du domaine, etc.

6.3.1.7 Autres approches

Dans le cadre d'une modélisation statistique classique, les valeurs $\{z_i \mid i = 1, \ldots, n\}$ sont vues comme étant n réalisations indépendantes d'une VA unique Z. Les intervalles de confiance calculés dans ce contexte sont trop étroits dans le cas de données positivement autocorrélées, à cause de la redondance d'information qui sous-estime la variabilité (Cressie 1991). En effet, dans la situation classique de l'autocorrélation positive, une valeur apporte moins d'information que dans le cas de l'indépendance spatiale parce qu'elle est partiellement prédictible d'après les valeurs voisines (Cliff & Ord 1981).

En absence de structure d'autocorrélation spatiale, la variance est calculée comme si les données étaient indépendantes (*e.g.*, Murray 1996). L'approche la plus simple pour éviter d'avoir à tenir compte de l'autocorrélation consisterait donc à minimiser la dépendance spatiale dans les données (Cardina *et al.* 1997), en adoptant un pas d'échantillonnage plus grand que la portée (Haining 1988, Midgarden *et al.* 1993), ou en utilisant uniquement les données séparées par une distance supérieure à la portée (Habib *et al.* 1991). Cette seconde solution n'est généralement pas recommandée parce qu'elle entraîne une perte d'information dont l'acquisition est souvent coûteuse en écologie (Legendre 1993).

Une autre solution consisterait à modifier la procédure paramétrique classique, notamment en cherchant à calculer une *taille d'échantillon efficace*. Intuitivement, en faisant référence à la quantité d'information apportée par l'échantillon, la taille efficace n' est définie comme la taille qu'aurait l'échantillon si les données étaient indépendantes : dans le cas de l'autocorrélation positive, on a donc $n' < n$. Par exemple, Barnes (1988) propose une heuristique pour mesurer la quantité d'information présente dans un échantillon spatialement autocorrélé sous la forme d'une taille d'échantillon efficace calculée à partir du variogramme estimé. En adoptant ce type d'approche, Dutilleul *et al.* (1993) considèrent l'autocorrélation spatiale (positive) comme un *paramètre de nuisance*[15] conduisant à sous-estimer la variabilité et par conséquent, à sous-estimer la largeur de l'intervalle de confiance de la moyenne à $100(1 - \alpha)$ %, calculé classiquement selon :

$$\left[\bar{z} - \frac{s}{\sqrt{n}} \cdot t_{n-1,1-\frac{\alpha}{2}}, \bar{z} + \frac{s}{\sqrt{n}} \cdot t_{n-1,1-\frac{\alpha}{2}}\right] \tag{6.89}$$

[15] Un paramètre de nuisance est un paramètre dont on doit tenir compte pour la validité de la procédure statistique mais qui n'est pas de premier intérêt pour le scientifique (Lindsay 1985).

avec s l'écart-type estimé de Z, et $t_{n-1,1-\frac{\alpha}{2}}$ la valeur critique du t de Student pour $n-1$ degrés de liberté et la probabilité $1-\frac{\alpha}{2}$. À l'aide d'une étude de Monte-Carlo, Dutilleul *et al.* (1993) examinent sept méthodes destinées à pallier l'effet de l'autocorrélation spatiale, en faisant référence aux travaux de Box (1954a, 1954b) et de Cliff & Ord (1975, 1981). Dutilleul *et al.* (1993) concluent qu'il est nécessaire à la fois d'utiliser une taille d'échantillon efficace et de modifier l'estimation de la variance.

Dans un cadre non paramétrique, Hall (1988) estime la distance minimale au-delà de laquelle il y a indépendance spatiale, partitionne l'échantillon en sous-ensembles qui sont traités comme s'ils étaient indépendants, puis utilise le bootstrap dans chaque ensemble. Les intervalles de confiance obtenus sont finalement combinés grâce à une inégalité de Bonferroni.

Dans ce qui suit, nous nous contentons d'évaluer la méthode *model-based* et la méthode intermédiaire que nous proposons, l'évaluation d'autres approches étant en dehors de notre propos.

6.3.1.8 Etudes de cas

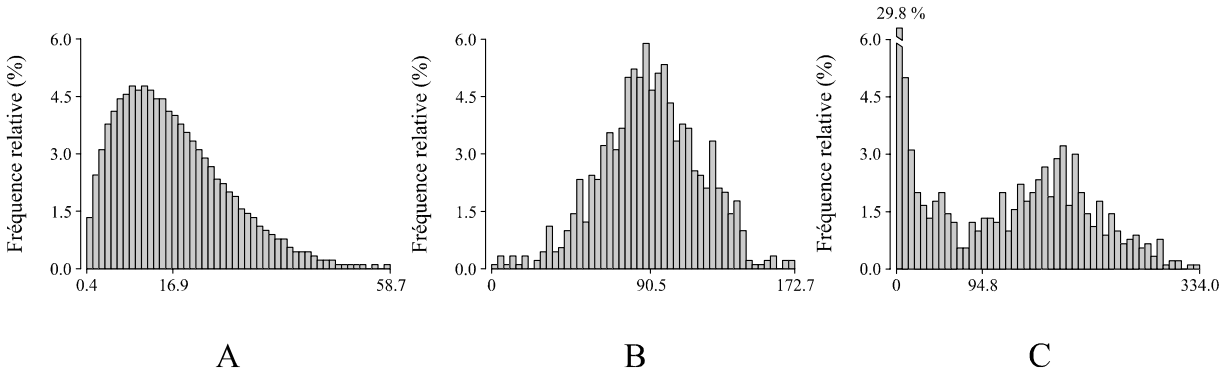
Les études de cas ne constituent pas des démonstrations mais permettent d'illustrer l'application de nouvelles méthodes, en l'occurrence l'estimation par intervalle de la moyenne globale z_D selon l'approche *model-based* et l'approche intermédiaire.

Populations Nous disposons de trois populations, chacune étant constituée de $N = 900$ quadrats, assimilables à des supports quasi-ponctuels, organisés selon une grille 30×30 de pas Δ discrétisant un domaine carré D de $L = 30$ unités de côté. Deux populations sont artificielles et sont obtenues par :

- simulation gaussienne d'une FAST-2 spécifiée par un modèle de variogramme isotrope, la réalisation simulée ayant pour supports les centres des 30×30 quadrats (Section 4.3.3.2),
- éventuelle anamorphose des valeurs simulées afin de s'écarter d'une distribution symétrique (Annexe E),
- permutation de certaines valeurs afin de respecter un modèle de variogramme pour toutes les distances $h \leq L$.

Une première population A est obtenue en simulant un modèle sphérique, en réalisant une anamorphose vers une loi bêta $\mathcal{B}(2, 10)$, suivie d'une multiplication par un facteur 100, puis en permutant certaines valeurs. L'anamorphose permet d'obtenir une distribution asymétrique (Fig. 6.5.A). Les permutations qui forcent le variogramme local à suivre un modèle pour $h \leq L$ introduisent un effet de pépite relatif $c_0/(c_0 + c) \simeq 0.50$, ce qui n'est pas exceptionnel en écologie.

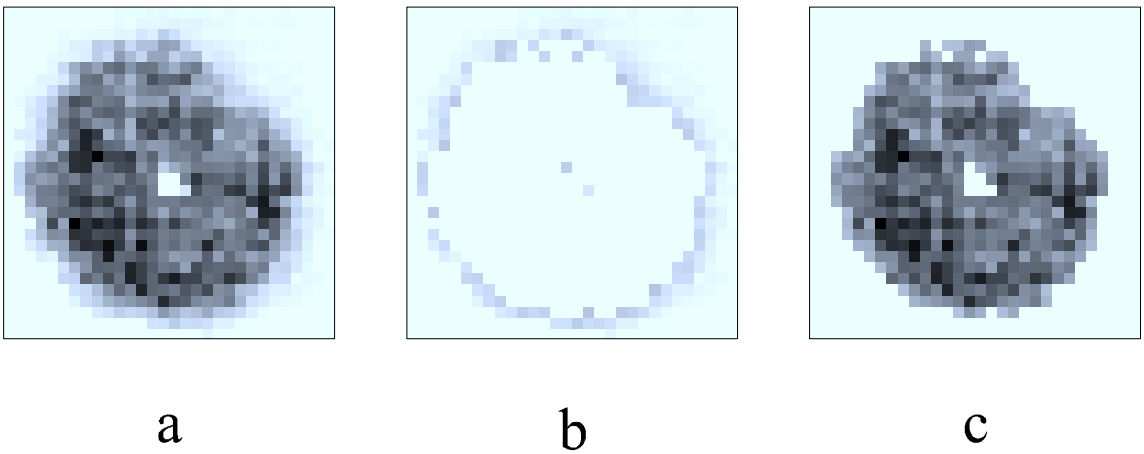
Une deuxième population B est obtenue en simulant un modèle périodique, sans anamorphose, ce qui conduit à une distribution approximativement gaussienne (Fig. 6.5.B). Les permutations jouent le même rôle que précédemment et introduisent un effet de pépite relatif comparable à celui du modèle sphérique ($c_0/(c_0 + c) \simeq 0.45$).



Figure˜6.5: Distributions statistiques des $N = 900$ valeurs des populations A, B et C. En abcisses : valeur minimale, moyenne, valeur maximale.

Les variogrammes dont l'allure correspond aux modèles sphérique et périodique se rencontrent dans des cas aussi divers que le pH du sol, les concentrations en azote, potassium, ou phosphore (Jackson & Caldwell 1993a), la conductivité électromagnétique du sol (Gascuel-Odoux & Boivin 1994), ou la biomasse d'oeufs de lépidoptères (Liebhold *et al.* 1991).

Une troisième population C est constituée par le nombre de glands tombés sur le sol, sous un chêne isolé (*Quercus petraea*), localisé au centre de recherche INRA d'Orléans (année 1986). Le tronc de l'arbre occupe le centre d'un domaine D carré, de côté $L = 15$ m, discrétisé par une grille de 30×30 quadrats carrés de 0.25 m^2. La variable régionalisée (additive) est la densité de glands par quadrat. Le nombre de glands a été compté deux fois par semaine, entre le début du mois de septembre et la fin du mois d'octobre 1986. Aucun gland ne tombe en dehors de D. Au total, 85305 glands ont été comptés, et 678 quadrats parmi les 900 de la grille contiennent au moins un gland, les quadrats vides étant localisés au centre de D (quatre quadrats correspondant au tronc), et dans les quatre coins du carré délimitant le domaine échantillonné D (Fig. 6.6.a).



Figure˜6.6: Partition spatiale associée à la décomposition de la distribution statistique bimodale de la population C. (a) Population totale. (b) Quadrats comptant moins de 80 glands. (c) Quadrats comptant plus de 80 glands.

La présence de nombreuses valeurs nulles et leur localisation dans D impliquent :

- une distribution des valeurs fortement asymétrique (Fig. 6.5.C), qui peut être vue comme résultant du mélange d'une distribution exponentielle négative et d'une distribution gaussienne, cette décomposition étant associé à une très nette partition spatiale (Fig. 6.6),
- une structure spatiale très caractéristique, associée à un variogramme montrant un effet de bord prononcé, mais sans pépite (Fig. 6.7.0C).

Le problème des valeurs nulles n'est pas exceptionnel et survient également dans le traitement des données halieutiques issues de mesures acoustiques (Guiblin 1997). Il y a évidemment deux options exclusives :

- soit supprimer les valeurs nulles, ce qui revient à introduire une incertitude concernant les limites de la population cible,
- soit conserver les valeurs nulles, ce qui introduit une source de non-homogénéité spatiale, perturbe l'analyse variographique, et empêche d'appliquer de façon satisfaisante l'anamorphose gaussienne.

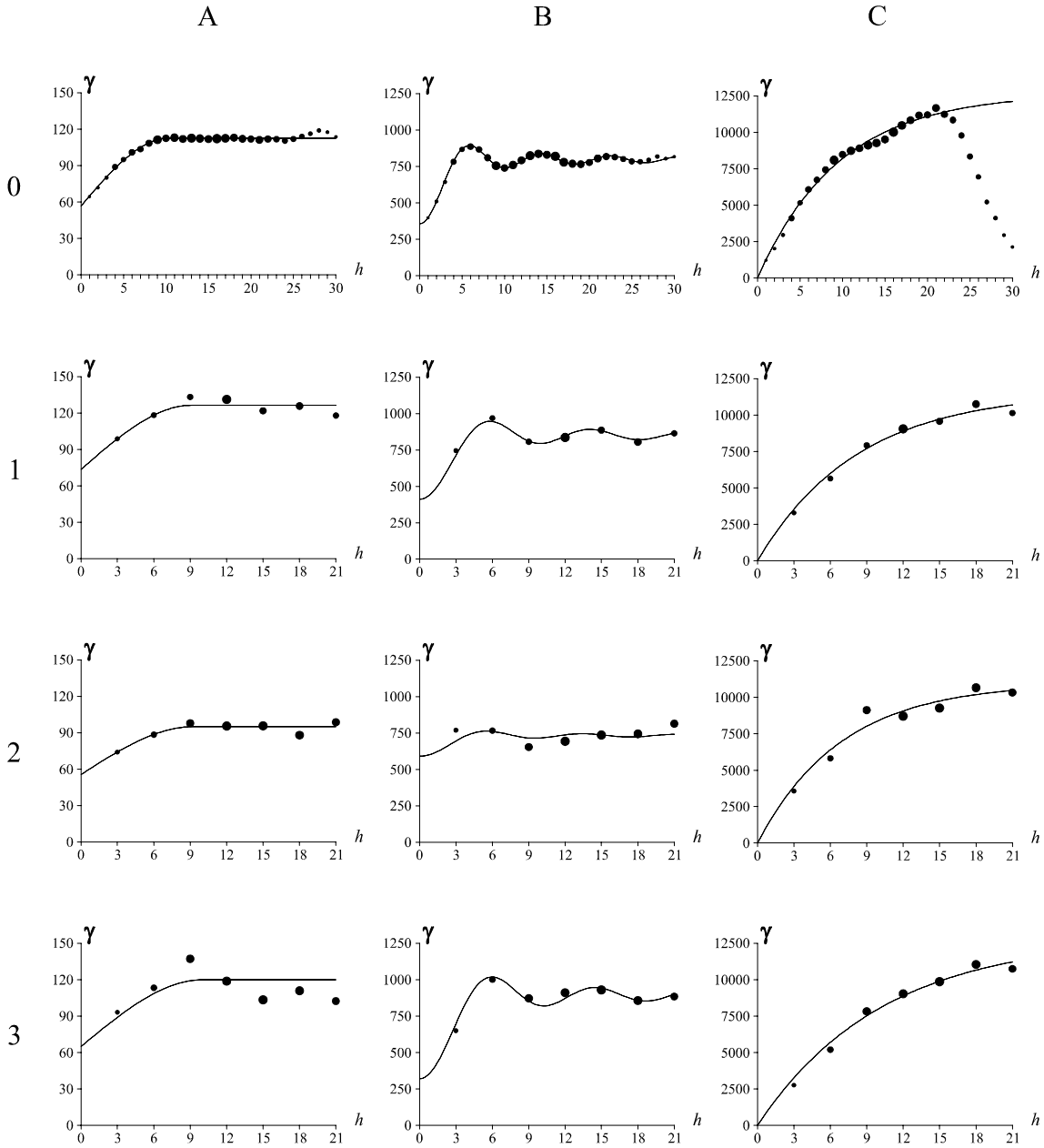
La suppression des valeurs nulles implique une redéfinition du domaine d'étude. Cette redéfinition serait exempte d'incertitude si nous connaissions complètement la population, ce qui n'est généralement pas le cas. En considérant un seul échantillon, l'omission des valeurs nulles nécessiterait de recourir à la géostatistique transitive pour tenir compte de l'incertitude liée à la définition du domaine occupé par la population cible. En outre, comme l'énonce Guiblin (1997) :

- que faire des valeurs nulles intérieures ?
- pourquoi ne pas retirer également les faibles valeurs, les 1 ou toutes les valeurs inférieures à 10, 20, etc. ?

Dans notre cas, il faudrait effectivement supprimer également les valeurs nulles correspondant au tronc du chêne, et les faibles valeurs qui participent à la distribution exponentielle négative que nous avons identifiée dans l'histogramme. Le problème qui se pose est en fait celui du traitement d'une population statistique que l'on peut considérer comme un mélange de deux populations : une population centrale (hormis les valeurs nulles correspondant au tronc) et une population périphérique (Fig. 6.6). De même que Guiblin (1997), nous avons pris le parti de conserver toutes les données, y compris les valeurs nulles. Dans le cadre de l'approche *model-based*, la population C permet ainsi d'évaluer la robustesse de la modélisation par une FAST-2 gaussienne isotrope lorsque tous les prérequis ne sont pas respectés :

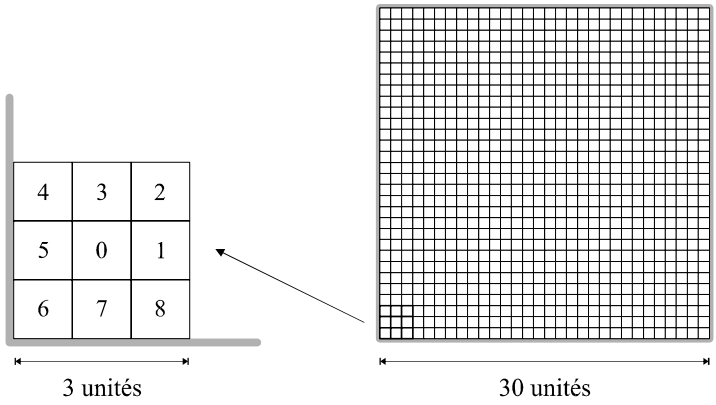
- perturbation de l'homogénéité spatiale par les valeurs nulles périphériques et centrales,
- modèle de variogramme valide pour une partie du domaine seulement,
- impossibilité d'approximer de façon satisfaisante une distribution gaussienne par une transformation non linéaire du type anamorphose.

Cependant, nous considérons que l'hypothèse d'isotropie de la FAST-2 est satisfaite. En effet, bien que l'examen de variogrammes directionnels (non figurés), et de l'image de la population (Fig. 6.6.a) montre qu'il existe une légère anisotropie selon la direction SO-NE, nous considérons néanmoins que cette anisotropie est tout à fait négligeable.



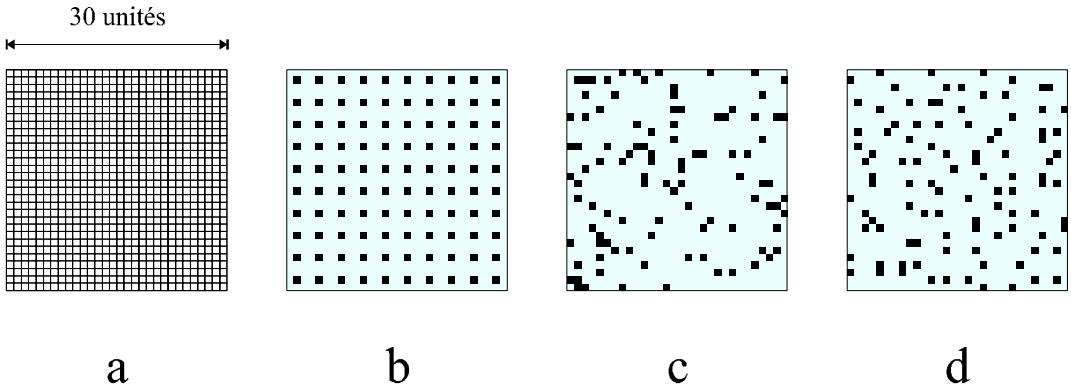
Figure˜6.7: Variogrammes omnidirectionnels et leurs modèles, pour les populations A, B et C. (0) Variogramme local. (1) Variogramme de l'échantillon systématique centré. (2) Variogramme de l'échantillon aléatoire simple. (3) Variogramme de l'échantillon aléatoire stratifié. Les cercles sont proportionnels au nombre de paires utilisées pour calculer chaque valeur. La proportionnalité diffère pour les variogrammes locaux et expérimentaux.

Echantillonnage Nous nous fixons une intensité d'échantillonnage à la fois compatible avec l'effort d'échantillonnage que l'on peut consentir en pratique, et avec la quantité de données nécessaire pour construire un modèle raisonnable, soit par exemple $f = 1/9$. Ainsi, chaque population peut être échantillonnée par un ES selon une grille 10×10, un EAS de taille $n = 100$, et un STR de taille $n = 100$, stratifié selon une grille 10×10. Pour une intensité d'échantillonnage $f = 1/9$, il est possible de prélever uniquement Card $(S_d) = 9$ échantillons systématiques différents organisés par une grille de 10×10 au sein d'une population spatiale de 30×30 quadrats. La seule randomisation de l'échantillonnage possible consiste à tirer au hasard, avec une loi de probabilité uniforme, l'origine de la grille 10×10 au sein d'une strate de 3×3 quadrats (Fig. 6.8).



Figure˜6.8: Localisation en bas à gauche de la grille 30×30 de la population spatiale et numérotation des origines pour les neuf échantillons systématiques 10×10 possibles.

Dans le cas de l'EAS, il est possible de prélever Card $(S_d) = \binom{N}{n}$ échantillons différents, ce qui donne ici Card $(S_d) \simeq 9.384 \times 10^{134}$, tandis que pour le STR, Card $(S_d) = f^{-n}$, soit ici Card $(S_d) \simeq 2.6561 \times 10^{95}$. L'estimation de z_D est effectuée à partir d'un échantillon de chaque type de dispositif au sein de la grille 30×30 : échantillon systématique centré (origine 0), échantillons quelconques générés par EAS et STR (Fig. 6.9). Pour simplifier, nous désignons ces échantillons par les dispositifs qui les ont générés.



Figure˜6.9: Population spatiale et motifs d'échantillonnage spatial. (a) Population spatiale 30×30. (b) Motif systématique centré 10×10. (c) Motif obtenu par échantillonnage aléatoire simple. (d) Motif obtenu par échantillonnage aléatoire stratifié.

L'échantillon EAS conduit, en valeur absolue, à la plus forte erreur d'estimation dans le cas des populations A et C, alors que l'ES donne une estimation assez précise dans tous les cas (Tab. 6.4). Par rapport à l'ES, le STR est meilleur pour la population C, et pire pour les populations A et B (Tab. 6.4).

	z_D	$\bar{z}_{ES}$	$\bar{z}_{EAS}$	$\bar{z}_{STR}$
A	16.91	16.78 (-0.13)	15.25 (-1.66)	17.57 (+0.66)
B	90.55	90.31 (-0.24)	89.81 (-0.74)	89.68 (-0.87)
C	94.78	91.00 (-3.78)	76.77 (-18.01)	91.81 (-2.97)

Tableau 6.4: Moyennes de population (z_D) et des échantillons ES ($\bar{z}_{ES}$), EAS ($\bar{z}_{EAS}$) et STR ($\bar{z}_{STR}$) pour les populations A, B et C. Erreurs d'estimation $\bar{z} - z_D$ entre parenthèses.

En ce qui concerne les variogrammes expérimentaux, l'ajustement des modèles est excellent dans les cas de l'ES, pour les trois populations (Fig. 6.7.1A, 6.7.1B & 6.7.1C), mais mauvais dans le cas de l'EAS pour B (Fig. 6.7.2B), et dans le cas du STR pour A (Fig. 6.7.3A). A cette étape, il est clair que la représentativité de l'ES centré est bonne vis-à-vis des trois populations considérées ici, alors qu'elle est dans l'ensemble plus médiocre pour l'EAS et le STR.

Variances Comme notre objectif n'est pas de discuter des différents estimateurs *design-based* utilisables, notamment dans le cas de l'ES et du STR, et puisque les populations sont connues, nous accédons directement aux p-variances en répliquant les dispositifs. Dans le cas de l'ES, Card (S_d) est très faible et la p-variance de l'estimateur $\bar{Z}$ peut être calculée exactement selon :

$$\mathrm{Var}_p\left[\bar{Z}\right] = \frac{1}{\mathrm{Card}\,(S_d)} \sum_{s \in S_d} \left\{\bar{z}\,(s) - z_D\right\}^2 \tag{6.90}$$

avec $\bar{z}\,(s)$ la moyenne pour l'échantillon s. Dans le cas de l'EAS et du STR, Card (S_d) est très élevé, aussi nous tirons un échantillon aléatoire avec remise dans S_d en répliquant m fois le dispositif, et nous estimons la p-variance par :

$$\widehat{\mathrm{Var}_p}\left[\bar{Z}\right] = \frac{1}{m-1} \sum_{i=1}^{m} \left(\bar{z}_i - z_D\right)^2 \tag{6.91}$$

Afin d'obtenir une estimation très précise, nous effectuons $m = 10^6$ réplications. Les valeurs obtenues sont considérées comme des références, au moins dans le cadre *design-based*. Dans l'approche intermédiaire, la population est interpolée par krigeage modifié. La p-variance estimée, notée σ_I^2, est calculée par (6.90) dans le cas de l'ES et par (6.91) dans le cas de l'EAS et du STR, avec $m = 10^4$. En ce qui concerne l'approche *model-based*, σ_E^2 est calculée à partir :

- du modèle de variogramme ajusté au variogramme local γ_D : σ_E^2 (1),
- du modèle de variogramme ajusté au variogramme expérimental $\widehat{\gamma}_D$ calculé pour 7 classes de pas $h = 3\Delta$ et de tolérance $\varepsilon = 0.5 \times h$, avec Δ le pas de la grille 30×30 de la population correspondante : σ_E^2 (2).

La variance conditionnelle σ_C^2 est calculée à partir de $L = 10^4$ simulations conditionnelles (Fig. 6.10). Les simulations sont produites par la méthode utilisant l'espérance conditionnelle $\boldsymbol{\mu}_{Z_2|z_1}$ et la décomposition de Cholesky de la matrice de covariance conditionnelle $\mathbf{C}_{Z_2|z_1}$ (Section 4.3.5.2). L'espérance et la covariance conditionnelles sont calculées à partir des données anamorphosées et du variogramme de ces données, les réalisations simulées subissant l'anamorphose réciproque (Annexe E).

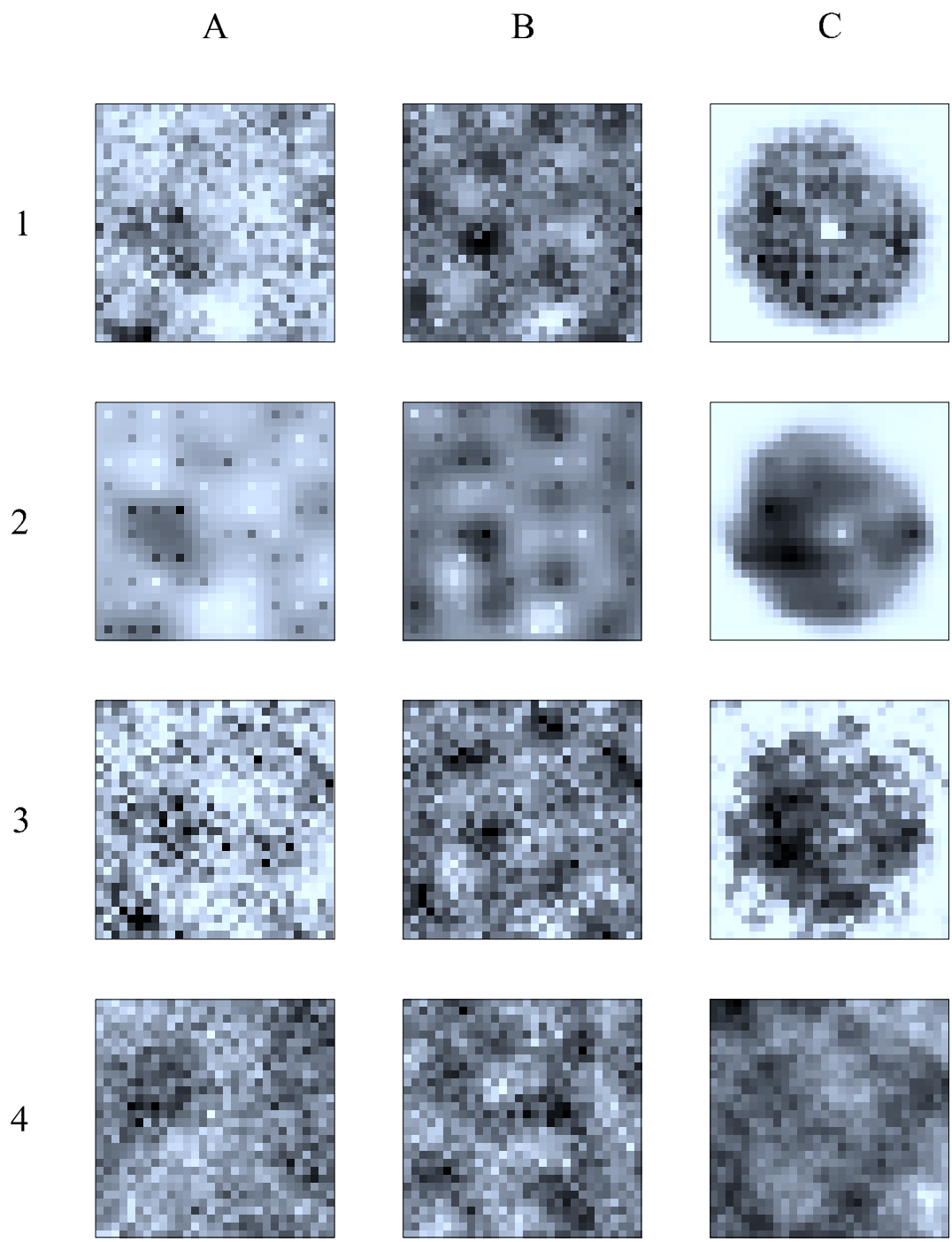
Pour les trois populations, la p-variance montre que les dispositifs ES et STR sont plus précis que l'EAS, mais que l'efficacité relative de l'ES et du STR dépend de la population (Tab. 6.5). Sauf dans le cas de l'EAS, la variance σ_I^2 sous-estime généralement la p-variance (hormis pour l'ES de la population C), ce qui montre la difficulté d'obtenir un modèle déterministe reproduisant fidèlement la variabilité de la population, même avec le krigeage modifié.

		Var_p	σ_I^2	σ_E^2 (1)	σ_E^2 (2)	σ_C^2
	ES	0.657	0.056	0.621	0.798	0.477
A	EAS	0.969	0.990	1.038	0.889	0.451
	STR	0.558	0.128	0.683	0.767	0.488
	ES	2.254	0.135	3.574	4.137	2.166
B	EAS	7.019	7.481	7.008	6.858	1.833
	STR	4.180	1.911	4.825	4.781	2.151
	ES	4.946	7.639	8.881	9.539	7.517
C	EAS	74.162	81.258	98.452	90.113	21.885
	STR	12.630	4.909	17.600	16.827	10.421

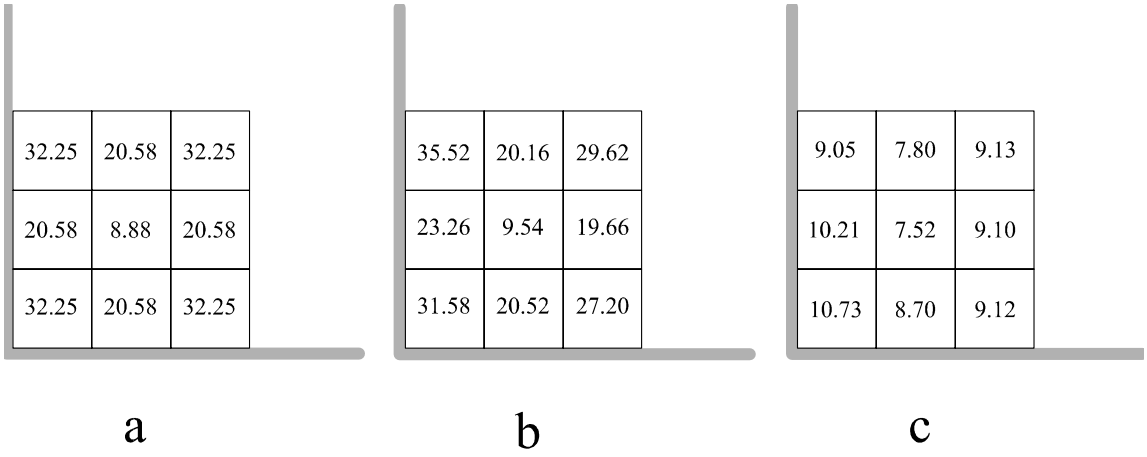
Tableau 6.5: p-variance de référence (Var_p), p-variance estimée par l'approche intermédiaire (σ_I^2), variances d'erreur d'estimation non conditionnées (σ_E^2 (1) et σ_E^2 (2)) et conditionnée (σ_C^2), pour les trois échantillons issus des dispositifs ES, EAS, STR, et les trois populations finies A, B, C (détails dans le texte).

L'impact de l'erreur d'estimation du variogramme dans le calcul de σ_E^2 peut être apprécié en comparant les valeurs de σ_E^2 (1) et de σ_E^2 (2) (Tab. 6.5, Fig. 6.11.a & 6.11.b). Les variations observées ne sont pas considérables et sont du niveau de celles dues à l'intégration de Monte-Carlo (Aubry & Debouzie 1999a). Ceci montre une relative robustesse de σ_E^2 vis-à-vis de l'erreur d'estimation du variogramme local, du moins pour les populations A, B, et C. La variance σ_E^2 apparaît du même ordre de grandeur que la p-variance, bien qu'il ne s'agisse pas d'un estimateur de $\text{Var}_p\left[\bar{Z}\right]$. A ce stade, σ_E^2 semble constituer une variance très intéressante pour les dispositifs ES et STR qui soulèvent des problèmes dans le cadre de l'inférence *design-based*.

Cependant, en examinant les neuf échantillons de l'ES, par exemple dans le cas de la population C, l'échantillon en position centrale (origine 0) apparaît globalement deux fois plus précis que les échantillons centrés en lignes ou en colonnes, *i.e.* en positions 1, 3, 5, ou 7, et trois fois plus précis que les échantillons non centrés, *i.e.* en positions 2, 4, 6, ou 8 (Fig. 6.11.b). Ce résultat découle de la nature non conditionnée de σ_E^2 et son interprétation est détaillée dans Aubry & Debouzie (1999a).



Figure˜6.10: Images se référant aux trois populations A, B et C. (1) Population. (2) Surface moyenne des 10^4 réalisations conditionnelles. (3) Exemple de réalisation conditionnelle. (4) Exemple de réalisation non conditionnelle. Les figures des lignes 2 & 3 ont été produites en conditionnant la simulation par l'échantillon systématique centré. Les surface 2A & 2B montrent des discontinuités dues à l'effet de pépite dans les variogrammes correspondant. La différence entre la simulation conditionnelle *vs.* non conditionnelle peut être appréciée en comparant les lignes 1, 3 & 4. La comparaison entre les lignes 1 & 4 montre que la simulation non conditionnelle ne respecte pas la structure spatiale de la population (*e.g.*, population C). Au contraire, les réalisations conditionnelles de la ligne 3 imitent bien les populations originelles correspondantes (ligne 1). En effet, les réalisations conditionnelles reflètent à la fois la variabilité spatiale de la population (ligne 1), et sa structure spatiale globale décrite par la surface moyenne (ligne 2).



Figure˜6.11: Variances non conditionnelles (σ^2_E) et conditionnelles (σ^2_C) pour les neuf échantillons systématiques 10×10 de la population C. (a) Variances σ^2_E (1) calculées d'après le modèle du variogramme local. (b) Variances σ^2_E (2) calculées d'après les modèles des variogrammes expérimentaux. (c) Variances σ^2_C calculées d'après les modèles des variogrammes expérimentaux et les valeurs des échantillons.

Au moins dans le cas de l'ES, la variance σ^2_E ne constitue pas une mesure de précision satisfaisante parce que la variation observée est sans véritable rapport avec les erreurs d'estimation des neuf échantillons. En effet, les erreurs absolues $\mid \bar{z} - z_D \mid$ varient seulement entre 0.38 et 3.78, le maximum étant atteint pour l'échantillon centré (origine 0). En revanche, la variance conditionnée σ^2_C est du même ordre de grandeur pour les neuf échantillons (Fig. 6.11.c), ce qui est conforme au fait que les neuf estimations sont d'une précision semblable. La variance σ^2_C est plus faible que σ^2_E du fait du conditionnement par les valeurs observées (Tab. 6.5, Fig. 6.11.b & 6.11.c).

Intervalles de confiance et de prédiction Les p-distributions obtenues par réplication des dispositifs ES, EAS, et STR permettent de calculer directement les intervalles de confiance des estimations $\bar{z}$. Dans le cas de l'ES, le recours à la loi normale pour calculer un intervalle de confiance comme en (6.58) n'a pas de sens puisque $\mathrm{Card}\,(S_d) = 9$, aussi nous donnons simplement l'intervalle de variation de $\bar{z}$ défini par (6.88). En revanche, dans le cas de l'EAS et du STR, l'utilisation de la loi normale est justifiée et nous calculons des intervalles de confiance proprement dits comme en (6.58), avec $\alpha = 0.05$.

Bien que définis dans un cadre *design-based*, ces intervalles d'estimation peuvent être vus comme des références dans la mesure où ils sont parfaitement objectifs. En adoptant ce point de vue, il ne s'agit pas pour nous de comparer les approches *design-based* et *model-based* (*e.g.*, Brus & de Gruijter 1997), mais d'examiner dans quelle mesure les intervalles d'estimation obtenus grâce aux modèles (approches intermédiaire et *model-based*) sont objectifs, et par conséquent, dans quelle mesure ces méthodes d'inférence peuvent s'avérer utiles en écologie.

Dans l'approche intermédiaire décrite dans la Section 6.3.1.6, les intervalles d'estimation $[a, b]$ sont obtenus sur la base des neuf échantillons ES par (6.88), et à partir des approximations des p-distributions pour l'EAS et le STR, en utilisant la méthode des pourcentages (6.86).

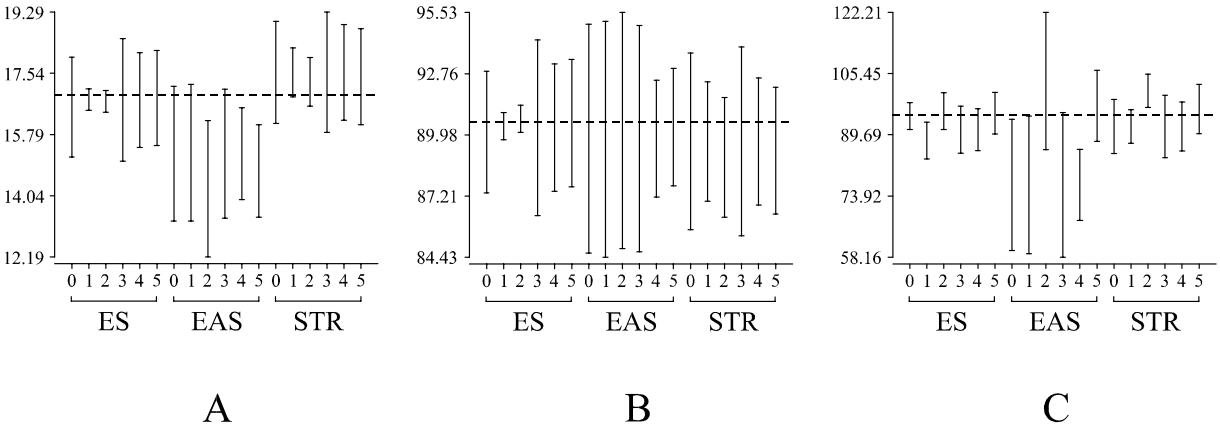
Deux intervalles peuvent être définis :

- en soustrayant l'erreur $\varepsilon = \widetilde{z}_D - \bar{z}$ aux limites de l'intervalle $[a, b]$, soit $[a - \varepsilon, b - \varepsilon]$, avec $\widetilde{z}_D$ la moyenne de la population interpolée par krigeage modifié et $\bar{z}$ la moyenne d'échantillon : IC_1,
- en utilisant directement $[a, b]$: IC_2.

Les intervalles IC_1et IC_2 sont évidemment de même largeur, mais l'intervalle IC_2 exploite davantage le modèle en considérant que $\widetilde{z}_D$ est une estimation plus proche de z_D que ne l'est la moyenne d'échantillon $\bar{z}$. IC_1 s'apparente à un intervalle de confiance de $\bar{z}$ tandis que IC_2 s'apparente davantage à un intervalle de prédiction de z_D, l'objectif étant dans les deux cas de fournir une estimation par intervalle de z_D.

Dans le cas de l'ES centré (origine 0), les résultats montrent que les IC_1 et IC_2 donnent une estimation par intervalle très précise pour les populations A et B (Fig. 6.12). Pour l'ES de la population C, l'IC_1 ne contient pas z_D parce que l'intervalle est centré sur $\bar{z}$ et est étroit. En revanche, l'IC_2 contient z_D, ce qui laisse penser que l'IC_2 peut être préférable à l'IC_1 lorsque la représentativité de l'échantillon est bonne, autrement dit, dans le cas d'un ES.

Ce résultat ne peut pas être étendu aux échantillons EAS. En effet, pour l'EAS-B, IC_1 et IC_2 contiennent tous les deux z_D, alors que l'IC_1 contient z_D pour l'EAS-A et pas pour l'EAS-C, tandis qu'à l'inverse, l'IC_2 contient z_D pour l'EAS-C et pas pour l'EAS-A (Fig. 6.12). Il convient de remarquer que l'IC_1 est très proche de l'intervalle de référence et que, si l'IC_1 ne contient pas z_D pour l'EAS-C, c'est également le cas de l'intervalle de référence (Fig. 6.12.C). Enfin, le cas du STR est intermédiaire entre celui de l'ES et celui de l'EAS (Fig. 6.12).



Figure˜6.12: Intervalles d'estimation de z_D pour les motifs d'échantillonnage ES, EAS, STR, et les populations A, B, C. (0) Intervalle de confiance à 95 % (EAS, STR), ou intervalle de variation (ES), définis dans le cadre *design-based*. (1) & (2) Intervalles IC_1 et IC_2 de l'approche intermédiaire. (3), (4) & (5) Intervalles IP_0, IP_1 et IP_2 de l'approche *model-based*. La ligne en pointillés décrit la moyenne de la population. L'intervalle de variation (ES) n'est pas nécessairement centré sur la moyenne d'échantillon.

Dans le cadre de l'approche *model-based*, trois intervalles de prédiction sont calculés comme en (6.58) :

- en utilisant la variance non conditionnée σ^2_E : IP_0,
- en utilisant la variance conditionnée σ^2_C : IP_1,
- en soustrayant le ξ-biais et en utilisant la variance conditionnée σ^2_C : IP_2.

Les intervalles IP_1 et IP_2 sont évidemment de même largeur, mais l'intervalle IP_2 exploite davantage le modèle en faisant référence au ξ-biais. Enfin, la variance σ^2_C étant plus faible que σ^2_E, les intervalles IP_1 et IP_2 sont plus étroits que l'intervalle IP_0 et conduisent donc à des estimations par intervalles plus précises.

Les IP_0 se révèlent généralement très proches des intervalles de référence dans le cas de l'EAS, mais peuvent être légèrement plus larges. Ainsi, alors que l'intervalle de confiance de l'EAS de la population C ne contient pas z_D, l'IP_0 correspondant est un peu plus large et contient z_D (Fig. 6.12.C). En revanche, dans le cas du STR de la population A, l'IP_0 est manifestement trop large (Fig. 6.12.A). Dans le cas des ES, la largeur de l'IP_0 est excessive (Fig. 6.12). En outre, IP_0 étant directement calculé à partir de σ^2_E, sa largeur dépend de la position de l'origine dans la strate élémentaire 3×3, les IP_0 des ES non centrés (origines 2, 4, 6, et 8) étant plus larges que ceux des ES simplement centrés (origines 1, 3, 5, et 7), eux-mêmes plus larges que celui de l'ES doublement centré (origine 0). Ces inconvénients sont corrigés par les IP_1 qui sont plus étroits et de largeur semblable quelle que soit l'origine.

Dans le cas de l'EAS des populations A et C, l'IP_1 ne contient pas z_D, ce qui est également le cas de l'intervalle de confiance de référence dans le cas de la population C (Fig. 6.12.A & 6.12.C). En faisant intervenir le ξ-biais (IP_2), il est possible de corriger ce problème dans le cas de la population C, mais pas dans le cas de la population A. Il convient de réaliser que l'utilisation du ξ-biais entraîne une exigence accrue de robustesse parce que l'utilisation du modèle est très poussée. Ainsi, il est possible que l'IP_2 à 95 % ne contienne pas z_D alors que l'IP_1 à 95 % contient bien z_D simplement parce que le ξ-biais ne va pas nécessairement dans le sens de l'erreur d'estimation réelle.

6.3.1.9 Recommandations

Les principes des méthodes, les propriétés des dispositifs d'échantillonnage, la structure d'autocorrélation spatiale des populations, et les résultats de l'étude de cas nous permettent de dégager quelques recommandations générales quant à l'estimation par intervalle de la moyenne globale z_D. Nous reprenons les notations utilisées dans l'étude de cas : IC_1, IC_1 pour les intervalles obtenus par l'approche intermédiaire, IP_0, IP_1, et IP_2 pour les intervalles calculés dans l'approche *model-based*.

L'inférence *design-based* ne pose aucune difficulté en ce qui concerne l'EAS, indépendamment de la structure d'autocorrélation spatiale, et représente l'approche la plus robuste parce qu'elle ne fait appel à aucune hypothèse. Cependant, il serait sans doute équivalent d'utiliser l'IC_1 de l'approche intermédiaire ou l'IP_0 de l'approche *model-based*, bien que cela ne présente aucun avantage théorique et demande en pratique davantage de temps de calcul.

L'utilisation d'un modèle de population déterministe (approche intermédiaire), ou stochastique (approche *model-based*), suppose avant toute chose que l'échantillon soit suffisamment représentatif de la population. En toute objectivité, cette représentativité ne peut pas être évaluée *a priori*, mais certains dispositifs d'échantillonnage spatial conduisent souvent à des échantillons représentatifs (ES), et d'autres conduisent généralement à des échantillons peu représentatifs (EAS).

L'approche intermédiaire est certainement l'approche la moins robuste parce qu'il est difficile de reconstituer la variabilité de la population réelle à partir d'un modèle déterministe, fût-il obtenu par krigeage modifié. En conséquence, cette approche devrait être réservée aux variables régionalisées qui présentent une relative régularité spatiale, *i.e.* dont le variogramme est de type gaussien, cubique, ou périodique, et sans pépite. Qui plus est, il est préférable de réserver cette méthode au cas de l'ES qui donne généralement des échantillons assez représentatifs pour que l'on puisse faire confiance au modèle et utiliser l'IC_2.

En ce qui concerne l'approche *model-based*, nous recommandons :

- de recourir à l'IP_0 dans les cas où la représentativité de l'échantillon est *a priori* médiocre, autrement dit, dans le cas d'un motif d'échantillonnage qui ne couvre pas de façon homogène le domaine D (*e.g.*, EAS),
- d'utiliser l'IP_1 dans le cas d'un motif d'échantillonnage régulier (*e.g.*, ES), ou du moins assez homogène (*e.g.*, STR).

Enfin, nous ne pouvons pas recommander l'utilisation du ξ-biais en toute circonstance, bien qu'une valeur très élevée puisse indiquer que l'échantillon est vraisemblablement peu représentatif de la population. Ce point mériterait une étude plus approfondie, notamment pour vérifier que le ξ-biais n'est pas un artefact essentiellement lié à l'étape d'anamorphose gaussienne.

6.3.2 Estimation locale

Soit une variable régionalisée $z(\cdot)$ définie de façon continue sur un domaine D. Les supports de $z(\cdot)$ forment une population spatiale $\mathcal{U} \subset D$, infinie ou finie selon qu'ils sont ponctuels ou surfaciques.

Indépendamment de la finitude de $\mathcal{U}$, l'estimation globale visait à estimer la valeur moyenne de $z(\cdot)$ sur D, autrement dit, un paramètre de $z(\cdot)$ vue comme une population de valeurs. Dans ce contexte, la précision statistique de l'estimation globale pouvait être quantifiée en faisant référence à différents paradigmes inférentiels, notamment les approches *design-based* et *model-based*. En revanche, l'estimation locale consiste à estimer (ou prédire) la valeur de $z(\cdot)$ pour tous les supports de $\mathcal{U}$, à partir d'un ensemble de mesures de supports $s = \{s_i \mid i = 1, \ldots, n\}$. Dans ce contexte, il n'est plus possible de recourir à l'approche *design-based*, parce que les estimateurs qu'elle propose ne concernent que des quantités globales (paramètres de la population), et pas les valeurs $z(u)$ elles-mêmes, avec $u \in \mathcal{U}$. Pour quantifier la précision des estimations locales $z^*(u)$, il est par conséquent nécessaire de recourir à un modèle de superpopulation, généralement une fonction aléatoire.

En pratique, même si $\mathcal{U}$ est infinie, la reconstitution de $z(\cdot)$ s'effectue sur un ensemble discret de supports. Par la suite, une carte isoplèthe peut être calculée afin de décrire la variation spatiale continue de $z(\cdot)$. Si $\mathcal{U}$ est finie, la reconstitution de $z(\cdot)$ donne directement lieu à une carte choroplèthe, en particulier de type raster (Section 2.3). Ainsi, en posant le problème de la précision des estimations locales, il est possible de distinguer le cas des cartes choroplèthes de celui des cartes isoplèthes, mais dans les deux cas, le recours à la simulation conditionnelle semble constituer la meilleure approche.

6.3.2.1 Précision des cartes choroplèthes

Dans le cas des cartes choroplèthes, la précision peut être localisée en chaque support $u \in \mathcal{U}$ centré en x, en considérant une variance d'erreur d'estimation $\mathrm{Var}_\xi [Z^*(x) - Z(x)]$ ou un intervalle de prédiction $[a(x), b(x)]$.

Lorsque la reconstitution de $z(\cdot)$ est effectuée par krigeage, il est courant de quantifier la précision des estimations locales au moyen de la variance de krigeage σ_K^2, autrement dit, la forme minimisée de σ_E^2. En outre, en invoquant l'hypothèse que l'erreur d'estimation suit une distribution gaussienne, un intervalle de prédiction calculé comme en (6.58) peut éventuellement être proposé. On sait que σ_E^2 est une variance non conditionnelle (Journel 1986a, Chauvet 1993), et qu'elle ne constitue pas à proprement parler une mesure de précision mais plutôt une mesure d'incertitude configuration-dépendante qui est souvent sans rapport avec l'amplitude de l'erreur d'estimation réelle (Journel & Rossi 1989). Dans ce contexte, la conception selon laquelle "[...] *a general indication of the quality of a map can be given, determined by the data and by the configuration of the observations as a whole*" (Stein & Corsten 1991) s'avère donc partiellement erronée puisque σ_E^2 ne dépend pas des données autrement qu'à travers l'estimation du variogramme. Bien que σ_E^2 puisse rendre de réels services dans le contexte de l'estimation globale, son utilisation pour calculer un intervalle de prédiction local nécessite de préciser à quelle interprétation du concept de superpopulation il est fait référence.

Dans le cas d'une population fixée, vis-à-vis de laquelle la superpopulation est vue comme un outil mathématique jouant un rôle strictement opératoire, l'utilisation de σ_E^2 pour calculer un intervalle de prédiction local ne peut pas être recommandée, parce que σ_E^2 n'est pas conditionnée par les données. En revanche, si la superpopulation est vue comme la représentation d'un processus réel, structurellement stable, se répétant dans le temps, alors le calcul d'un intervalle de prédiction local à partir de σ_E^2 trouve sa justification dans la répétition **réelle** du processus générateur. Cela dit, la question de la distribution de référence reste en suspens, la loi de Gauss ne constituant qu'un choix par défaut pas nécessairement justifié. Une autre approche consiste à calculer directement un intervalle de prédiction, sans faire référence à une valeur estimée particulière, à σ_E^2 et à la distribution gaussienne.

Soit $Z(x)$ la variable aléatoire modélisant l'incertitude concernant la valeur $z(x)$. Notons pour simplifier $(n) = \{z(x_i) \mid i = 1, \ldots, n\}$ l'ensemble des données disponibles pour estimer $z(x)$. Afin d'estimer $z(x)$ par intervalle, il est nécessaire de considérer la fonction de répartition conditionnelle (Journel 1983a, 1989, 1996, Goovaerts 1997) :

$$F(x\ ;\ z \mid (n)) = \Pr[Z(x) \leq z \mid (n)] \tag{6.92}$$

qui permet de calculer un intervalle de prédiction $[a, b]$ tel que :

$$\Pr\left[Z\left(x\right) \in [a, b] \mid (n)\right] = F\left(x\ ;\ b \mid (n)\right) - F\left(x\ ;\ a \mid (n)\right) \tag{6.93}$$

Cet intervalle est indépendant d'une prédiction particulière $z^*\left(x\right)$ de la valeur inconnue $z\left(x\right)$, et dépend uniquement du modèle probabiliste et des données (n) (Goovaerts 1997). Dans ce contexte, le problème est d'approximer la fonction (6.92) pour chaque support x. Cette approximation peut s'effectuer au moyen d'indicatrices (Journel 1983a, 1989, Isaaks & Srivastava 1989, Goovaerts 1994b, Tercan & Dowd 1995), dans le cadre des méthodes de la géostatistique non linéaire. Il existe plusieurs méthodes d'approximation de (6.92), notamment celles qui font référence au krigeage d'indicatrices (KI), au krigeage de probabilités (KP), ou au krigeage disjonctif (KD) (Goovaerts 1999). En passant en revue ces trois méthodes, Lajaunie (1990) conclut que le KD constitue la procédure la plus appropriée. Par exemple, Lark & Bolam (1997) utilisent le KD en conjonction avec la théorie des ensembles flous afin de traiter simultanément l'incertitude de la prédiction et l'incertitude de l'interprétation dans le cas des variables édaphiques.

Cependant, cette approche présente deux inconvénients majeurs. En premier lieu, l'ensemble des intervalles de prédiction des estimations locales n'est pas équivalent à une prédiction par intervalle de la carte entière. Autrement dit, cette méthode ne rend pas compte de l'incertitude conjointe pour un ensemble de valeurs prédites (Deutsch & Journel 1992, Goovaerts 1997). En second lieu, et sans entrer dans les détails techniques (*cf.* Rivoirard 1991, 1994), le krigeage disjonctif est assez complexe (von Steiger *et al.* 1996, Goovaerts 1999). La lourdeur de la mise en oeuvre du KD nous semble disproportionnée par rapport aux enjeux habituels des études écologiques. Une solution à ces deux problèmes consiste à recourir à la simulation conditionnelle, ce qui nous conduira à ébaucher la notion de *carte stochastique*.

6.3.2.2 Précision des cartes isoplèthes

Lorsque la variation spatiale de $z\left(\cdot\right)$ est représentée par une carte isoplèthe, la précision porte sur les isolignes elles-mêmes, la question qui est posée étant de savoir quel crédit il faut accorder à un tracé particulier d'isolignes. En pratique, le manque d'analyse d'incertitude au sein des programmes de calcul des cartes isoplèthes laisse l'utilisateur sans interprétation alternative (Wingle & Poeter 1993). Le calcul d'une carte isoplèthe s'effectue généralement en deux étapes (Ripley 1981, Davis 1986, Myers 1994b) :

- interpolation de $z\left(\cdot\right)$ aux noeuds d'une grille régulière Ω dans D, au moyen d'un certain algorithme $\mathcal{A}_1$,
- tracé des isolignes au moyen d'un certain algorithme $\mathcal{A}_2$.

L'incertitude associée à une carte isoplèthe résulte donc de la combinaison de l'incertitude associée à l'étape d'interpolation et de celle associée à l'étape de calcul des isolignes (Myers 1994b). Nous considérons par la suite que les cartes isoplèthes sont obtenues au moyen des mêmes algorithmes $\mathcal{A}_1$ et $\mathcal{A}_2$, à partir d'une même grille Ω, afin de ne pas avoir à tenir compte d'une autre source de variabilité que celle induite par l'échantillonnage de $z\left(\cdot\right)$.

Dans ce contexte, compte tenu de l'information fragmentaire apportée par l'échantillon, il est légitime de nuancer l'interprétation d'une carte isoplèthe particulière. Le problème de la confiance statistique que l'on peut attribuer aux isolignes était déjà signalé par Gabriel & Sokal (1969) : depuis, plusieurs approches ont été envisagées.

Par exemple, afin d'apprécier la précision de la carte isoplèthe de l'acidité des pluies dans l'Est des USA, Eynon & Switzer (1982, *op. cit.* Diaconis & Efron 1989) simulent des échantillons par bootstrap. Cependant, on s'accorde généralement sur le fait que les procédure du type bootstrap sont difficiles à implémenter de façon valide dans le cas de données spatialement autocorrélées[16] (*e.g.*, Cressie 1991, p. 100, Guttorp 1994). Il est vraisemblable que le bootstrap pose encore plus de difficultés pour l'estimation locale que pour l'estimation globale, de sorte que les résultats présentés par Eynon & Switzer (1982) sont sujets à caution (Cressie 1991, p. 491).

Par ailleurs, Bregt *et al.* (1991) calculent une carte isoplèthe du déficit hydrique des sols de la région de Mander (Pays-bas) par krigeage ordinaire à partir d'un jeu de données simulé comportant 398 points. Un sous-ensemble aléatoire de 75 points est soustrait aux données afin de servir de test. Tout en reconnaissant que la variance de krigeage ne constitue pas une mesure de précision locale et ne permet pas de calculer des intervalles de confiance fiables pour les isolignes, les auteurs proposent de modifier de façon *ad hoc* la variance de krigeage afin qu'elle soit plus "réaliste", autrement dit, davantage conforme à la variance d'erreur d'estimation estimée à partir des 75 données non utilisées. Avec ce type d'approche, il est difficile de savoir ce que représente la variance qui est finalement utilisée pour calculer les intervalles de confiance des isolignes.

Lindgren & Rychlik (1995) proposent quant à eux d'associer des bandes de confiance à des isolignes au moyen d'une approche paramétrique utilisant un modèle de fonction aléatoire. Néanmoins, ces auteurs insistent sur le fait que leur méthode suppose que la FA est gaussienne et dérivable. Cette double hypothèse constitue une restriction sur les variables régionalisées considérées. En particulier, l'hypothèse de dérivabilité de la FA correspond à une forte régularité spatiale de la VR (Section 4.2.2.2, p. 75), ce qui se traduit par l'utilisation d'un modèle de variogramme dont le comportement est parabolique à l'origine. Ainsi, Lindgren & Rychlik (1995) considèrent la concentration en ozone dans le Nord-Est des USA et utilisent un modèle de covariance gaussien, ce qui correspond effectivement à une FA dérivable en moyenne quadratique. Il est clair que cette approche ne peut pas être utilisée de façon universelle en écologie parce que les variogrammes expérimentaux présentent généralement un comportement linéaire à l'origine et/ou un fort effet de pépite.

Une solution générale consiste à simuler des réalisations d'un modèle de FA aux noeuds de la grille Ω, conditionnellement aux données, puis à appliquer l'algorithme de tracé $\mathcal{A}_2$ pour chaque réalisation, afin d'examiner la variabilité des isolignes (Guttorp 1994). En outre, cette approche permet de probabiliser tous les attributs d'une carte isoplèthe, notamment la position des extrema locaux de $z(\cdot)$ dans D.

[16] A cet égard, la proposition de Solow (1985) consistant à décorréler les données, à les bootstrapper, puis à les recorréler, n'est certainement pas valide.

6.3.2.3 Notion de carte stochastique

Dans Aubry (1996b, p. 56), nous avons considéré qu'un modèle de structure spatiale est avant tout une image (carte en mode raster), ou plus généralement une tessellation susceptible d'être consommée par des opérateurs destinés :

- à en extraire un certain nombre d'attributs,
- à produire de nouveaux objets.

Il convient d'ajouter que les cartes peuvent constituer des données en entrée d'un modèle de processus écologique, tel que l'évapotranspiration (*e.g.*, Phillips & Marks 1996), ou d'un système d'aide à la décision, par exemple en foresterie (*e.g.*, Liu & Herrington 1996).

Dans la plupart des cas, une partie seulement des cartes utilisées est issue d'images obtenues par télédétection, l'autre partie résultant généralement d'une certaine procédure d'interpolation à partir de données fragmentaires. Dans tous les cas, les cartes sont déterministes, même si elles dérivent d'outils qui peuvent être définis dans un cadre probabiliste, comme c'est le cas du krigeage.

Les cartes étant partiellement erronées, il faut se poser la question de l'impact de ces erreurs sur les résultats des modèles de processus ou des systèmes d'aide à la décision (Phillips & Marks 1996, Liu & Herrington 1996). En effet, il est particulièrement dangereux d'intégrer dans une seule carte en sortie l'information provenant de différentes cartes en entrées, chacune caractérisée par sa propre structure d'erreur (Arbia 1993). Ce problème fondamental est connu comme celui de la *propagation de l'incertitude*[17] (Arbia 1993, Haining & Arbia 1996, Caloz & Collet 1997, Mowrer 1997). Les solutions adoptées consistent par exemple à :

- perturber stochastiquement les entrées en ajoutant un terme d'erreur gaussien, cette procédure étant éventuellement suivie d'un filtrage passe-bas (Liu & Herrington 1996),
- utiliser la variance de krigeage comme indice d'incertitude des valeurs interpolées, puis propager cette mesure d'incertitude dans le modèle (Phillips & Marks 1996).

D'une façon générale, la carte d'une VR $z(\cdot)$ obtenue par une procédure d'interpolation telle que le krigeage présente deux inconvénients majeurs dans le cadre de son utilisation comme entrée d'un modèle spatial $\mathcal{M}(\cdot)$:

- elle conduit à un biais lorsque $\mathcal{M}(\cdot)$ est non linéaire puisque généralement on a (Heuvelink & Pebesma 1999) :

$$\mathcal{M}\left(\mathrm{E}_{\xi}\left[Z(x)\right]\right) \neq \mathrm{E}_{\xi}\left[\mathcal{M}\left(Z(x)\right)\right] \tag{6.94}$$

- elle sous-estime la variabilité réelle de la VR de sorte que la variabilité locale des phénomènes naturels est mal modélisée (Mowrer 1997), et les considérations statistiques portant sur les valeurs estimées ne peuvent pas être étendues aux valeurs réelles inconnues (Chauvet 1985).

[17] Ce problème a été retenu comme thème central du symposium *Accuracy 2000* organisé par l'Université d'Amsterdam (www.gis.wau.nl/Accuracy2000).

Pour éviter les biais et apprécier l'incertitude des résultats d'un modèle spatial, il convient donc de ne pas considérer une seule carte mais plutôt un grand nombre de cartes alternatives, équiprobables, dont l'ensemble constitue finalement une *carte stochastique*. Cette approche revient à appliquer le modèle $\mathcal{M}(\cdot)$ à une distribution de probabilités plutôt qu'à une espérance conditionnelle. Autrement dit, l'utilisation d'une carte stochastique conduit *ipso facto* à une simulation de Monte-Carlo[18], et cela quelle que soit la nature de $\mathcal{M}(\cdot)$ (déterministe ou stochastique).

La simulation conditionnelle de fonctions aléatoires constitue actuellement l'approche la plus utilisée pour définir des cartes stochastiques, et en particulier des images stochastiques (Journel 1996). Chaque réalisation est simulée en tenant compte de l'information disponible, *i.e.* de valeurs, de contraintes d'inégalité, de l'autocorrélation spatiale, etc. Cependant, la question de savoir si une carte stochastique doit être obtenue par simulation conditionnelle ou non conditionnelle dépend de la nature du phénomène étudié, et en particulier de sa stabilité spatiale.

Quoi qu'il en soit, l'intérêt d'une carte stochastique est qu'elle peut être soumise à tout un ensemble d'opérateurs, statistiques, géomatiques, morphologiques, dont les résultats se trouvent *ipso facto* probabilisés.

Considérons un modèle écologique $\mathcal{M}(\cdot)$ sous la forme d'une boîte noire consommant un ensemble de cartes $\mathcal{C}$ et produisant un ensemble de résultats $\mathcal{R}$, soit formellement $\mathcal{R} = \mathcal{M}(\mathcal{C})$. Si $\mathcal{M}(\cdot)$ est un modèle déterministe, le recours à des cartes stochastiques en entrée $(\mathcal{C})$ garantit la stochasticité des sorties $(\mathcal{R})$: les résultats seront davantage conformes à la variabilité observée au sein des écosystèmes. Mais si $\mathcal{M}(\cdot)$ est un modèle stochastique et si $\mathcal{C}$ est constitué de cartes déterministes, la variabilité des résultats de $\mathcal{R}$ — qui provient uniquement de la stochasticité de $\mathcal{M}(\cdot)$ — risque de sous-évaluer la variabilité réelle, et les conclusions qui découleront de l'utilisation de $\mathcal{M}(\cdot)$ seront donc partiellement erronées.

Sans tenir compte du coût en temps de calcul, une solution appropriée consiste à substituer des cartes stochastiques aux cartes déterministes. Ainsi, que le modèle écologique ou le système d'aide à la décision soit lui-même déterministe ou stochastique, nous recommandons dans tous les cas d'utiliser des cartes stochastiques. Par exemple, Mowrer (1997) utilise les cartes de trois variables en entrée d'un SIG afin de délimiter des aires potentielles de forêts anciennes subalpines. Cet auteur génère 500 réalisations indépendantes de chaque variable par simulation gaussienne séquentielle afin de définir des régions de confiance à 90 %, 95 %, et 99 %. Ce type d'approche est appelé à se développer, essentiellement au rythme de l'augmentation de la puissance de calcul des ordinateurs.

[18] Heuvelink & Pebesma (1999) considèrent également le problème du changement de support lors de l'application du modèle, sujet que nous n'abordons pas ici.

Chapitre 7

Variogramme

> "$2\gamma(h)$ *has been called the variogram by Matheron* [...] *a structure function by Yaglom* [...] *in probability and Gandin* [...] *in meteorology, and a mean-squared difference by Jowett* [...] *in time series*" (Cressie 1989)

La géostatistique utilise comme modèles de superpopulations des fonctions aléatoires dont la structure d'autocorrélation spatiale est résumée sous la forme de *fonctions structurales*. Ces fonctions structurales diffèrent selon la classe de FA (Section 4.2.2) :

- covariance pour les FAST-2,
- variogramme pour les FAI-0,
- covariance généralisée pour les FAI-k.

La classe des FAI-k inclut celles des FAI-0, cette dernière incluant à son tour la classe des FAST-2. A notre connaissance, l'application de la géostatistique à l'écologie considère généralement les FAST-2 et les FAI-0, et plus rarement les FAI-k ($k > 0$) (*e.g.*, Cesaroni *et al.* 1997). En conséquence, dans ce chapitre nous considérons essentiellement le variogramme qui fait office de fonction structurale dans le cas des FAI-0 strictes, et permet de modéliser la covariance dans le cas des FAST-2. Nous examinerons successivement l'estimation du variogramme, sa modélisation et la question de sa précision.

7.1 Estimation du variogramme

Le problème de l'estimation du variogramme est généralement posé en termes d'estimation du variogramme théorique $\gamma(\cdot)$, *i.e.* du variogramme d'une superpopulation modélisée par une fonction aléatoire $Z(\cdot)$. Dans cette optique, comme les données ne concernent généralement qu'une seule population — une seule réalisation de $Z(\cdot)$ — et que cette population est connue de façon partielle à travers un échantillon, l'estimation de $\gamma(\cdot)$ s'effectue en deux temps :

- approximation discrète $\widehat{\gamma}_D(\cdot)$ du variogramme local $\gamma_D(\cdot)$ à partir de l'échantillon,
- estimation du variogramme théorique $\gamma(\cdot)$ par le variogramme local $\gamma_D(\cdot)$.

La première estimation ne nécessite qu'une hypothèse de relative homogénéité spatiale de la population (*i.e.*, la VR) tandis que la seconde requiert classiquement des hypothèses de stationnarité et d'ergodicité concernant $Z(\cdot)$. Finalement, le variogramme théorique $\gamma(\cdot)$ est estimé par l'approximation $\widehat{\gamma}_D(\cdot)$. Cette démarche est parfaitement objective dans le cadre de l'étude d'une superpopulation, *e.g.* dans le cadre d'une étude de Monte-Carlo reposant sur la simulation numérique de $Z(\cdot)$. En revanche, dans le cadre de l'étude de données réelles, l'utilisation d'un modèle superpopulationnel ne constitue le plus souvent qu'un choix opératoire. Dans ce cas, il n'existe pas nécessairement de variogramme théorique immanent qui serait sous-jacent au phénomène étudié (*e.g.*, la topographie). Il convient donc de bien distinguer les opérations qui concernent l'approximation de $\gamma_D(\cdot)$ à partir d'un échantillon, sanctionnées par une erreur d'estimation parfaitement objective, des opérations concernant l'estimation de $\gamma(\cdot)$, liées à la notion de fluctuation, qui revêtent le plus souvent un caractère subjectif.

Dans ce qui suit, nous adoptons un point de vue centré autour de l'approximation du variogramme local et de sa modélisation. Dans cette optique, nous identifions le variogramme local $\gamma_D(\cdot)$ au variogramme théorique $\gamma(\cdot)$, *i.e.* à son espérance $\mathrm{E}_\xi[\gamma_D(\cdot)]$, de sorte que l'approximation discrète $\widehat{\gamma}_D(\cdot)$ estime directement $\gamma(\cdot)$. Dans ce contexte, l'hypothèse d'ergodicité est sans objet, et l'hypothèse de stationnarité ne concerne que le choix du type de FA, en fonction de l'homogénéité spatiale de la VR traitée (Journel 1985).

7.1.1 Estimateurs

Considérons une variable régionalisée isotrope ; à partir d'un échantillon s de supports ponctuels x, l'estimateur classique obtenu par la méthode des moments peut s'écrire (Matheron 1695) :

$$\widehat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \{z(x+h) - z(x)\}^2 \tag{7.1}$$

Cet estimateur est sans biais mais se révèle très sensible à l'asymétrie de la distribution statistique de la VR. En fait, dans le cadre d'une superpopulation — lorsque la VR considérée est effectivement une réalisation d'une FAI-0 stricte ou d'une FAST-2 — l'utilisation du variogramme en tant que fonction structurale suppose implicitement la normalité (Chauvet 1985).

L'estimateur (7.1) ne constitue qu'une approximation discrète du variogramme local $\gamma_D(\cdot)$ défini comme l'intégrale d'espace[1] (Matheron 1965, Journel & Huijbregts 1978) :

$$\gamma_D(h) = \frac{1}{2[D \cap \tau_h D]} \int_{D \cap \tau_h D} \{z(x+h) - z(x)\}^2 \, dx \tag{7.2}$$

avec τ_h l'opérateur de translation par les vecteurs de module h appliqué au domaine D.

En identifiant le variogramme local $\gamma_D(\cdot)$ à son espérance $\mathrm{E}_\xi[\gamma_D(\cdot)]$, autrement dit, en se plaçant dans un cadre *non ergodique* (Journel 1985), le problème consiste à approximer le mieux possible l'intégrale d'espace (7.2). C'est dans ce contexte que Isaaks

[1] La question de savoir si cette intégrale doit être vue comme une intégrale de Riemann ou de Lebesgue est discutée par Isaaks & Srivastava (1988).

& Srivastava (1988) proposent d'estimer la covariance spatiale, plutôt que le variogramme (Section 3.3.3). Une étude approfondie de toutes les pratiques d'estimation de la structure d'autocorrélation spatiale étant hors de notre propos, nous considérons dans ce qui suit uniquement les estimateurs du variogramme.

L'estimateur classique (7.1) se révèle sensible à la présence de valeurs extrêmes (*outliers*). La sensibilité aux valeurs extrêmes concerne généralement toutes les mesures d'autocorrélation spatiale, mais elle s'avère particulièrement grande dans le cas du variogramme. Pour certaines variables, l'addition ou le retrait d'une valeur particulière parmi plusieurs centaines peut entraîner une modification de 10% ou 20 % (Chauvet 1985). Afin de pallier l'absence de robustesse de l'estimateur classique (7.1), des estimateurs robustes ont été proposés. Pour obtenir un estimateur *robuste* il faut, selon Guiblin (1997) :

- trouver une transformation de la VR ou de l'estimateur classique du variogramme,
- introduire un facteur correctif pour annuler le biais dû à la transformation,
- admettre une loi de distribution de la VR qui autorise ces manipulations.

Dans le cadre classique de l'estimation du variogramme théorique $\gamma(\cdot)$, *i.e.* en considérant un ensemble de variables aléatoires $Z(x)$ corrélées, Cressie & Hawkins (1980) ont étudié 10 estimateurs concurrents, incluant l'estimateur classique (7.1). Parmi ces estimateurs, Cressie (1985, 1991) retient finalement les estimateurs robustes suivants :

$$2\widehat{\gamma}(h) = \left[\frac{1}{N(h)}\sum_{i=1}^{N(h)}|z(x+h)-z(x)|^{1/2}\right]^4 \Big/ \left[0.457+\frac{0.494}{N(h)}\right] \tag{7.3}$$

$$2\widehat{\gamma}(h) = \left[\operatorname{med}\left\{|z(x+h)-z(x)|^{1/2}\right\}\right]^4 /0.457 \tag{7.4}$$

avec med un opérateur calculant la médiane d'une séquence de valeurs. Ces estimateurs sont obtenus en considérant que la distribution statistique est pratiquement normale dans sa région centrale et peut être assimilée à une loi normale polluée (Cressie & Hawkins 1980). Les dénominateurs corrigent asymptotiquement le biais introduit par la transformation en exploitant le fait que la variable aléatoire $Y(x) = \{Z(x+h) - Z(x)\}^2$ suit une loi proportionnelle au χ^2_1 lorsque $Z(x)$ suit une distribution gaussienne (Cressie & Hawkins 1980, Cressie 1991). En toute rigueur, les estimateurs (7.3) et (7.4) ne sont valides que si les $Y(x)$ sont indépendants, ce qui n'est pas le cas. Néanmoins, Cressie & Hawkins (1980) estiment que la corrélation des $Y(x)$ est pratiquement négligeable, sauf pour une petite proportion de couples.

Goulard (1988, p. 45) juge cette approche très limitative, et Guilbin (1997) considère que les estimateurs (7.3) et (7.4) ne peuvent pas être utilisés à la place de l'estimateur classique lorsque les données présentent une distribution fortement dissymétrique comme c'est souvent le cas en biométrie. Guilbin (1997, pp. 13-14) propose notamment une formule calculant le variogramme à partir de celui des données log-translatées. En fait, cette formule est utilisée de façon heuristique bien qu'elle soit établie dans le cadre de modèles bien précis, ce qui peut conduire à une estimation biaisée (Guiblin 1997).

D'autres approches destinées à construire un estimateur robuste du variogramme sont énumérées dans Cressie (1991), notamment l'estimation robuste de l'échelle des incréments $z(x+h) - z(x)$. Dans ce cadre, il est possible d'appliquer la théorie des M-estimateurs d'échelle afin d'obtenir un estimateur robuste du variogramme (Genton 1998a). Des simulations réalisées dans $\mathbb{R}$ semblent montrer que l'estimateur obtenu est beaucoup plus robuste que l'estimateur classique (7.1) ou que son principal concurrent (7.3) (Genton 1998a).

Dans ce qui suit, nous ne faisons plus référence aux estimateurs robustes dont il vient d'être question, pour au moins quatre raisons :

- Le problème de l'estimation robuste du variogramme est fondamentalement différent selon que le variogramme à estimer est vu comme un paramètre superpopulationnel (variogramme théorique) ou comme une intégrale d'espace (variogramme local). Or, à notre connaissance, aucune étude concernant l'estimation robuste du variogramme local n'a été publiée, alors que c'est précisément l'estimation d'une grandeur objective, à la différence du variogramme théorique qui joue souvent un rôle strictement opératoire.
- Matheron (1978) est assez critique envers les estimateurs robustes que proposent les statisticiens. En effet, ces estimateurs sont souvent plus robustes vis-à-vis des données mais moins robustes vis-à-vis du modèle que des estimateurs plus simples. Par exemple, l'estimateur robuste (7.3) nécessite que les couples $Z(x), Z(x+h)$ suivent une loi gaussienne bivariée.
- L'utilisation conjointe d'un estimateur robuste pondérant implicitement les données et de la validation croisée (*e.g.*, Maravelias *et al.* 1996) est critiquable en ce que les *outliers* sont toujours présents dans les données. Dans ces conditions, comment juger objectivement de la validité d'un modèle ajusté à une estimation robuste dans des conditions où les poids appliqués aux données sont inconnus ?
- D'une façon générale, nous considérons qu'il est préférable d'identifier explicitement les *outliers* (éventuellement à l'aide d'algorithmes) et de laisser la décision de les conserver ou de les supprimer à l'écologiste lui-même. En effet, en écologie il faut s'attendre à la présence d'*outliers* qui ne sont pas des valeurs "aberrantes" mais font sens (Rossi *et al.* 1992). Ainsi, les *outliers* sont parfois les données les plus intéressantes pour un écologiste (Gaines & Denny 1993, Halvorson *et al.* 1995) ou simplement typiques du phénomène étudié (*e.g.*, Barange & Hampton 1997).

A notre sens, le traitement du problème des *outliers* requiert davantage une analyse exploratoire des données et une analyse de sensibilité des résultats plutôt que la recherche d'estimateurs robustes qui seront utilisés par la suite comme des boîtes noires. L'analyse exploratoire peut être effectuée en utilisant la *nuée variographique* ou des *h-scattergrammes*.

7.1.2 Nuée variographique

La *nuée variographique* est un nuage de points représentant $d(h) = \frac{1}{2}\{z(x+h) - z(x)\}^2$ en fonction de h, sans discrétisation, autrement dit, sans avoir à définir des classes de distances. L'usage de la nuée variographique a été introduit par Chauvet (1982) afin de réaliser une analyse structurale fine, critique, dans laquelle l'objet étudié n'est pas le variogramme expérimental $\widehat{\gamma}(\cdot)$ lui-même mais l'ensemble des valeurs utilisées dans le calcul de $\widehat{\gamma}(\cdot)$.

Le calcul du variogramme expérimental $\widehat{\gamma}(\cdot)$ à partir de la nuée variographique consiste à diviser le nuage en classes de distances contiguës, puis à calculer une valeur moyenne pour chaque classe. Le choix des classes de distances, ainsi que la présence d'*outliers* affectent considérablement l'allure de $\widehat{\gamma}(\cdot)$. En manipulant la nuée variographique, il devient possible d'évaluer l'impact de certaines valeurs $d(h)$ afin de mettre en lumière certaines instabilités dans le calcul de $\widehat{\gamma}(\cdot)$. Afin de disposer d'un outil d'analyse exploratoire interactif, il est possible de concevoir un système de *graphiques dynamiques* reliant la représentation cartographique des données, l'histogramme des valeurs et la nuée variographique (Haslett *et al.* 1991, Bradley & Haslett 1992). Cependant, pour n valeurs représentées cartographiquement, il faut représenter $\binom{n}{2} = n(n-1)/2$ points sur la nuée variographique, ce qui devient rapidement prohibitif. Comme une nuée variographique présente une forte asymétrie positive et que seules les valeurs extrêmes sont finalement intéressantes à examiner, une solution consiste à ne représenter que la partie supérieure de la nuée (Bradley & Haslett 1992).

L'analyse exploratoire fondée sur la manipulation de tels graphiques dynamiques représente certainement l'approche la plus judicieuse pour connaître de façon intime un jeu de données. Néanmoins, si la nuée variographique constitue un outil d'analyse exploratoire extrêmement utile, elle ne fournit pas une information structurale synthétique, et reste difficile à modéliser sans hypothèses fortes (Chauvet 1993).

7.1.3 h-scattergrammes

Déjà moins complets et moins objectifs que la nuée variographique, les *h-scattergrammes* (Rossi *et al.* 1992, Liebhold *et al.* 1993) constituent toutefois un moyen simple d'exprimer la similarité (ou la dissimilarité) entre paires de valeurs $\{z(x), z(x+h)\}$. Pour une distance h donnée, un h-scattergramme est obtenu en représentant $z(x)$ en fonction de $z(x+h)$. Soit un ensemble de données $\{z_i \mid i = 1, \ldots, n\}$, un h-scattergramme est construit en représentant z_j en fonction de z_i pour une classe de distances regroupant un ensemble de couples (z_i, z_j), avec $i \neq j$. L'interprétation d'un h-scattergramme est la suivante (Rossi *et al.* 1992) :

- Les points alignés sur la bissectrice principale correspondent à des valeurs identiques. En conséquence, un point éloigné de cette droite est une valeur extrême — en référence uniquement à la classe de distance considérée — qui doit être examinée et éventuellement supprimée ou corrigée s'il s'agit d'une valeur aberrante (erreur de mesure, erreur de transcription, etc.).
- Plus la distance entre les points augmente, plus le nuage est diffus autour de la bissectrice principale, ce qui traduit l'affaiblissement de l'autocorrélation spatiale.

- Une asymétrie du nuage de part-et-d'autre de la bissectrice principale doit faire suspecter des différences entre moyennes et/ou variances locales.

Le moment d'inertie I_Δ d'un nuage de m points — chacun de poids m^{-1} — par rapport à la bissectrice principale Δ du nuage s'écrit :

$$I_\Delta = \frac{1}{m}\sum_{i=1}^{m} d_i^2 = \frac{1}{2m}\sum_{i=1}^{m} (x_i - y_i)^2 \tag{7.5}$$

avec d_i la distance entre un point et la droite Δ. Le moment d'inertie I_Δ mesure l'aplatissement du nuage et caractérise le manque d'indépendance entre les abscisses et les ordonnées. Dans le cas d'un h-scattergramme, le moment d'inertie n'est pas autre chose que $\widehat{\gamma}(h)$ (Journel 1989, Isaaks & Srivastava 1989). Le variogramme $\widehat{\gamma}(\cdot)$ constitue par conséquent une représentation synthétique d'une famille de h-scattergrammes pour une discrétisation donnée, chaque valeur $\widehat{\gamma}(h)$ étant un résumé d'un h-scattergramme entier. Il convient d'apprécier à la fois la perte d'information qui résulte du calcul de $\widehat{\gamma}(\cdot)$ et sa sensibilité à la présence d'*outliers* (points éloignés de Δ).

Produire une famille de h-scattergrammes nécessite de choisir une discrétisation en classes de distances. Ce choix — assez arbitraire lorsqu'il n'est pas imposé par une répartition régulière des supports — est crucial parce qu'il conditionne en grande partie l'allure de $\widehat{\gamma}(\cdot)$.

7.1.4 Découpage en classes de distances

L'estimation du variogramme local $\gamma_D(\cdot)$ à partir d'un ensemble fini de supports s ne peut s'effectuer que pour les distances induites par s. Pour une cardinalité de s courante (*i.e.*, inférieure à 200 supports), et *a fortiori* si le motif d'échantillonnage est irrégulier, il est peu probable de disposer de beaucoup de valeurs pour chaque distance h induite par s. Ainsi, l'utilisation de classes de distances est indispensable pour estimer chaque valeur $\gamma_D(h)$ à partir d'un grand nombre de couples $\{z(x), z(x+h)\}$.

Il n'existe pas d'argument décisif pour choisir les classes de distances — qui peuvent être d'amplitudes différentes — pas plus qu'il n'existe de critère indiscutable pour décider du nombre de classes. Néanmoins, il est possible de définir une procédure qui s'avère généralement très satisfaisante en pratique.

7.1.4.1 Définition des classes de distances

La procédure classique consiste à fixer *a priori* des classes de distances de pas constant Δ de la forme $h \pm \varepsilon(h)$, où $\varepsilon(h)$ est une tolérance et h un multiple de Δ. Le plus simple est de considérer que la tolérance est constante, soit $\varepsilon(h) = \varepsilon$ (Journel & Huijbregts 1978). Dans ce cas, les classes sont définies par le couple (Δ, ε).

Le regroupement en classes de distances entraîne un lissage du variogramme expérimental et ce n'est plus exactement $\gamma_D(\cdot)$ qui est estimé mais plutôt une combinaison linéaire (David 1977, Journel & Huijbregts 1978). Afin de pouvoir négliger l'effet introduit par le regroupement en classes de distances, il faut que la tolérance ε soit petite vis-à-vis de la portée du variogramme (Journel & Huijbregts 1978). En outre, il est souhaitable

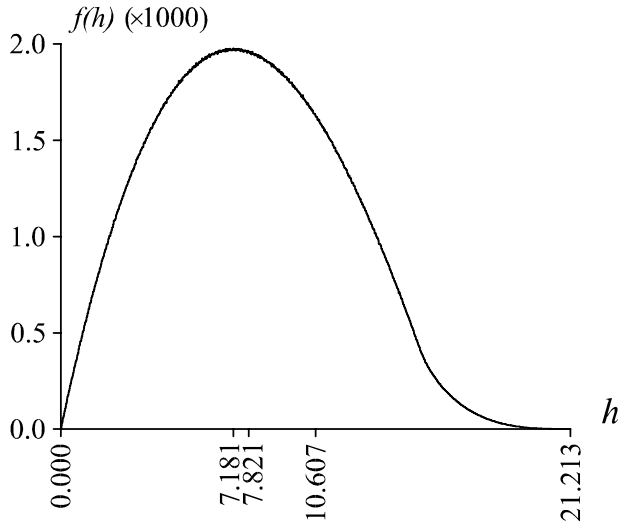
d'estimer $\gamma_D(\cdot)$ en utilisant toute l'information disponible (*i.e.*, tous les supports de s), ce qui nécessite de définir les classes de distances avec une tolérance maximale $\varepsilon = 0.5 \times \Delta$: cette tolérance est d'utilisation courante (*e.g.*, Russo & Jury 1987, Isaaks & Srivastava 1989, p. 148, Carrat & Valleron 1992, Aubry & Debouzie 1999a).

D'une façon générale, le lissage facilite l'ajustement d'un modèle $\widetilde{\gamma}(\cdot)$ au variogramme expérimental $\widehat{\gamma}(\cdot)$ en évitant les allures erratiques. Ainsi, lorsque l'estimation du variogramme est effectuée à des fins de calculs géostatistiques, il s'avère utile d'utiliser des classes de distances même lorsque le motif d'échantillonnage est régulier.

7.1.4.2 Nombre de classes de distances

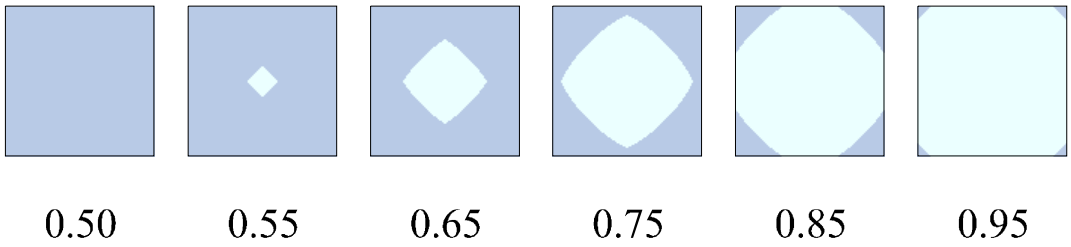
Soit D un domaine et $L_{\max}$ la plus grande distance dans D. Il est classiquement recommandé d'estimer le variogramme en se limitant à des distances $h \leq L_{\max}/2$ (Journel & Huijbregts 1978). L'argument classique invoque le fait que la fluctuation du variogramme local autour du variogramme théorique devient tellement grande au-delà de $L_{\max}/2$ que n'importe quelle allure peut être choisie pour $\gamma(\cdot)$ (Journel & Huijbregts 1978, pp. 193-194). Comme nous considérons que la fluctuation est hors de propos en pratique, nous préférons invoquer un argument géométrique.

Considérons un domaine carré d'arête L pour lequel $L_{\max}$ se calcule simplement comme $L_{\max} = \sqrt{2}L$, par exemple, $L = 15$ ($L_{\max} \simeq 21.213$). La densité de probabilité des distances dans D croît jusqu'à son mode situé approximativement vers 7.181, puis décroît (Fig. 7.1).



Figure˜7.1: Fonction densité de probabilité des distances entre points aléatoires situés dans un carré de 15 unités de côté. En abscisses : distance nulle, distance correspondant au mode, distance moyenne, distance $L_{\max}/2$ et distance $L_{\max}$.

Au-delà de $L_{\max}/2 \simeq 10.607$ on a $D \cap \tau_h D$ Ã D, autrement dit, l'intersection du domaine D avec ses translatés par les vecteurs de module h ne recouvre plus entièrement D. Ainsi, considérer les grandes distances dans D (*e.g.*, $h \geq 0.75 \times L_{\max}$) revient à étudier exclusivement les coins de D, ce qui n'est généralement pas l'objectif recherché (Fig. 7.2).



Figure˜7.2: Intersection (figurée en gris) d'un domaine carré D avec ses translatés par les vecteurs de module $h = \alpha \times L_{\max}$ avec $\alpha \in \{0.50, 0.55, 0.65, 0.75, 0.85, 0.95\}$. Au delà de $\alpha = 0.50$, l'intersection ne recouvre plus entièrement D.

En vertu de l'argument géométrique qui est parfaitement objectif, il convient donc de se restreindre à des distances $h \leq L_{\max}/2$ (Journel 1985, Rossi *et al.* 1992). Il faut alors choisir un nombre de classes à la fois suffisamment élevé pour pouvoir disposer de plusieurs valeurs $\widehat{\gamma}(h)$, et suffisamment faible pour pouvoir disposer de nombreuses valeurs dans chaque classe. Par exemple, pour des échantillons comportant 100 supports, l'utilisation de 7 ou 8 classes de distances constitue un bon compromis.

7.1.5 Optimisation des classes de distances

La littérature exhibe de nombreux exemples de variogrammes $\widehat{\gamma}(\cdot)$ particulièrement erratiques (*e.g.*, Freire *et al.* 1992, Fig. 4, 1993, Fig. 2 & 4, Rahman *et al.* 1996, Fig. 2, Liu & Rossini 1996, Fig. 3 & 4, Vischetti *et al.* 1997, Fig. 2 & 3), ce qui peut être dû, notamment :

- à un trop petit nombre de supports (Webster & Oliver 1992a),
- à la présence d'*outliers* (Liebhold *et al.* 1993),
- à un mauvais découpage en classes de distances (Armstrong 1984, Russo 1984),
- à une structure d'anisotropie sous-jacente (Russo 1984).

En partant du principe que l'échantillon est de taille suffisante et que l'influence des *outliers* a été appréciée, par exemple en explorant la nuée variographique, l'utilisation de classes de distances de pas Δ et de tolérance $\varepsilon = 0.5 \times \Delta$ permet de limiter considérablement les allures erratiques. *A posteriori, c*ette pratique se révèle bien adaptée, surtout lorsque le découpage en classes respecte la distribution des distances induites par s. Aussi, nous proposons d'optimiser la définition des classes afin de respecter au mieux la distribution des distances. La présentation de cette optimisation nécessite de considérer successivement :

- le choix d'un critère W mesurant la qualité des classes de distances,
- le choix d'un algorithme $\mathcal{A}$ destiné à minimiser W,
- une étude de cas illustrant la validité des deux choix précédents.

7.1.5.1 Critère à minimiser

Soit $\mathcal{G}$ l'ensemble ordonné des distances $h \in]0, h_{\max}]$ induites par s, avec $h_{\max} \leq L_{\max}$. Soit P une partition de $\mathcal{G}$ en k classes disjointes $\{C_i \mid i = 1, \ldots, k\}$. Nous proposons d'évaluer la qualité de la partition P au moyen d'un critère W de la forme :

$$W(P) = \sum_{i=1}^{k} I(C_i) \tag{7.6}$$

avec

$$I(C_i) = \sum_{h \in C_i} \omega(h) \cdot \{G(C_i) - h\}^2 \tag{7.7}$$

$$G(C_i) = \frac{1}{\omega(C_i)} \sum_{h \in C_i} \omega(h) \cdot h \tag{7.8}$$

$$\omega(C_i) = \sum_{h \in C_i} \omega(h) \tag{7.9}$$

La classe C_i est munie d'un poids total $\omega(C_i)$ et d'un barycentre $G(C_i)$. Pour une distance h, il est naturel de choisir comme poids la fréquence absolue, soit $\omega(h) = N(h)$. La valeur du variogramme $\widehat{\gamma}(C_i)$ correspondant à la distance moyenne $G(C_i)$ se calcule selon :

$$\widehat{\gamma}(C_i) = \frac{1}{N(C_i)} \sum_{h \in C_i} N(h) \cdot \widehat{\gamma}(h) \tag{7.10}$$

avec

$$N(C_i) = \sum_{h \in C_i} N(h) \tag{7.11}$$

L'expression (7.7) a la forme d'une inertie calculée sur les distances de la classe C_i de sorte que $W(P)$ représente l'inertie intra-classe de la partition P. Minimiser l'inertie intra-classe $W(P)$ revient donc à calculer un variogramme expérimental pour k classes de distances les plus homogènes possible.

7.1.5.2 Algorithme d'optimisation

Soit $P_{m,k}$ l'ensemble des partitions P de k classes qu'il est possible de construire pour $\mathrm{Card}(\mathcal{G}) = m$. Le dénombrement combinatoire de $P_{m,k}$ est donné par le *nombre de Stirling de seconde espèce* (Abramowitz & Stegun 1972, Diday *et al.* 1982) :

$$\mathrm{Card}(P_{m,k}) = \frac{1}{k!} \sum_{i=0}^{k} (-1)^{k-i} \binom{k}{i} i^m \tag{7.12}$$

avec

$$\binom{k}{i} = \frac{k!}{i!\,(k-i)!} \tag{7.13}$$

Compte tenu de l'explosion combinatoire révélée par l'expression (7.12) — à titre indicatif, $\mathrm{Card}(P_{20,5}) \simeq 7.492 \times 10^{11}$ — il s'avère impossible d'examiner toutes les partitions afin de déterminer la partition optimale P^* telle que :

$$W(P^*) = \min_{P \in P_{m,k}} W(P) \tag{7.14}$$

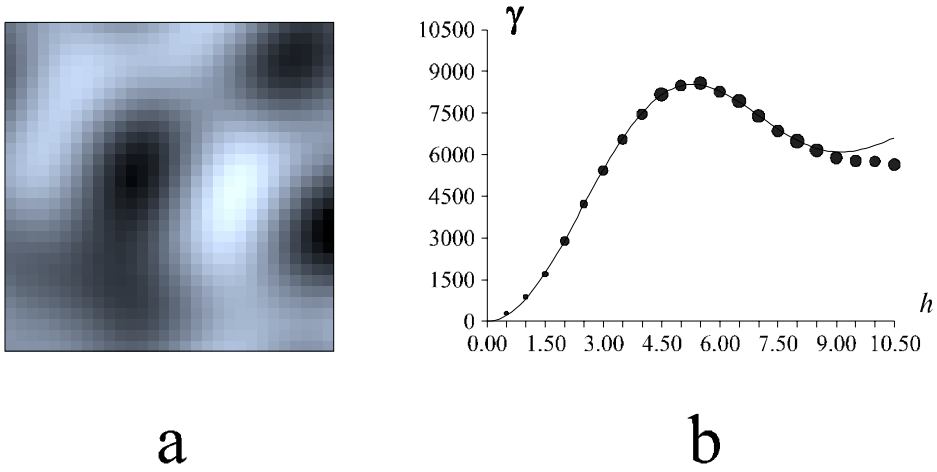
Toutefois, dans un espace à une dimension (ici l'espace des distances h), la partition optimale peut être construite par *programmation dynamique*[2] grâce à l'*algorithme de Fisher* (Fisher 1958). Cet algorithme exploite l'existence de l'ordre sur l'ensemble des distances pour calculer par récurrence une séquence de partitions optimales P_f^i de $\{h_i, h_{i+1}, \ldots, h_m\}$ en f classes. L'algorithme de Fisher peut s'écrire (Diday *et al.* 1982) :

1. Poser $P_1^i = \{h_i, h_{i+1}, \ldots, h_m\}$ pour $i = 1, \ldots, m$.
2. Pour $f = 2, \ldots, k-1$, calculer la partition $P_f^i = \left(\{h_i, h_{i+1}, \ldots, h_j\}, P_{f-1}^{j+1}\right)$ où la partition P_{f-1}^{j+1} à $f-1$ classes de $\{h_{j+1}, \ldots, h_m\}$ a été calculée à l'étape $f-1$ et où j est choisi dans $\{i, i+1, \ldots, m-f+1\}$ de façon à ce que $I(\{h_i, \ldots, h_j\}) + W\left(P_{f-1}^{j+1}\right)$ soit minimum.
3. A l'étape k, construire la partition $P_k^1 = \left(\{h_1, \ldots, h_j\}, P_{k-1}^{j+1}\right)$ où j est choisi dans $\{1, 2, \ldots, m-k+1\}$ de façon à minimiser $I(\{h_1, \ldots, h_j\}) + W\left(P_{k-1}^{j+1}\right)$.

On démontre que la partition finale P_k^1 de $\mathcal{G} = \{h_1, \ldots, h_m\}$ en k classes est optimale au sens de la minimisation de W (Diday *et al.* 1982).

7.1.5.3 Etude de cas

Nous simulons une réalisation d'une FA de modèle Perio$(1, 7999, 5)$, aux noeuds d'une grille 30×30 centrée dans un domaine carré D de 15 unités de côté : la maille de la grille vaut donc $\Delta = 0.5$ (Fig. 7.3.a). Calculons le variogramme local de $z(\cdot)$ pour h compris entre 0.5 et une distance maximale $h_{\max} = 10.5$. Pour tenir compte de toutes les données qui satisfont à $h \leq h_{\max}$ et obtenir un variogramme suffisamment lisse, nous utilisons une tolérance $\varepsilon = 0.25$. Le variogramme local $\gamma_D(\cdot)$ est donc calculé pour $k = 21$ classes $(0.5, 0.25)$. Un modèle périodique $\widetilde{\gamma}_D(\cdot)$ est ajusté à $\gamma_D(\cdot)$ pour $h \leq 8$ (Fig. 7.3.b).



Figure˜7.3: Variable régionalisée simulée sur une grille 30×30 d'après un modèle périodique. (a) Image de la population. (b) Variogramme local et son modèle périodique.

[2]La *programmation dynamique* est une méthode d'optimisation telle que, dans une séquence optimale de décisions, quelle que soit la première décision prise, les décisions suivantes forment une sous-séquence optimale, compte tenu des résultats de la première décision (Sakarovitch 1984).

La qualité de l'ajustement peut être mesurée par l'écart quadratique moyen (MSE) :

$$\mathrm{MSE}\left(\widetilde{\gamma}_D, \gamma_D\right) = \frac{1}{k}\sum_{i=1}^{k}\left\{\widetilde{\gamma}_D\left(h_i\right) - \gamma_D\left(h_i\right)\right\}^2 \tag{7.15}$$

avec h_i la distance représentant la classe C_i. En considérant désormais que le modèle $\widetilde{\gamma}_D(\cdot)$ représente le variogramme local de $z(\cdot)$ sur D, il est possible de juger la qualité d'une estimation $\widehat{\gamma}_D(\cdot)$ en calculant le rapport $R = \mathrm{MSE}\left(\widetilde{\gamma}_D, \widehat{\gamma}_D\right) / \mathrm{MSE}\left(\widetilde{\gamma}_D, \gamma_D\right)$. Dans ce contexte, la valeur $R = 1$ correspond à une connaissance exhaustive ($\widehat{\gamma}_D(\cdot) = \gamma_D(\cdot)$) : pour un échantillon, on a donc nécessairement $R > 1$.

Soit un échantillon systématique s constitué d'une grille 10×10 centrée dans D, de maille $\Delta = 1.5$. L'échantillonnage systématique selon une grille constitue un cas pour lequel le découpage en classes de distances peut être donné directement par la maille de la grille. Nous allons confronter le résultat de cette procédure très simple à deux autres approches : l'utilisation des barycentres des classes de distances à la place des multiples de la maille, et l'optimisation des classes de distances. Par conséquent, nous considérons les variogrammes $\widehat{\gamma}_D^{(i)}(\cdot)$, $i = 1, 2, 3$, calculés d'après s :

1. pour $k = 7$ classes $(1.5, 0.75)$, avec $h_j = 1.5 \times j$, $j = 1, \ldots, k$,
2. comme en 1 mais avec $h_j = G(C_j)$, $j = 1, \ldots, k$, et $G(C_j)$ le barycentre de la classe C_j,
3. pour la partition optimale P^* obtenue en minimisant W grâce à l'algorithme de Fisher, avec $h_{\max} = 11.25$ et $k = 7$.

Les trois variogrammes $\widehat{\gamma}_D^{(1)}(\cdot)$, $\widehat{\gamma}_D^{(2)}(\cdot)$ et $\widehat{\gamma}_D^{(3)}(\cdot)$ respectent tous assez bien $\widetilde{\gamma}_D(\cdot)$ (Fig. 7.4). Cependant, le calcul de R donne $R^{(1)} = 2.7679$, $R^{(2)} = 1.8631$ et $R^{(3)} = 1.8607$. Le classement $R^{(1)} < R^{(2)} < R^{(3)}$ correspond exactement à la prise en compte de plus en plus exacte et complète de la distribution des distances induite par s, dans l'intervalle $[0.5, h_{\max}]$.

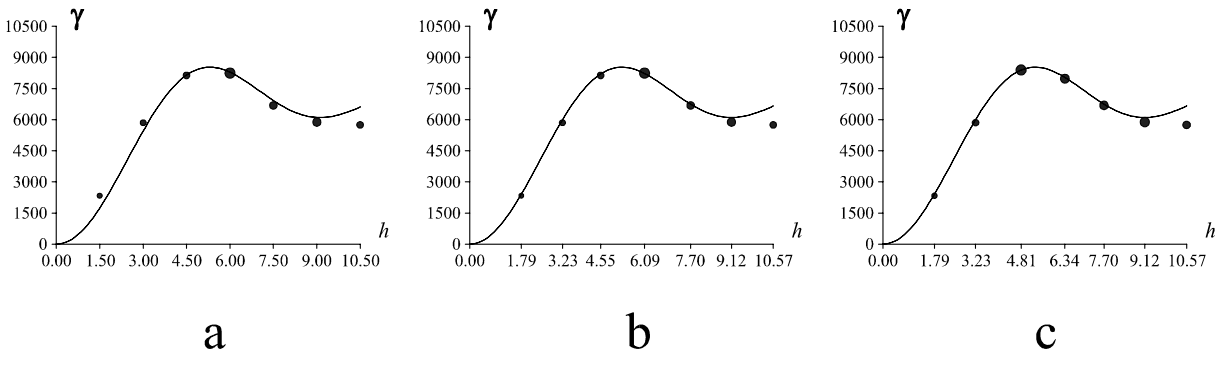


Figure 7.4: Comparaison du modèle de variogramme local et des variogrammes expérimentaux obtenus selon les trois procédures (détails dans le texte). (a) Variogramme $\widehat{\gamma}_D^{(1)}(\cdot)$. (b) Variogramme $\widehat{\gamma}_D^{(2)}(\cdot)$. (c) Variogramme $\widehat{\gamma}_D^{(3)}(\cdot)$.

Dans le cadre d'une grille d'échantillonnage de maille carrée, la recherche de la partition optimale pour k fixé présente évidemment peu d'intérêt comme en témoigne la très faible différence entre $R^{(2)}$ et $R^{(3)}$, mais elle peut s'avérer très utile dans le cas d'un échantillonnage irrégulier pour lequel il est toujours difficile de proposer *a priori* un découpage en classes qui respecte au mieux la distribution des distances. En examinant l'histogramme des distances et en faisant varier le nombre de classes, l'optimisation des classes de distances doit permettre de faciliter l'estimation du variogramme local.

7.1.6 Intégrale du semivariogramme

Le variogramme expérimental usuel $\widehat{\gamma}(\cdot)$ n'est pas entièrement objectif puisqu'il nécessite de définir des classes de distances (Şen 1989). Une solution naturelle consisterait à construire une fonction qui s'affranchirait du découpage en classes de distances. En outre, le variogramme peut être mal défini ou fortement erratique lorsque l'échantillon est petit et/ou lorsque les supports sont répartis de façon très irrégulière. En principe, le variogramme cumulé $\gamma_c(\cdot)$ — calculé en cumulant les valeurs de $\gamma(h)$ pour des distances croissantes — permettrait de pallier ces problèmes puisque :

- le calcul de $\gamma_c(\cdot)$ ne nécessite pas de définir des classes de distances,
- $\gamma_c(\cdot)$ est défini pour davantage de valeurs que le variogramme expérimental $\widehat{\gamma}(\cdot)$,
- $\gamma_c(\cdot)$ présente une allure plus continue que celle de $\widehat{\gamma}(\cdot)$.

Tous ces avantages font que la modélisation du variogramme cumulé serait à la fois plus objective et plus facile que celle de $\widehat{\gamma}(\cdot)$. En fait, l'utilisation de $\gamma_c(\cdot)$ serait surtout utile dans le cas où la répartition des supports est très irrégulière (Şen 1989, 1992, Delay & de Marsily 1994). D'un point de vue qualitatif, le variogramme cumulé permettrait de comparer les nuées variographiques car il est beaucoup plus parlant de comparer des fonctions de répartition que des histogrammes, même si, en théorie, ils contiennent la même information (Chauvet, comm. pers. 1997).

L'utilisation du variogramme cumulé se heurte cependant à des difficultés telles que celles rencontrées dans l'intégration des fonctions fortement oscillantes (Chauvet, comm. pers. 1997). Par ailleurs, il existe une polémique autour du calcul de $\gamma_c(\cdot)$ et de son utilisation pour estimer l'intégrale du variogramme $\gamma_i(\cdot)$ ou pour remplacer le variogramme dans le krigeage (Şen 1989, Şen 1992, Myers 1994c, Şen 1994, Delay & de Marsily 1994).

De la même façon que pour le variogramme, l'intégrale du variogramme $\gamma_i(\cdot)$ peut être définie comme un paramètre d'une superpopulation, *i.e.* en faisant référence au variogramme théorique $\gamma(\cdot)$ (Delay & de Marsily 1994) :

$$\gamma_i(h) = \int_0^h \gamma(u)\,du = \frac{1}{2}\int_0^h \mathrm{Var}_\xi\left[Z(x+u) - Z(x)\right] du \tag{7.16}$$

ou bien, comme nous le proposons, en faisant référence au variogramme local $\gamma_D(\cdot)$, ce qui peut s'écrire :

$$\gamma_i(h) = \int_0^h \gamma_D(u)\,du = \frac{1}{2\left[D \cap \tau_h D\right]}\int_0^h du \int_{D\cap\tau_h D} \left\{z(x+h) - z(x)\right\}^2 dx \tag{7.17}$$

Conformément à l'introduction de cette section, et afin d'éviter tout malentendu, nous préférons faire référence à la définition locale (7.17) plutôt qu'à la définition superpopulationnelle (7.16).

Soit la séquence des $d(h_i) = \frac{1}{2}\{z(x+h_i) - z(x)\}^2$ telle que les m distances h_i induites pas l'échantillon s sont triées par ordre croissant, avec i le rang dans la séquence. Dans le cas d'un motif d'échantillonnage régulier (*e.g.*, une grille de maille carrée), plusieurs valeurs élémentaires $d(h) = \frac{1}{2}\{z(x+h) - z(x)\}^2$ sont définies pour une même distance h, de sorte que nous en calculons la valeur moyenne. Le variogramme cumulé défini dans Şen (1989) se calcule selon :

$$\gamma_c(h_k) = \sum_{i=1}^{k} d(h_i) \tag{7.18}$$

avec $k = 1, \ldots, m$. Cependant, le variogramme cumulé $\gamma_c(\cdot)$ calculé d'après (7.18) ne peut approximer $\gamma_i(\cdot)$ qu'à une constante multiplicative près, et si l'écart entre les distances successives est constant (Delay & de Marsily 1994). Ainsi, il conviendrait plutôt d'utiliser une approximation discrète du type de celles utilisées dans l'intégration numérique, par exemple la méthode des trapèzes (Delay & de Marsily 1994) :

$$\widehat{\gamma}_i(h_k) = \frac{1}{2}\sum_{i=1}^{k-1}\{d(h_i) + d(h_{i+1})\} \cdot (h_{i+1} - h_i) \tag{7.19}$$

pour $k = 2, \ldots, m$. Delay & de Marsily (1994) estiment qu'il est préférable d'utiliser une tolérance afin de regrouper les $d(h)$ définis pour des distances très proches. Bien que cette tolérance soit plus petite que dans le cas du calcul de $\widehat{\gamma}(\cdot)$, nous n'avons pas suivi cette recommandation afin de ne pas introduire à nouveau le choix d'un paramètre arbitraire.

Bazuhair & Şen (1994) suggèrent de modéliser $\gamma_c(\cdot)$ et d'utiliser le modèle dans les calculs du krigeage. Outre le fait que $\gamma_c(\cdot)$ n'estime par correctement $\gamma_i(\cdot)$, Delay & de Marsily (1994) précisent que si l'intégrale du variogramme est substituée au variogramme dans les calculs du krigeage, l'estimateur conserve sa propriété d'universalité (absence de ξ-biais), mais n'est plus optimal dans le modèle. Enfin, en accord avec Delay & de Marsily (1994), et en désaccord avec Şen (1989, 1992), il nous semble très difficile de choisir un type de modèle en examinant uniquement l'allure de $\widehat{\gamma}_i(\cdot)$, de sorte que le calcul de $\widehat{\gamma}(\cdot)$ reste nécessaire pour effectuer ce choix. L'intérêt de l'intégrale du variogramme n'est donc pas de remplacer le variogramme lui-même mais plutôt de venir enrichir la gamme des techniques permettant de le modéliser correctement. Dans cette optique, l'intégrale du variogramme $\gamma_i(\cdot)$ est donc estimée directement à partir de la nuée variographique, sans aucun découpage en classes de distances, puis modélisée par l'intégrale d'un modèle de variogramme afin d'obtenir les valeurs des paramètres du modèle de variogramme correspondant (Annexe F).

7.2 Modélisation du variogramme

Le variogramme expérimental $\widehat{\gamma}(\cdot)$ fournit une estimation du variogramme uniquement pour un ensemble discret de k distances $\{\widehat{\gamma}(h_i) \mid i = 1, \ldots, k\}$. Or toutes les procédures géostatistiques fondées sur le variogramme imposent de pouvoir calculer $\gamma(\cdot)$ pour toute valeur de h, ce qui nécessite d'utiliser un modèle analytique $\widetilde{\gamma}(\cdot)$ à la place de $\widehat{\gamma}(\cdot)$.

7.2.1 Choix du modèle

A des fins de calcul, le choix du modèle de variogramme est tout d'abord contraint par des exigences mathématiques. En effet, le modèle de variogramme doit être conditionnellement semi-défini négatif afin d'assurer que les variances ne sont pas négatives (Section 4.2.2.1). L'utilisation d'un modèle *autorisé* est une condition nécessaire pour garantir l'unicité des solutions des systèmes de krigeage, ainsi que des variances de krigeage non négatives. En outre, la simulation d'une FAST-2 par une méthode directe (*e.g.*, factorisation de Cholesky) requiert que sa matrice de covariance soit semi-définie positive (Annexe D). A des fins de calcul, il n'est donc pas possible d'envisager la modélisation des valeurs $\{\widehat{\gamma}(h_i) \mid i = 1, \ldots, k\}$ par un polynôme quelconque.

A de rares exceptions près (*e.g.*, Anh *et al.* 1997), le choix du type de modèle n'est pas dicté par des considérations concernant la genèse du phénomène étudié (McBratney & Webster 1986, Posa 1989). Pour autant, tous les modèles ne devraient pas être considérés comme plausibles dès lors qu'ils s'ajustent bien au variogramme expérimental. En effet, il faut distinguer deux catégories de modèles selon que leur comportement à l'origine est linéaire (modèles pentasphérique, sphérique et exponentiel) ou parabolique (modèles périodique, gaussien et cubique). Au contraire des modèles de la première catégorie, les modèles de la seconde catégorie sont différentiables à l'origine et correspondent donc à des VR d'assez grande régularité spatiale. Par conséquent, le choix entre les deux catégories de modèles est crucial pour la suite des opérations, qu'il s'agisse d'estimation ou de simulation. En outre, les matrices de covariance associées aux modèles différentiables à l'origine (*e.g.*, modèle gaussien) sont moins bien conditionnées que les autres (Section 6.2.3.2). Ainsi, même lorsque le choix d'un modèle est dicté par la nature du phénomène, il peut s'avérer préférable d'utiliser un modèle conduisant à des procédures plus robustes, notamment dans le cas du krigeage (Posa 1989, Davis & Morris 1997). En accord avec McBratney & Webster (1986), nous considérons que le choix d'un modèle nécessite un bon jugement basé sur la connaissance du phénomène et la compréhension des propriétés mathématiques de chaque fonction. En conséquence, nous ne recommandons pas les procédures de comparaison automatique de modèles (*e.g.*, Jian *et al.* 1996).

Lorsque plusieurs modèles sont comparés, il convient de disposer d'un critère objectif afin de les classer. Des modèles simples ayant le même nombre de paramètres peuvent être comparés en utilisant l'erreur quadratique moyenne :

$$\mathrm{MSE}\left\{\widetilde{\gamma}(h;\theta), \widehat{\gamma}(h)\right\} = \frac{1}{k}\sum_{i=1}^{k}\left\{\widetilde{\gamma}(h_i;\theta) - \widehat{\gamma}(h_i)\right\}^2 \tag{7.20}$$

L'ajustement du modèle $\widetilde{\gamma}(h;\theta)$ à $\widehat{\gamma}(h)$ pour $h \in \{h_i \mid i = 1, \ldots, k\}$ est d'autant meilleur que (7.20) est faible. Il est toujours possible d'améliorer l'ajustement en augmentant le nombre de paramètres du modèle, *i.e.* en le complexifiant. Néanmoins, le principe de parcimonie suggère qu'un modèle simple est préférable à un modèle qui s'ajuste mieux, mais qui est aussi plus complexe. Le compromis entre la qualité de l'ajustement et la simplicité du modèle peut être quantifié à l'aide du critère d'information d'Akaike (AIC), ou du critère d'information Bayesien (BIC) (Webster & McBratney 1989). Le critère d'Akaike est plus souvent utilisé que le BIC (McBratney & Webster 1986, Oliver *et al.* 1989a, Jian *et al.* 1996) bien qu'il ait tendance à sélectionner des modèles plus complexes (Webster & McBratney 1989).

Dans la plupart des cas, il nous semble préférable d'utiliser des modèles simples. En effet, le calcul d'un variogramme expérimental — et qui plus est, souvent omnidirectionnel — n'est pas une opération entièrement objective de sorte que l'on peut s'interroger sur l'opportunité d'un ajustement respectant au mieux les subtilités de $\widehat{\gamma}(\cdot)$. On devrait éviter d'ajuster des modèles emboîtés pour ajuster des variations sans véritable signification (Jian *et al.* 1996). Enfin, les procédures de krigeage s'avèrent très robustes vis-à-vis du modèle, ce qui relativise l'intérêt d'une modélisation très fine de $\widehat{\gamma}(\cdot)$. La question qui se pose alors est essentiellement celle de la validité du modèle pour un objectif donné plutôt que la qualité de l'ajustement.

7.2.2 Validation d'un modèle

Divisons un échantillon comportant n valeurs en deux sous-ensembles disjoints Ω_1 et Ω_2. La *validation croisée* consiste à estimer les valeurs de Ω_1 à partir de celles de Ω_2 au moyen d'une certaine procédure. La qualité de cette procédure est appréciée en comparant les valeurs estimées et les valeurs réelles de Ω_1. Dans le cadre de la validation d'un modèle de variogramme, la procédure d'estimation doit nécessairement utiliser $\widetilde{\gamma}(h;\theta)$, ce qui est notamment le cas du krigeage ordinaire. Dans ce cadre, la validation croisée exploite principalement le fait que le krigeage est un interpolateur exact, et accessoirement l'absence de ξ-biais — le krigeage satisfait à la condition d'universalité $\mathrm{E}_\xi[Z^*(x_i) - Z(x_i)] = 0$ — et la minimisation de la variance d'erreur d'estimation $\sigma_E^2(x_i)$.

La validation croisée considère classiquement que Ω_1 est un singleton et consiste alors à estimer successivement chaque donnée $z(x_i)$ à partir de Ω_2 (voisinage unique) ou d'un sous-ensemble de Ω_2 (voisinage glissant). L'utilisation naïve du krigeage dans le cadre de la validation croisée en voisinage unique est rapidement prohibitive puisqu'elle nécessite de résoudre n systèmes de krigeage. En exploitant la dualité du système de krigeage, Dubrule (1983b) montre qu'il suffit en fait d'inverser une seule matrice $n \times n$. La validation en voisinage unique peut donc s'avérer finalement plus rapide que la validation en voisinage glissant.

L'évaluation de la qualité du modèle peut consister à examiner les résidus $z(x_i) - z^*(x_i)$ comme dans n'importe quelle procédure de régression. Cependant, les résidus sont généralement corrélés d'une façon complexe qui dépend du modèle, ainsi que de la configuration de l'échantillon (Kitanidis 1991). Pour pallier ce problème, Kitanidis (1991) expose une méthode permettant d'obtenir des résidus indépendants, sous l'hypothèse que le modèle est correct. Cet auteur propose également plusieurs statistiques afin de tester l'absence de tendance dans le graphe des résidus. Toutefois, en pratique, ces propositions ne semblent pas être suivies. En général, l'évaluation globale des erreurs d'estimation $z^*(x_i) - z(x_i)$ s'effectue en calculant leur moyenne M, l'erreur absolue moyenne (MAE), l'erreur quadratique moyenne (MSE) ou le coefficient de corrélation de Pearson (r) entre les $z^*(x_i)$ et les $z(x_i)$ (Dubrule 1983b, Russo & Jury 1987, Isaaks & Srivastava 1989, Delay & de Marsily 1994). En outre, la qualité de la variance de krigeage peut être appréciée en calculant (Dubrule 1983b, Russo & Jury 1987, Delay & de Marsily 1994) :

$$R = \frac{1}{n}\sum_{i=1}^{n} \frac{\{z^*(x_i) - z(x_i)\}^2}{\sigma_K^2(x_i)} \tag{7.21}$$

avec $\sigma_K^2(x_i)$ la variance de krigeage au point x_i.

L'idéal serait d'avoir $M = 0$, MAE $= 0$, MSE $= 0$, $r = 1$ et $R = 1$. Cependant, il convient de distinguer l'impact des différents paramètres d'un modèle sur les différentes statistiques. En effet, les estimations ne dépendent pas du seuil du variogramme ($c_0 + c$), mais essentiellement du comportement à l'origine du variogramme, du rapport c_0/c et de la portée a, tandis que la variance de krigeage dépend étroitement du seuil $c_0 + c$. Ainsi, il est possible que de très bonnes estimations soient obtenues bien que le critère R soit médiocre, voire même très mauvais. La validation croisée doit par conséquent être utilisée en faisant référence à l'objectif de la modélisation du variogramme. S'il s'agit d'obtenir la meilleure estimation possible, pour un jeu de données unique, le critère R est sans objet. En revanche, s'il s'agit de calculer des variances d'erreurs d'estimation, le modèle de variogramme doit être satisfaisant à la fois du point de vue des critères mesurant la qualité de l'estimation, et du critère R.

La variance de krigeage σ_K^2 étant très souvent vue comme une mesure de précision de l'estimation (*e.g.*, Lam 1983, Robertson 1987), il peut sembler paradoxal que le critère R qui fait intervenir σ_K^2 soit sans objet lorsqu'il est fait référence à la qualité de l'estimation. Il n'y a cependant aucun paradoxe, simplement parce que $\sigma_K^2(x)$ n'est pas une mesure de précision qui serait à la fois propre à la position x et à la réalisation (*i.e.,* la VR). En effet, la validation croisée permet justement de montrer qu'il n'y a pas de corrélation entre les rangs des résidus absolus et les rangs de σ_K^2 (Journel & Rossi 1989, Aubry 1996b). Lorsque le critère R est proche de 1, cela n'implique pas nécessairement que chaque rapport élémentaire $\{z^*(x_i) - z(x_i)\}^2 / \sigma_K^2(x_i)$ soit proche de 1 : en pratique c'est même rarement le cas. Si R est proche de 1, cela signifie uniquement qu'il y a globalement compensation dans le calcul de R (Aubry 1996b). La variance de krigeage n'est pas une mesure de précision locale, pour une réalisation particulière, mais une mesure d'incertitude calculée en moyenne pour toutes les réalisations de la FA, telle qu'on espère une estimation plus fiable pour un support situé dans une zone bien échantillonnée plutôt que pour un support isolé, situé dans une zone mal connue.

La validation croisée telle qu'elle est pratiquée en géostatistique vise uniquement à juger la qualité du modèle vis-à-vis de l'échantillon (Solow 1990), et pas à réduire le biais de l'estimation du variogramme. Par conséquent, la validation croisée doit être distinguée du jackknife[3] (Davis 1987, Cressie 1991, p. 102), même si ces deux termes sont régulièrement utilisés comme synonymes dans la littérature (*e.g.*, Deutsch & Journel 1992, p. 90).

L'utilisation de la validation croisée pour vérifier la validité d'un modèle ajusté au préalable est encore rare en écologie (*e.g.*, Cardina *et al.* 1995, Aubry & Debouzie 1999a).

7.2.3 Modélisation

La modélisation du variogramme consiste à déterminer les paramètres d'un modèle $\widetilde{\gamma}(h;\theta)$. Nous ne considérons ici que le cas des modèles simples que nous avons retenus, paramétrés par $\theta = (c_0, c, a)^{\mathrm{T}}$ (Annexe F). Déterminer les paramètres du modèle peut s'effectuer de différentes façons selon que le problème est posé en termes :

- de validation croisée,
- d'estimation des paramètres d'un modèle statistique,
- d'ajustement d'un modèle analytique à une courbe expérimentale.

[3] Pour un exposé statistique du jackknife et de la validation croisée, voir Efron & Gong (1983).

7.2.3.1 Modélisation par validation croisée

La validation croisée étant une méthode qui permet de choisir entre plusieurs modèles $\widetilde{\gamma}(h;\theta)$ au sens de certains critères de qualité, il est possible de s'en servir directement pour déterminer les valeurs des paramètres de θ. Dans cette approche, il n'est pas nécessaire de calculer le variogramme, le choix des paramètres étant effectué directement à partir des données, par optimisation d'un critère de qualité. Comme l'optimisation des paramètres utilisant la validation croisée est rapidement prohibitive en temps de calcul (Ripley 1981, p. 57), Marcotte (1995) propose d'utiliser une technique efficace issue de la théorie des splines et connue sous le nom de *validation croisée généralisée* ou GCV (*Generalized Cross-Validation*). Les auteurs qui préconisent l'utilisation de la validation croisée pour déterminer les paramètres de θ considèrent que cette méthode est surtout utile lorsque les échantillons sont petits et que le variogramme expérimental n'est pas bien défini ou pas suffisamment fiable pour faire l'objet d'un ajustement (Lamorey & Jacobson 1995, Marcotte 1995).

7.2.3.2 Modélisation par estimation

Considérons le modèle de fonction aléatoire tel que le vecteur $\mathbf{z}$ de n valeurs observées $z(x_i)$ est vu comme une réalisation d'une superpopulation modélisée par un vecteur $\mathbf{Z}$ de n variables aléatoires $Z(x_i)$. Dans ce cadre, le variogramme — ou ce qui revient au même sous l'hypothèse de stationnarité d'ordre 2, la covariance — est vu comme un paramètre de la superpopulation. Considérons à présent le modèle universel[4] :

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{7.22}$$

avec $\mathbf{X}$ une matrice $n \times p$ de fonctions de bases (*e.g.*, des monômes dans le cas d'une dérive de forme polynomiale), $\boldsymbol{\beta}$ le vecteur $p \times 1$ des coefficients de la dérive, $\boldsymbol{\varepsilon}$ le vecteur $n \times 1$ des résidus. Les n résidus sont de moyenne nulle et de matrice de covariance $\mathbf{C}(\theta)$. Dans ce contexte, θ peut être déterminé en utilisant des techniques classiques d'estimation statistique telles que (Stein 1987) :

- l'estimation au maximum de vraisemblance ou estimation ML (*Maximum Likelihood estimation*),
- l'estimation quadratique de norme minimale ou estimation MINQ (*Minimum Norm Quadratic estimation*).

Estimation au maximum de vraisemblance Déterminer θ peut s'effectuer en termes d'estimation ML (Mardia 1980, Kitanidis 1983, Mardia & Marshall 1984, Cressie 1991, Todini & Ferraresi 1996). Cette méthode ne requiert pas le calcul du variogramme expérimental des résidus $\varepsilon(x_i)$ mais suppose que les $Z(x_i)$ suivent une distribution gaussienne multivariée $\mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{Q}(\theta))$, avec la factorisation $\mathbf{C}(\theta) = \sigma^2\mathbf{Q}(\theta)$. Il s'agit de l'hypothèse par défaut la plus logique, maintenue uniquement pour l'estimation de θ et pas nécessairement pour d'autres opérations (Pardo-Igúzquiza 1998a).

[4] En statistique on parle plutôt de *modèle linéaire généralisé*.

La fonction de log-vraisemblance négative ou NLLF (*Negative Log-Likelihood Function*) s'écrit (Pardo-Igúzquiza 1997a, 1998a) :

$$L\left(\boldsymbol{\beta},\sigma^2,\theta|\ \mathbf{z}\right) = \frac{n}{2}\ln\left(2\pi\right) + n\ln\left(\sigma\right) + \frac{1}{2}\ln|\mathbf{Q}| + \frac{1}{2\sigma^2}\left(\mathbf{z}-\mathbf{X}\boldsymbol{\beta}\right)^{\mathrm{T}}\mathbf{Q}^{-1}\left(\mathbf{z}-\mathbf{X}\boldsymbol{\beta}\right) \quad (7.23)$$

L'estimation ML des paramètres de la dérive $\widehat{\boldsymbol{\beta}}$ est donnée par :

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{z} \quad (7.24)$$

et celle de la variance $\widehat{\sigma}^2$ est donnée par :

$$\frac{1}{n}\left(\mathbf{z}-\mathbf{X}\boldsymbol{\beta}\right)^{\mathrm{T}}\mathbf{Q}^{-1}\left(\mathbf{z}-\mathbf{X}\boldsymbol{\beta}\right) \quad (7.25)$$

Les paramètres estimés peuvent s'obtenir numériquement par minimisation de la NLLF exprimée comme une fonction des paramètres de la corrélation seule (Pardo-Igúzquiza 1997a, 1998a) :

$$L\left(\widehat{\boldsymbol{\beta}},\widehat{\sigma}^2,\theta|\ \mathbf{z}\right) = \frac{n}{2}\left(\ln\left(2\pi\right)+1-\ln\left(n\right)\right)+\frac{1}{2}\ln|\mathbf{Q}|+\frac{n}{2}\left(\mathbf{z}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)^{\mathrm{T}}\mathbf{Q}^{-1}\left(\mathbf{z}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right) \quad (7.26)$$

Lorsque la dérive de la FA n'est pas nulle, l'estimation ML simultanée de $\boldsymbol{\beta}$ et de θ conduit à une estimation biaisée de la covariance (Cressie 1991, Dietrich 1991). Il serait préférable d'utiliser l'estimation du maximum de vraisemblance restreinte ou estimation REML (*Restricted Maximum Likelihood estimation*) qui opère sur des incréments[5] afin de filtrer la dérive (Zimmerman 1989b, Cressie 1991, Dietrich 1991), bien qu'en pratique les deux méthodes puissent donner des résultats comparables (Pardo-Igúzquiza 1998a). Par ailleurs, l'estimation ML est coûteuse en temps de calcul, ce qui conduit à rechercher des méthodes approximatives, plus efficaces (*e.g.*, Pardo-Igúzquiza & Dowd 1998).

Outre la question du temps de calcul, le principal problème posé par l'estimation ML (ou REML) de θ provient de la multimodalité de la fonction de vraisemblance, notamment dans le cas des modèles sphérique (Warnes & Ripley 1987, Mardia & Watkins 1989, Dietrich 1991, Watkins 1992), exponentiel (Warnes & Ripley 1987) et gaussien (Dietrich 1991). La fréquence de la multimodalité est reliée à la discontinuité des dérivées de la fonction de vraisemblance par rapport à la portée (Mardia & Watkins 1989, Dietrich 1991), et elle augmente avec la taille d'échantillon (Dietrich 1991). Ce problème se pose très vraisemblablement pour d'autres modèles que ceux étudiés, notamment les modèles pentasphérique, périodique et cubique. Pour surmonter le problème de la multimodalité lors de la maximisation de la vraisemblance, Pardo-Igúzquiza (1997a) propose d'utiliser une méthode de recherche globale telle que le recuit simulé (Section 8.2.4). Cependant, le maximum global est souvent éloigné de la vraie valeur des paramètres à estimer et la fonction de vraisemblance peut constituer un mauvais résumé de l'information concernant les paramètres de la covariance (Warnes & Ripley 1987, Ripley 1988a, p. 19).

[5] En statistique on parle plutôt de *contrastes*.

Estimation quadratique de norme minimale Déterminer θ peut également s'effectuer en termes d'estimation MINQ (Kitanidis 1983, 1985, Cressie 1991). Cette méthode ne requiert pas le calcul du variogramme expérimental des résidus $\varepsilon(x_i)$ mais nécessite que la covariance soit une fonction linéaire en ses paramètres, de forme générale (Kitanidis 1985, Cressie 1991) :

$$\mathbf{C}(\theta) = \sum_{i=1}^{m} \mathbf{C}_i \theta_i \tag{7.27}$$

Parmi les covariances linéaires en leurs paramètres, citons (Kitanidis 1985) :

- une covariance généralisée de forme polynomiale ou bien spline[6],
- une covariance dont on connaît déjà la structure de corrélation mais pas la variance,
- le modèle sphérique de portée connue.

Les auteurs considèrent généralement la méthode MINQ(U,I), *i.e.* la méthode MINQ sans biais (U, *Unbiased*), et invariante par rapport au vecteur $\boldsymbol{\beta}$ (I, *Invariant*) (Stein 1987, Kitanidis 1985). De plus, si l'hypothèse gaussienne est invoquée, l'estimation MINQ devient l'estimation quadratique de variance minimale ou estimation MIVQ (*Minimum Variance Quadratic estimation*) (Stein 1987). Le principe de la méthode MIVQ(U,I) est d'estimer les m paramètres comme des fonctions quadratiques des données (Kitanidis 1985) :

$$\widehat{\theta}_j = \mathbf{z}^{\mathrm{T}} \mathbf{F}_j \mathbf{z} \tag{7.28}$$

avec $j = 1, \ldots, m$ et $\mathbf{F}_j$ une matrice $n \times n$ sélectionnée telle qu'il y ait absence de biais :

$$\theta_j = \mathrm{E}\left[\widehat{\theta}_j\right] \tag{7.29}$$

et invariance par rapport à $\boldsymbol{\beta}$:

$$\widehat{\theta}_j = \mathbf{z}^{\mathrm{T}} \mathbf{F}_j \mathbf{z} = (\mathbf{z} + \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} \mathbf{F}_j (\mathbf{z} + \mathbf{X}\boldsymbol{\beta}) \tag{7.30}$$

L'estimation MIVQ(U,I) peut également être interprétée comme l'ajustement de $\mathbf{WCW}^{\mathrm{T}}$ à $\mathbf{Wzz}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}$, en minimisant la somme des carrés des écarts de tous les éléments de la matrice $\mathbf{Wzz}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}} - \mathbf{WCW}^{\mathrm{T}}$, avec $\mathbf{W}$ une matrice $n \times n$ telle que la transformation des données $\mathbf{y} = \mathbf{Wz}$ filtre la dérive et donne un vecteur d'incréments autorisés $\mathbf{y}$ (Kitanidis 1985).

Enfin, pour que le modèle de covariance obtenu soit semi-défini positif, il est nécessaire d'imposer des contraintes au cours de la minimisation, par exemple en utilisant la programmation quadratique (Kitanidis 1985).

7.2.3.3 Modélisation par ajustement

La modélisation du variogramme par ajustement d'une fonction analytique au variogramme expérimental est l'approche la plus ancienne et la plus utilisée parce qu'elle est peu exigente en temps de calcul.

[6] L'inférence des covariances généralisées est traitée notamment par Kitanidis (1983), Zimmerman (1989b) et Pardo-Igúzquiza (1997b).

Les géostatisticiens ajustent souvent les modèles visuellement plutôt qu'à l'aide des moindres carrés ou LS (*Least Squares*) (*e.g.*, Journel & Huijbregts 1978, Delay & de Marsily 1994). Néanmoins, l'utilisation de l'ajustement LS tend à se généraliser même si certains auteurs considèrent que cette méthode est trop restrictive (Goovaerts 1999), ou n'offre aucun avantage pas rapport à l'ajustement visuel (Hosseini *et al.* 1994). Il est juste de dire qu'il n'y a aucune évidence que l'ajustement LS donne des modèles davantage conformes au variogramme local que des modèles ajustés visuellement, mais l'ajustement LS présente l'avantage d'éviter l'interprétation individuelle et d'être répétable (Delcourt *et al.* 1996). En conséquence, dans ce qui suit, nous ne considérons pas l'ajustement visuel qui n'est pas suffisamment objectif ni reproductible, mais uniquement l'ajustement LS : cette approche nécessite de choisir un critère à minimiser, ainsi qu'une procédure itérative de minimisation.

Notons $\{\widehat{\gamma}(h_i) \mid i = 1, \ldots, k\}$ l'ensemble des valeurs du variogramme expérimental $\widehat{\gamma}(h)$ et $\{\widetilde{\gamma}(h_i;\theta) \mid i = 1, \ldots, k\}$ l'ensemble des valeurs correspondantes pour le modèle $\widetilde{\gamma}(h;\theta)$. Le principe de l'ajustement aux moindres carrés généralisés ou GLS (*Generalized Least Squares*) consiste à minimiser la quantité :

$$[\widehat{\gamma}(h) - \widetilde{\gamma}(h;\theta)]^{\mathrm{T}} \mathbf{V}^{-1} [\widehat{\gamma}(h) - \widetilde{\gamma}(h;\theta)] \tag{7.31}$$

où $\mathbf{V}$ est la matrice de variance-covariance des résidus.

Dans ce cadre formel général, l'ajustement au sens des moindres carrés ordinaires ou OLS (*Ordinary Least Squares*) considère $\mathbf{V} = \sigma^2\mathbf{I}$, avec $\mathbf{I}$ la matrice identité. Les moindres carrés ordinaires supposent donc que les résidus sont indépendants et de variance constante σ^2. Dans le cas du variogramme, ces hypothèses ne sont généralement pas satisfaites. Une approche plus satisfaisante consisterait à utiliser les moindres carrés généralisés. Comme $\mathbf{V}$ est inconnue[7], il faut recourir à une procédure itérative de mise-à-jour de $\mathbf{V}$ à partir d'une estimation initiale $\mathbf{V}_0$, ce qui nécessite une inversion matricielle à chaque itération.

Afin d'obtenir un compromis entre la méthode des moindres carrés ordinaires qui est simple mais inadaptée, et la méthode des moindres carrés généralisés qui n'est pas très efficace, il suffit de considérer uniquement la diagonale principale de $\mathbf{V}$, ce qui revient à négliger la dépendance des résidus mais à tenir compte des différentes variances. Cette approche, connue comme l'ajustement aux moindres carrés pondérés ou WLS (*Weighted Least Squares*), consiste à minimiser (*e.g.*, Goovaerts 1997, p. 105) :

$$\sum_{i=1}^{k} \omega(h_i) \cdot \{\widehat{\gamma}(h_i) - \widetilde{\gamma}(h_i;\theta)\}^2 \tag{7.32}$$

avec $\omega(h_i) = 1/\sigma_i^2$ et σ_i^2 la variance de la VA modélisant le résidu $\{\widehat{\gamma}(h_i) - \widetilde{\gamma}(h_i;\theta)\}$. Comme les variances σ_i^2 sont inconnues, d'autres poids heuristiques ont été proposés afin de pondérer les moindres carrés :

- David (1977) propose de pondérer directement par $N(h_i)$,
- Beliaeff & Cochard (1995) utilisent une pondération par l'inverse de h_i,

[7] Dans le cas de l'estimation du variogramme d'une FA Gaussienne à partir d'un variogramme expérimental, la matrice de variance-covariance admet cependant une formulation explicite (*cf.* Genton 1998b, Bogaert & Russo 1999).

- Jian *et al.* (1996) choisissent une pondération par l'inverse de la variance des $N(h_i)$ valeurs utilisées dans le calcul de $\widehat{\gamma}(h_i)$,
- Brus *et al.* (1996) pondèrent par l'inverse de $N(h_i)$ (ce choix trouvera sa justification dans la Section 7.3.2.1, p. 218).

Un autre schéma — obtenu après de longs développements statistiques — consiste à minimiser (Cressie 1985) :

$$\sum_{i=1}^{k} \frac{N(h_i)}{\{\widetilde{\gamma}(h_i;\theta)\}^2} \{\widehat{\gamma}(h_i) - \widetilde{\gamma}(h_i;\theta)\}^2 \tag{7.33}$$

De même que la pondération par $N(h_i)$, ce schéma favorise les classes de distances pour lesquelles le nombre de valeurs est le plus élevé, mais tend également à favoriser les faibles distances dans la mesure où $\gamma(h_i)$ est faible lorsque h_i est faible (Gotway 1991, Zhang *et al.* 1995). Cette pondération présente néanmoins certains inconvénients. En premier lieu, la croissance monotone de $\gamma(\cdot)$ n'est pas une propriété partagée par les modèles périodiques et par conséquent, la pondération en $\{\widetilde{\gamma}(h_i;\theta)\}^{-2}$ décroît, croît, décroît, et ainsi de suite jusqu'à l'amortissement. Il convient également de remarquer (Zhang *et al.* 1995) :

- que le poids est lui-même une fonction des paramètres à déterminer,
- que la somme des poids diffère d'une itération à l'autre,
- que le poids d'un résidu absolu $|\widehat{\gamma}(h_i) - \widetilde{\gamma}(h_i;\theta)|$ diffère selon qu'il est positif ou négatif.

De même que Zhang *et al.* (1995), Goulard (1988) considère que la pondération proposée par Cressie (1985) n'est pas bonne parce qu'elle dépend des paramètres à estimer. Compte tenu des inconvénients du schéma de pondération (7.33), il conviendrait plutôt de minimiser (Zhang *et al.* 1995) :

$$\sum_{i=1}^{k} \frac{N(h_i)}{h^{\lambda}} \{\widehat{\gamma}(h_i) - \widetilde{\gamma}(h_i;\theta)\}^2 \tag{7.34}$$

où λ est un exposant déterminé selon l'application. Le schéma de pondération de (7.34) présente cependant l'inconvénient d'avoir à choisir λ.

La pondération proposée par Cressie (1985) et recommandée par McBratney & Webster (1986) semble poser des problèmes lors de l'ajustement WLS. Gotway (1991) considère que les problèmes sont dus à la variabilité des valeurs du variogramme et propose de considérer uniquement les premières classes de distances, tandis que Zhang *et al.* (1995) affirment que la forme de la pondération est la cause d'une convergence lente, souvent vers un optimum local, voire même d'une divergence. Quoi qu'il en soit, les propriétés de la pondération proposée par Cressie (7.33) ne plaident pas en sa faveur, et il faudrait effectivement lui préférer celle de (7.34). En fait, le choix le plus classique consiste à pondérer simplement par $N(h_i)$ (*e.g.*, Pelletier & Parma 1994), en considérant que la fiabilité de $\widehat{\gamma}(h_i)$ est d'autant plus grande que $N(h_i)$ est grand (Webster & McBratney 1989, Goovaerts 1997, p. 105).

Afin de faire la part des problèmes d'ajustement qui proviennent de la procédure utilisée de ceux qui proviennent du schéma de pondération lui-même, il convient avant tout de fixer la procédure d'ajustement. Parmi les méthodes d'ajustement non linéaire, la méthode de Levenberg-Marquardt est reconnue comme celle qui donne les meilleurs résultats (Bard 1970). De même que Jian *et al.* (1996), mais de façon indépendante, nous avons utilisé dans Aubry (1996b) la version de la méthode de Levenberg-Marquardt proposée par Press *et al.* (1989).

La méthode de Levenberg-Marquardt nécessite de calculer les dérivées partielles du modèle par rapport à ses paramètres (Press *et al.* 1989). Bates & Watts (1988) recommandent d'utiliser des dérivées analytiques[8] plutôt que numériques pour une question de précision. Haas (1990) propose d'ajuster automatiquement un modèle identifié au préalable à partir de valeurs initiales des paramètres obtenues grâce à des heuristiques. Dans le même esprit, Jian *et al.* (1996) proposent d'ajuster en série tout un ensemble de modèles concurrents à partir de valeurs initiales des paramètres calculées automatiquement, puis de sélectionner le meilleur modèle au sens du critère d'Akaike. Nous recommandons plutôt un ajustement interactif autorisant (Aubry 1996b) :

- le choix des valeurs initiales des paramètres, qui conditionne le succès de l'ajustement non linéaire, *i.e.* sa convergence vers un optimum global (Bates & Watts 1988),
- la possibilité de fixer la valeur d'un paramètre tout en ajustant les autres (Press *et al.* 1989) ; il s'agit en pratique essentiellement de la pépite c_0,
- un ajustement prenant en compte les distances h_i en deçà d'une distance limite $h_{\max}$ (Lamorey & Jacobson 1995).

Dans la majorité des cas, un choix grossier des valeurs initiales des paramètres par examen visuel du variogramme expérimental suffit à garantir la convergence de la méthode de Levenberg-Marquardt. Néanmoins, à l'issue d'une tentative d'ajustement automatique, il se peut qu'un paramètre soit négatif. En particulier, il n'est pas exceptionnel que la pépite c_0 soit négative, ce qui confirme bien que la valeur de la pépite est due en partie au choix du modèle, ainsi qu'à son ajustement, qu'il soit manuel ou automatique (Marcotte 1995, Atkinson 1996). Dans le cas où un paramètre ajusté automatiquement s'avère négatif, sa valeur est forcée à zéro, et l'ajustement est de nouveau effectué pour les autres paramètres (Haas 1990, Aubry 1996b).

En dernier recours, dans les cas les plus pathologiques pour lesquels il est impossible d'obtenir la convergence de l'algorithme de Levenberg-Marquardt, il reste toujours possible de fixer manuellement la valeur des paramètres. Dans ce cas, il devient intéressant de disposer de critères pour choisir les valeurs des paramètres. Par défaut, la pépite peut être définie comme dans Jian *et al.* (1996) :

$$c_0 = \max\left[0, \widehat{\gamma}(h_1) - \frac{h_1}{h_2 - h_1}\left(\widehat{\gamma}(h_2) - \widehat{\gamma}(h_1)\right)\right] \tag{7.35}$$

autrement dit, comme l'ordonnée à l'origine de la droite passant par les deux premiers points du variogramme expérimental, du moins lorsque cette ordonnée est positive. Le choix de la portée peut être effectué en examinant le p-gramme associé au variogramme (Section 3.5.5).

[8] Les dérivées partielles des modèles de variogrammes utilisés dans ce mémoire ainsi que celles des intégrales des modèles sont données dans l'Annexe F.

En ce qui concerne le seuil $c_0 + c$, il est classiquement identifié à la variance d'échantillon s^2_{n-1} (*e.g.*, David 1977, Rossi *et al.* 1995). Cependant, Barnes (1991) fait remarquer que s^2_{n-1} surestime ou sous-estime le seuil du variogramme en fonction du motif d'échantillonnage. Si les supports sont concentrés dans une zone de taille inférieure ou égale à la portée, l'estimateur s^2_{n-1} sous-estime le seuil et ne doit pas être utilisé comme heuristique lors de la modélisation du variogramme. Au contraire, si les supports sont répartis dans un vaste domaine, s^2_{n-1} surestime le seuil apparent (Barnes 1991). En conséquence, nous ne recommandons pas d'identifier automatiquement s^2_{n-1} avec le seuil du variogramme, mais plutôt d'utiliser l'appréciation visuelle.

7.2.3.4 Choix d'une méthode de modélisation

Le choix d'une méthode de modélisation nécessite d'abord un examen *a priori* des avantages et des inconvénients des différentes propositions, puis un examen *a posteriori*, effectué grâce à des études de cas basées sur des données réelles ou simulées. Nous ne considérons pas ici les propriétés asymptotiques des différents estimateurs, sujet qui intéresse surtout les statisticiens et qui est généralement éloigné des préoccupations des biométriciens "de terrain". Signalons cependant que dans le contexte spatial, les propriétés asymptotiques de l'estimateur ML sont étudiées par Mardia & Marshall (1984), et celles de l'estimation MINQ(U,I) par Stein (1987).

La modélisation par validation croisée est considérée comme une alternative à l'ajustement LS lorsque le variogramme n'est pas bien défini ou pas fiable, *i.e.* lorsque l'échantillon est petit (Lamorey & Jacobson 1995, Marcotte 1995), ou bien comme une optimisation d'un ajustement LS effectué au préalable (Phillips & Marks 1996). Afin d'éviter d'introduire d'autres sources de variation que celles qui concernent les données, nous considérons que la validation croisée est effectuée par krigeage ordinaire, en voisinage unique.

Le recours à la validation croisée comme méthode de modélisation du variogramme soulève quelques critiques. En effet, s'il est toujours possible d'optimiser un modèle vis-à-vis d'un certain critère de qualité, le problème est de savoir si cette pratique a un sens. D'abord, la modélisation dépend du critère retenu et par conséquent, la validation croisée est avant tout un outil dépendant de l'objectif (Isaaks & Srivastava 1989, pp. 364-368). Ensuite, toute tentative pour pallier le manque de fiabilité du variogramme expérimental dans le cas des petits échantillons grâce à une boîte noire qui détermine à la fois le modèle et ses paramètres est vouée à l'échec, parce que le problème du manque d'information reste entier. Enfin, dans le cadre de l'estimation locale par krigeage, il ne suffit pas que le modèle décrive correctement les données comme la validation croisée permet de s'en assurer, car l'objectif est justement d'interpoler en dehors des données, celles-ci n'étant pas nécessairement très représentatives du domaine d'étude (Goovaerts 1997, p. 106).

Isaaks & Srivastava (1989, p. 533) montrent comment un modèle de variogramme optimisé par validation croisée conduit à des résultats qui sont pires que ceux obtenus avec le modèle originel (Isaaks & Srivastava, pp. 514-517). En outre, Aubry (1996b) donne un exemple de modèle optimisé par validation croisée, au sens du critère R (toutes choses égales par ailleurs), qui ne passe même pas à travers le variogramme expérimental[9].

[9] Un autre exemple de modèle optimisé par validation croisée qui ne respecte pas le variogramme expérimental se trouve dans Brannan & Hamlett (1998, Fig. 5).

Bien que l'utilisation de la validation croisée pour déterminer automatiquement le type de modèle de variogramme ainsi que ses paramètres ne soit pas récente — puisqu'elle est déjà signalée par David (1977) — cette approche n'a toujours pas fait ses preuves et nous ne la recommandons pas, bien qu'en pratique elle soit parfois utilisée (*e.g.*, Maravelias *et al.* 1996, Rahman *et al.* 1996, Brannan & Hamlett 1998).

Les méthodes d'estimation ML ou MINQ considèrent que les paramètres à estimer sont ceux d'une superpopulation[10]. En conséquence, les hypothèses qui sont faites concernent la superpopulation et ne peuvent pas être testées *a priori* à partir d'une seule réalisation (*i.e.,* la VR), par exemple la normalité multivariée dans le cas de l'estimation ML (Pardo-Igúzquiza 1998a).

La dépendance cruciale de l'estimation ML vis-à-vis de l'hypothèse gaussienne conduit Cressie (1985, 1991) à rejeter cette approche. Néanmoins, Pardo-Igúzquiza (1998a) justifie l'utilisation de la distribution gaussienne multivariée en faisant référence à la méthode du maximum d'entropie et donne une interprétation heuristique cohérente de ce choix. En outre, le modèle gaussien multivarié est le seul modèle qui puisse être construit de façon consistante pour tous les modèles de covariance (Armstrong 1992, Journel 1992).

D'autres aspects de l'estimation ML sont certainement plus problématiques. En premier lieu, la multimodalité des fonctions de vraisemblance des modèles de variogrammes usuels révélée par Warnes & Ripley (1987), Ripley (1988a), Mardia & Watkins (1989) et Dietrich (1991), constitue une propriété défavorable. En second lieu, la complexité de la procédure d'estimation ML (ou REML) la rend très coûteuse en temps de calcul[11], et son utilisation se révèle prohibitive pour $n > 150$ (Mardia & Marshall 1984, Pardo-Igúzquiza & Dowd 1998). Par exemple, Knotters *et al.* (1995) cherchent à estimer les paramètres de la covariance généralisée d'une FAI-2 en utilisant la méthode REML, mais le nombre des supports s'avère trop élevé ($n = 117$). En conséquence, ces auteurs divisent leurs données en deux sous-ensembles, estiment les paramètres séparément pour chacun des sous-ensembles, puis calculent la moyenne des coefficients. Bien que l'évolution des ordinateurs devrait permettre de résoudre ce type de problème à plus ou moins court terme, l'estimation ML est actuellement limitée au cas des petits échantillons pour lesquels le variogramme expérimental ne s'avère pas suffisamment bien défini ou fiable. Néanmoins, il est légitime de douter de l'intérêt de la géostatistique lorsque les données sont trop peu nombreuses pour calculer un variogramme expérimental bien défini, si ce n'est fiable. Le risque est d'obtenir des résultats assez subjectifs, davantage dus au modèle de superpopulation utilisé qu'à l'information réellement disponible.

L'estimation MINQ n'est pas toujours applicable et ne présente vraiment d'intérêt que dans le cas des FAI-0 strictes et des FAI-k pour $k > 0$. Toutefois, Cressie (1991, p. 94) considère qu'il n'est pas raisonnable de penser que la combinaison linéaire de certaines fonctions de covariance généralisée (polynomiales ou splines) soit suffisamment flexible pour approximer n'importe quelle matrice de covariance $\mathbf{C}(\theta)$. En fait, l'estimation MINQ serait généralisable aux modèles non linéaires en leurs paramètres bien que cette voie n'ait pas été poursuivie (Kitanidis 1985, Stein 1987).

[10] Ce peut aussi être le cas pour l'ajustement GLS (Genton 1998b) ou WLS (Beckers & Bogaert 1998).

[11] Dans le cas de supports organisés selon une grille régulière, il est cependant possible d'exploiter la structure Toeplitz par bloc de la matrice de covariance généralisée afin d'économiser du temps de calcul (*cf.* Zimmerman 1989b).

L'ajustement du modèle $\widetilde{\gamma}(h;\theta)$ à $\widehat{\gamma}(h)$ pour $h \in \{h_i \mid i = 1, \ldots, k\}$ en utilisant la procédure de régression non linéaire de Levenberg-Marquardt ne présente pas d'inconvénient dès lors que $\widehat{\gamma}(h)$ a été calculé dans les règles de l'art. Le schéma de pondération le plus populaire reste certainement la pondération directe par $N(h)$ (Webster & McBratney 1989). Cette méthode ne nécessite aucune hypothèse et donne généralement des résultats très satisfaisants tout en étant relativement facile à mettre en oeuvre pour n'importe quel modèle de variogramme autorisé.

Dans les exemples qu'ils considèrent, Russo & Jury (1987) constatent que les valeurs estimées du seuil et de la portée sont plus élevées avec la méthode REML qu'avec la méthode ML, et plus élevées avec la méthode ML qu'avec la méthode LS, mais que les résultats sont tous consistants avec les données. Dans une étude de Monte-Carlo, Zimmerman & Zimmerman (1991) comparent les performances de sept méthodes de type ML, REML, MIVQ et LS. Zimmerman & Zimmerman (1991) concluent que les estimations par ajustement LS sont relativement faciles à calculer et sont généralement aussi bonnes que celles du type ML, REML ou MIVQ, plus coûteuses en temps de calcul.

Enfin, la même méthode d'ajustement WLS peut servir à ajuster l'intégrale du variogramme : il suffit pour cela d'adapter le schéma de pondération. En plus de l'interactivité de l'ajustement, cette extension augmente encore les possibiltiés d'obtenir un bon modèle, au sens d'un certain objectif (krigeage, calcul de la variance d'erreur d'estimation, simulation). En conséquence, nous utilisons par la suite uniquement l'ajustement interactif par la méthode de Levenberg-Marquardt.

7.2.4 Etude de cas

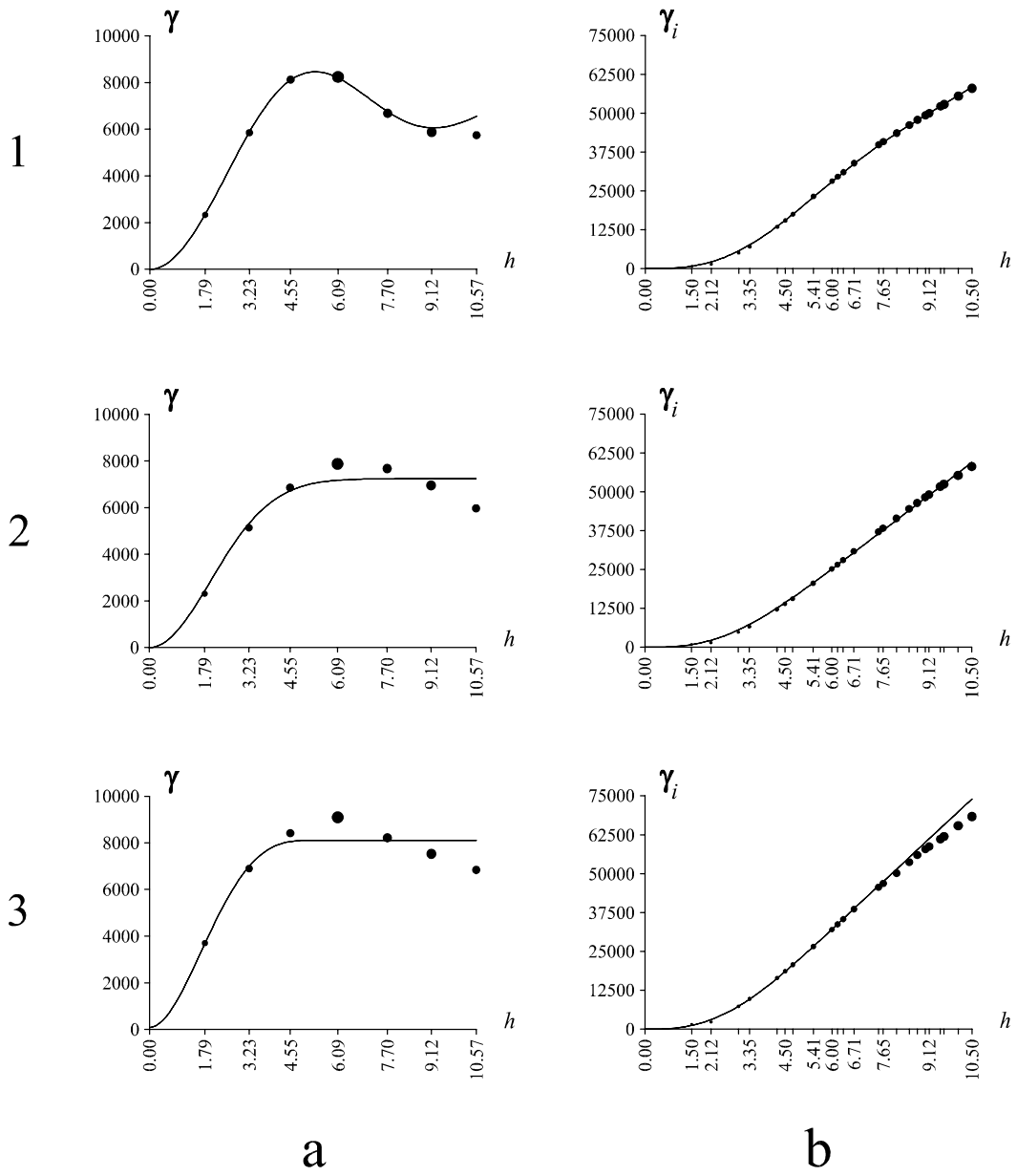
Afin d'illustrer la modélisation du variogramme par ajustement WLS, nous considérons six VR simulées sur une grille 30×30, de maille carrée $\Delta = 0.5$, centrée dans un domaine carrée D de $L = 15$ unités de côté. Chaque VR correspond à l'un des six modèles que nous avons retenus : périodique, gaussien, cubique, sphérique, pentasphérique et exponentiel. Les modèles sont tous paramétrés par $\theta = (1, 7999, 5)$.

Nous examinons la question de la modélisation du variogramme local et ne cherchons pas à identifier le modèle théorique, essentiellement parce que nous n'avons aucun moyen objectif de juger de l'écart entre un modèle et le variogramme théorique lorsque nous connaissons uniquement une réalisation (*i.e.*, la VR). Toutefois, afin de limiter l'étude de cas, nous utilisons le type de modèle correspondant à chaque simulation.

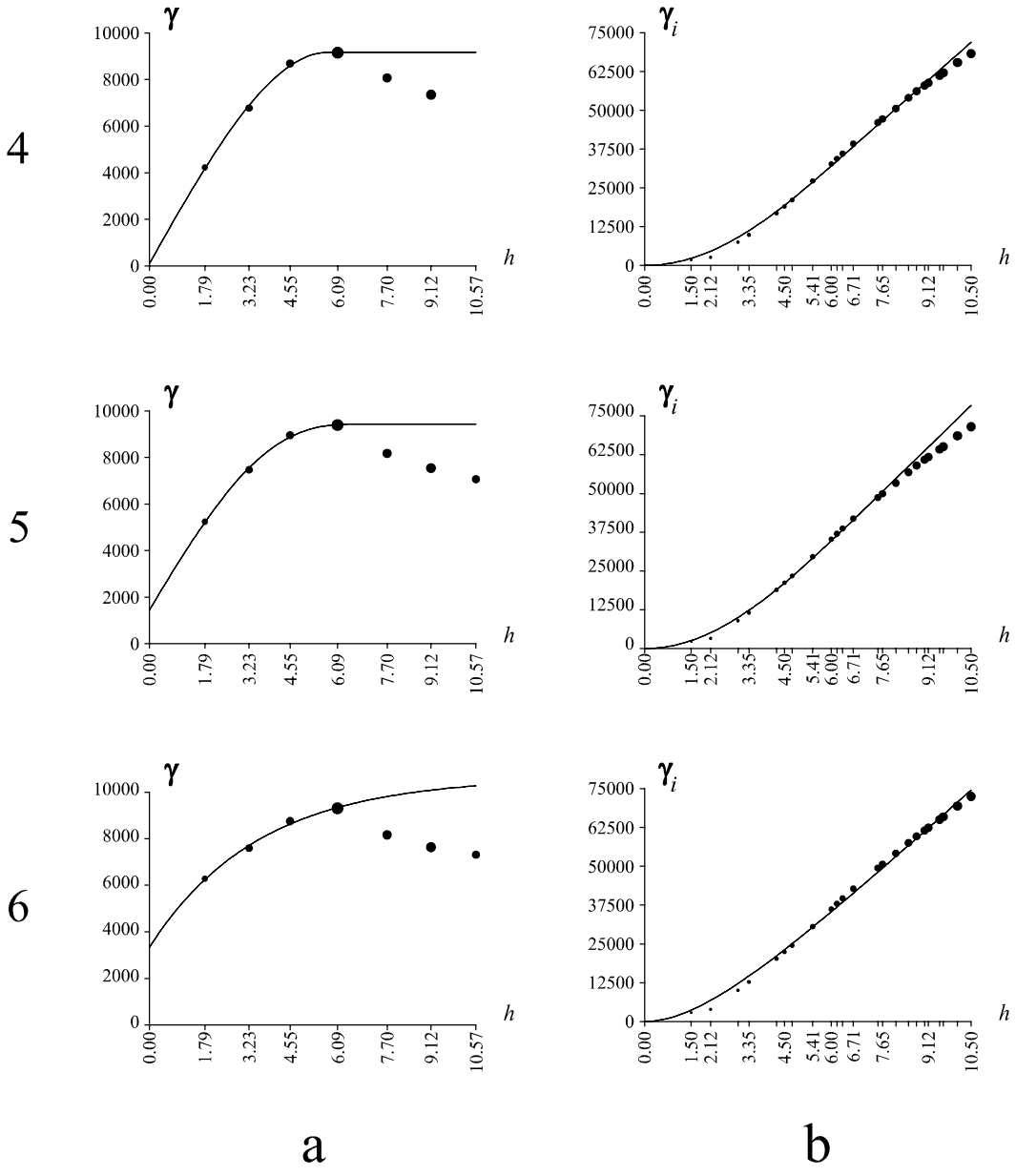
Chaque VR est échantillonnée par un motif d'échantillonnage systématique (ES) selon une grille 10×10 centrée dans D, et par un motif d'échantillonnage aléatoire simple (EAS) de taille $n = 100$. Les motifs ES et EAS sont les mêmes pour les six VR, et les distances induites respectivement par chaque type d'échantillon sont par conséquent identiques.

Pour les ES et les EAS des six VR, le variogramme expérimental est calculé pour $k = 7$ classes $(1.5, 0.75)$. Chaque point $\widehat{\gamma}(h_i)$ est situé au barycentre de la classe correspondante. L'intégrale du variogramme est estimée par (7.19). L'ajustement WLS des modèles de variogrammes (SV) ou d'intégrales de variogrammes (ISV) est effectué par la méthode de Levenberg-Marquardt, avec un paramétrage par défaut $\theta_0 = (0, 8000, 5)$, et une des quatre distances maximales $h_{\max} \in \{6.10, 7.70, 9.125, 10.60\}$.

Pour chaque échantillon, la validation croisée par krigeage ordinaire en voisinage unique est utilisée afin de choisir parmi les quatre modèles SV, puis parmi les quatre modèles ISV, ceux qui conduisent à la meilleure prédiction des données (Fig. 7.5 & 7.6).



Figure~7.5: Modèles du variogramme et de son intégrale conduisant à la meilleure prédiction des données par krigeage ordinaire en voisinage unique (cas de l'échantillon systématique). (1) Modèle périodique. (2) Modèle gaussien. (3) Modèle cubique. (a) Variogramme. (b) Intégrale du variogramme.



Figure˜7.6: Modèles du variogramme et de son intégrale conduisant à la meilleure prédiction des données par krigeage ordinaire en voisinage unique (cas de l'échantillon systématique). (4) Modèle sphérique. (5) Modèle pentasphérique. (6) Modèle exponentiel. (a) Variogramme. (b) Intégrale du variogramme.

Les critères utilisés pour classer les modèles sont donc les critères M, MAE, MSE et r. Le critère R est calculé mais n'est pas utilisé dans le choix du meilleur modèle SV ou ISV.

En pratique, le critère M ne s'avère pas très utile par rapport aux critères MAE, MSE et r (Tab. 7.1, 7.2, 7.3 & 7.4). En considérant le coefficient de corrélation r, on vérifie que la prédiction par krigeage est excellente pour les VR spatialement régulières, avec même $r > 0.99$ dans le cas du modèle périodique (Tab. 7.1). A l'autre extrémité, *i.e.* pour le modèle exponentiel, $r = 0.62$ pour l'EAS, et $r = 0.46$ pour l'ES (Tab. 7.4 & 7.2). Par ailleurs, on vérifie également qu'une bonne prédiction ne s'accompagne pas nécessairement d'une valeur de R proche de 1. Par exemple, dans le cas du modèle gaussien, la prédiction peut s'avérer très bonne avec $r \simeq 0.98$, tandis que $R \simeq 73.34$ (Tab. 7.1). Pour l'ES, les performances des modèles SV et ISV se révèlent équivalentes (Tab. 7.1 & 7.2), tandis que pour l'EAS, les modèles ISV sont toujours un peu meilleurs que les modèles SV (Tab. 7.3 & 7.4). Il est vraisemblable que cet écart doit s'accentuer lorsque la modélisation SV s'effectue dans des conditions de moins en moins favorables, mais cela reste à vérifier.

7.3 Précision du variogramme

Le problème de la précision du variogramme est un sujet qui est abordé de façon récurrente depuis Matheron (1965, Chapitre XIII). Cependant, il s'agit d'un problème difficile pour lequel on dispose, encore actuellement, de relativement peu de résultats généraux et bien fondés (revue dans Brus & de Gruijter 1994).

L'objet de cette section est de discuter succinctement de l'influence de la taille d'échantillon sur la précision du variogramme et de présenter quelques résultats concernant le problème du calcul d'un intervalle de confiance pour chaque valeur du variogramme. Nous considérons de façon classique l'estimateur de Matheron (7.1), ainsi que la modélisation par ajustement WLS, en utilisant la méthode de Levenberg-Marquardt (Section 7.2.3.3).

7.3.1 Influence de la taille de l'échantillon

Afin d'estimer le variogramme — omnidirectionnel — avec suffisamment de confiance, une règle classique préconise de disposer d'au moins 30 à 50 valeurs $d(h)$ pour tout $h \leq L_{\max}/2$ (*e.g.*, Journel & Huijbregts 1978, Shotzko & O'Keeffe 1989, Liebhold *et al.* 1993). Webster & Oliver (1992a, 1992b) considèrent que cette règle donne un faux sens de sécurité parce qu'elle aurait été établie dans $\mathbb{R}$ plutôt que dans $\mathbb{R}^2$.

En utilisant une population fictive générée par simulation non conditionnelle, Webster & Oliver (1992a, 1992b, 1993) concluent qu'une taille d'échantillon $n = 50$ ne permet pas d'estimer correctement le variogramme, que $n = 100$ constitue un minimum, et qu'il serait préférable d'avoir $n = 150$, voire même $n = 200$. Gascuel-Odoux & Boivin (1994) considèrent que le problème de l'échantillonnage pour l'estimation du variogramme ne doit pas être examiné uniquement à l'aide de données simulées. Les auteurs préfèrent donc examiner un jeu de données réel de taille $n = 561$ dont le variogramme est utilisé comme référence. Gascuel-Odoux & Boivin (1994) tirent indépendamment 5 séries de 20 sous-échantillons aléatoires, pour 5 cardinalités différentes $n \in \{50, 75, 100, 150, 200\}$.

	Périodique		gaussien		Cubique	
	SV	ISV	SV	ISV	SV	ISV
$h_{\max}$	7.70	10.60	10.60	10.60	10.60	6.10
M	0.01	−0.01	−0.03	0.01	−0.26	−0.06
MAE	1.87	2.11	7.85	11.45	33.14	32.01
MSE	6.41	9.17	90.33	215.54	1757.05	1601.60
r	0.99	0.99	0.99	0.98	0.87	0.88
R	2.03	2.88	6.73	73.34	1.77	3.34

Tableau 7.1: Résultats de la validation croisée par krigeage ordinaire en voisinage unique pour les modèles spatialement réguliers et l'ES (détails dans le texte).

	Sphérique		Pentasphérique		Exponentiel	
	SV	ISV	SV	ISV	SV	ISV
$h_{\max}$	6.10	9.12	6.10	7.70	6.10	10.60
M	−0.12	−0.12	−0.14	−0.10	−0.12	−0.12
MAE	46.29	46.29	53.82	54.39	61.93	63.40
MSE	3161.44	3128.51	4395.89	4485.87	5998.48	6222.83
r	0.75	0.75	0.65	0.64	0.46	0.45
R	1.07	1.24	0.96	1.56	0.94	1.38

Tableau 7.2: Résultats de la validation croisée par krigeage ordinaire en voisinage unique pour les modèles spatialement irréguliers et l'ES (détails dans le texte).

	Périodique		gaussien		Cubique	
	SV	ISV	SV	ISV	SV	ISV
$h_{\max}$	6.10	7.70	6.10	7.70	6.10	7.70
M	0.17	0.12	0.64	0.64	0.00	−0.20
MAE	7.26	4.11	8.61	8.61	26.62	22.40
MSE	100.32	36.23	256.37	256.31	1475.26	996.79
r	0.99	0.99	0.97	0.97	0.88	0.92
R	0.10	0.08	8.26	8.32	0.68	1.32

Tableau 7.3: Résultats de la validation croisée par krigeage ordinaire en voisinage unique pour les modèles spatialement réguliers et l'EAS (détails dans le texte).

	Sphérique		Pentasphérique		Exponentiel	
	SV	ISV	SV	ISV	SV	ISV
$h_{\max}$	6.10	10.60	10.60	10.60	6.10	6.10
M	0.08	−0.46	0.03	−0.72	−0.43	−0.81
MAE	41.58	40.12	47.32	45.95	53.46	53.33
MSE	3133.78	2964.44	3977.51	3753.23	4817.88	4816.07
r	0.76	0.77	0.68	0.70	0.62	0.62
R	0.71	1.06	0.79	1.17	0.96	1.13

Tableau 7.4: Résultats de la validation croisée par krigeage ordinaire en voisinage unique pour les modèles spatialement irréguliers et l'EAS (détails dans le texte).

Chaque série de 20 sous-échantillons permet de calculer un variogramme moyen qui est comparé au variogramme de référence. Les modèles ajustés aux variogrammes sont également comparés entre eux. En accord avec Webster & Oliver (1992a, 1992b), Gascuel-Odoux & Boivin (1994) recommandent de prélever des échantillons de taille $n = 150$ ou $n = 200$, mais précisent qu'en général $n = 100$ est une taille d'échantillon considérée comme suffisante pour l'analyse géostatistique. D'après notre expérience en écologie, $n = 100$ constitue effectivement une taille d'échantillon très correcte pour estimer le variogramme. En revanche, une taille aussi faible que $n = 9$ (Vischetti *et al.* 1997), $n = 11$ (Wingle & Poeter 1993), $n = 19$ (Chang *et al.* 1998), $n = 20$ (Legendre *et al.* 1989) ou $n = 26$ (Joffre *et al.* 1996) ne devrait raisonnablement pas donner lieu à l'utilisation de la géostatistique, à moins de pallier le manque de données par une expertise du phénomène et du domaine échantillonné (Wingle & Poeter 1993, Carle & Fogg 1996, Goovaerts 1999).

L'estimation fiable des variogrammes unidirectionnels en vue de modéliser l'anisotropie demande évidemment davantage de données que celle du variogramme omnidirectionnel. Il est parfois recommandé de disposer d'un échantillon de taille minimale $n = 300$ (Oliver *et al.* 1989b). Par exemple $n = 529$ points sont utilisés par Guertal & Elkins (1996) dans leur étude de la variation spatiale des radiations photosynthétiquement actives dans une serre. Cook & Coles (1997) insistent sur le fait que plusieurs centaines de supports sont nécessaires pour étudier l'anisotropie, ce qui représente un effort d'échantillonnage élevé. L'anisotropie exhibée par des variogrammes unidirectionnels calculés à partir de petits échantillons (*e.g.*, Chang *et al.* 1998, Fig. 3) peut être considérée comme purement artefactuelle (Goovaerts 1997). Par exemple, il ne nous semble pas raisonnable de calculer des variogrammes directionnels pour 8 classes d'angles à partir d'un échantillon comportant seulement $n = 33$ supports comme le fait Holdaway (1996). Dans le cas des petits échantillons, il faut se contenter de calculer le variogramme omnidirectionnel, même lorsque l'anisotropie est certaine[12] (*e.g.*, Lecoustre *et al.* 1989).

Il nous semble cependant impossible de discuter davantage de la taille d'échantillon requise *a priori* pour une estimation fiable du variogramme indépendamment du motif d'échantillonnage spatial. Dans ce qui suit, nous tenons compte du motif d'échantillonnage et considérons par défaut que l'isotropie constitue une hypothèse raisonnable, et que l'échantillon est de taille $n = 100$.

7.3.2 Intervalle de confiance du variogramme

Le calcul classique d'un intervalle de confiance pour chaque valeur $\widehat{\gamma}(h)$ nécessite de pouvoir lui associer une variance $\sigma^2_{\widehat{\gamma}(h)}$ et de recourir à une distribution statistique de référence, *e.g.* la loi normale. Cependant, il est essentiel de distinguer les niveaux de représentativité de l'échantillon afin de savoir quelle est la source d'erreur qui est considérée (Section 5.3.3.2).

Dans la littérature, la question de la précision de $\widehat{\gamma}(\cdot)$ est parfois traitée par simulation non conditionnelle d'un modèle de variogramme (*e.g.*, Russo & Jury 1987). Cette approche indique que les auteurs s'intéressent davantage à l'erreur de fluctuation qu'à l'erreur d'estimation. Parfois, les deux types de variations sont considérés simultanément

[12] Ceci n'empêche pas de réintroduire la connaissance *a priori* de cette anisotropie et d'adopter un modèle de variogramme qui en tienne compte (Goovaerts 1999).

(*e.g.*, Morris 1991). D'autres situations sont moins claires. Par exemple, Webster & Oliver (1992a, 1992b) proposent de calculer l'intervalle de confiance du variogramme en rééchantillonnant une population générée par simulation non conditionnelle à partir d'un modèle plausible de variogramme. Webster & Oliver (1992a, 1992b) se réfèrent explicitement à l'erreur d'estimation du variogramme local, mais utilisent néanmoins comme référence le variogramme théorique ayant servi à simuler leurs populations fictives (images 256×256). Il serait de toute façon impossible de calculer les variogrammes locaux de façon classique compte tenu de la taille de leurs images[13] (une image 256×256 induit plus de 2×10^9 distances). Brus & de Gruijter (1994) considèrent que l'approche de Webster & Oliver (1992a, 1992b) ne peut pas être utilisée en pratique, parce que :

- le modèle de variogramme utilisé dans la simulation est obtenu à partir du variogramme expérimental,
- il n'y a aucune garantie que le variogramme local de la réalisation simulée soit égal au modèle utilisé.

En dehors de l'étude d'un modèle de superpopulation, la question de l'erreur de fluctuation n'est pas objective en ce que, pour une variable régionalisée donnée $z(\cdot)$, il est souvent impossible d'invoquer l'immanence d'un variogramme théorique $\gamma(\cdot)$ sous-jacent au variogramme local $\gamma_D(\cdot)$. Ainsi, dans ce qui suit nous ne faisons pas référence à l'erreur de fluctuation mais bien à l'erreur d'estimation de $\gamma_D(\cdot)$ par son approximation discrète $\widehat{\gamma}_D(\cdot)$. Pour l'estimateur classique (7.1), nous considérons successivement :

- l'intervalle de confiance défini dans le cadre *design-based*,
- l'intervalle de prédiction calculé selon l'approche *model-based*,
- l'intervalle de confiance du jackknife.

7.3.2.1 Approche design-based

Dans le cadre de l'inférence *design-based* du variogramme, on considère exclusivement l'erreur d'estimation de $\gamma_D(\cdot)$ par son approximation discrète $\widehat{\gamma}(\cdot)$ puisque le concept de superpopulation n'intervient pas.

Brus & de Gruijter (1994) utilisent la théorie de l'échantillonnage probabiliste afin d'estimer la variance d'échantillonnage $\sigma^2_{\widehat{\gamma}(h)}$ associée à chaque valeur $\widehat{\gamma}(h)$ (l'indice D est omis car il n'y a pas de confusion possible avec l'estimation du variogramme théorique). La population statistique échantillonnée n'est plus la population des supports $\mathcal{U}$ mais le produit cartésien $\mathcal{U} \times \mathcal{U}$. Autrement dit, la solution proposée par Brus & de Gruijter (1994) nécessite de recourir à un certain dispositif d'échantillonnage probabiliste (*e.g.*, l'EAS) portant sur l'ensemble des couples de supports. A notre connaissance, il n'existe pas de théorie *design-based* permettant d'estimer $\sigma^2_{\widehat{\gamma}(h)}$ sur la base d'un dispositif d'échantillonnage portant sur la population spatiale $\mathcal{U}$ elle-même.

Nous proposons néanmoins quelques résultats élémentaires obtenus grâce à une simulation du même type que celle utilisée dans le cas de la moyenne globale (Section 6.3.1.4, p. 157). Comme nous nous contentons de traiter du variogramme omnidirectionnel, nous utilisons le modèle isotrope de la demi-sphère centrée dans un domaine carré 1×1 (Section 6.3.1.4, p. 157).

[13] Le calcul de la fonction d'autocorrélation ou du variogramme dans le cas des images 256×256 peut être traité dans le domaine spectral grâce à la FFT (Pfleiderer *et al.* 1993, Marcotte 1996).

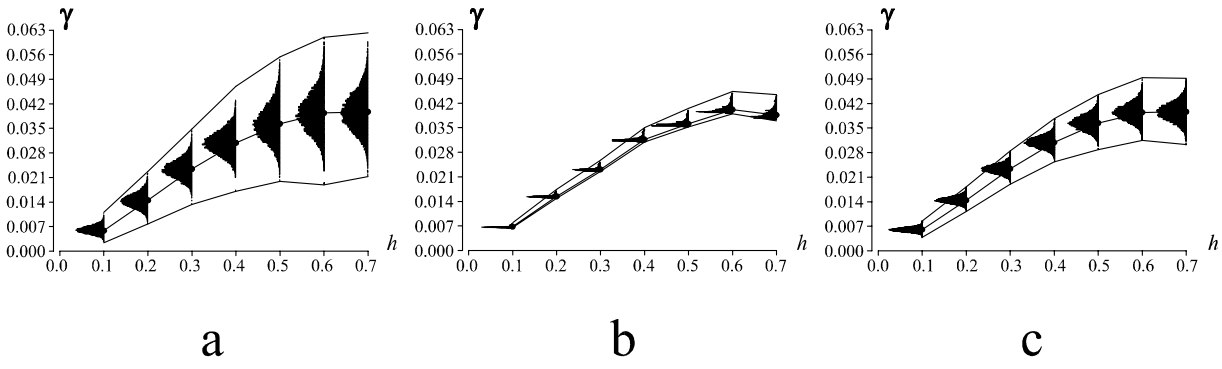
Les supports échantillonnés sont ponctuels et la population est donc infinie. Trois dispositifs d'échantillonnage sont comparés :

- échantillonnage aléatoire simple avec $n = 100$ (EAS),
- échantillonnage systématique selon une grille 10×10 (ES),
- échantillonnage stratifié selon une grille 10×10, à un point par maille (STR).

Dans chaque cas, la distribution d'échantillonnage du variogramme $\widehat{\gamma}(\cdot)$ est approximée par une distribution empirique dérivée de 10^4 réplications du dispositif d'échantillonnage (Fig. 7.7). Les résultats montrent que :

- l'ES s'avère plus précis que le STR, lui même plus précis que l'EAS,
- la distribution d'échantillonnage est asymétrique dans le cas de l'ES, et symétrique dans le cas du STR et de l'EAS,
- dans tous les cas, l'imprécision de l'estimation augmente avec la distance.

Ainsi, un estimateur de la variance d'échantillonnage $\sigma^2_{\widehat{\gamma}(h)}$ pour une valeur $\widehat{\gamma}(h)$ doit évidemment tenir compte du type de dispositif.



Figure˜7.7: Distribution d'échantillonnage du variogramme $\widehat{\gamma}(\cdot)$ approximée par une distribution empirique dérivée de 10^4 réplications du dispositif d'échantillonnage. (a) Dispositif EAS. (b) Dispositif ES. (c) Dispositif STR.

Par ailleurs, l'imprécision de l'estimation a tendance[14] à augmenter avec le nombre de valeurs $N(h)$ utilisées dans le calcul de $\widehat{\gamma}(h)$ (Fig. 7.8). Cette tendance peut sembler contraire à l'intuition puisqu'il est classique de considérer que l'estimation d'un paramètre θ et d'autant plus précise que le nombre de valeurs utilisées dans le calcul de la valeur estimée θ^* est grand. En fait, $N(h)$ est lié à h en fonction de la définition des classes et de l'ensemble des distances $\mathcal{G}$ induit par le motif d'échantillonnage. La distribution des distances dans $\mathcal{G}$ approxime la densité de probabilité $f_D(h)$ des distances dans D. Même si certains motifs impliquent une mauvaise approximation de $f_D(h)$ (*e.g.*, une grille de points), les caractéristiques générales de $f_D(h)$ sont respectées. Or, la fonction $f_D(h)$ (ici pour un domaine carré) croît lorsque h augmente, jusqu'à un maximum inférieur à $L_{\max}/2$ (Fig. 7.1, p. 193).

[14] Cette tendance n'est pas testée à cause de l'autocorrélation entre les écarts-types.

Ainsi, la tendance de $\sigma_{\widehat{\gamma}(h)}$ à croître lorsque $N(h)$ augmente ne fait que traduire de façon indirecte et imparfaite l'augmentation de $\sigma_{\widehat{\gamma}(h)}$ avec la distance h. En ce qui concerne l'ajustement WLS d'un modèle de variogramme à $\widehat{\gamma}(h)$, ce résultat suggère qu'une pondération en $[N(h)]^{-1}$ serait davantage justifiée que celle en $N(h)$, cette pondération étant précisément celle utilisée par Brus *et al.* (1996). Néanmoins, en utilisant une tolérance de la forme $\varepsilon = 0.5 \times \Delta$, les valeurs de $N(h)$ sont généralement du même ordre de grandeur, et la relation positive entre $\sigma_{\widehat{\gamma}(h)}$ et $N(h)$ n'étant pas linéaire (Fig. 7.8), ce changement de pondération nous semble présenter peu d'intérêt en pratique.

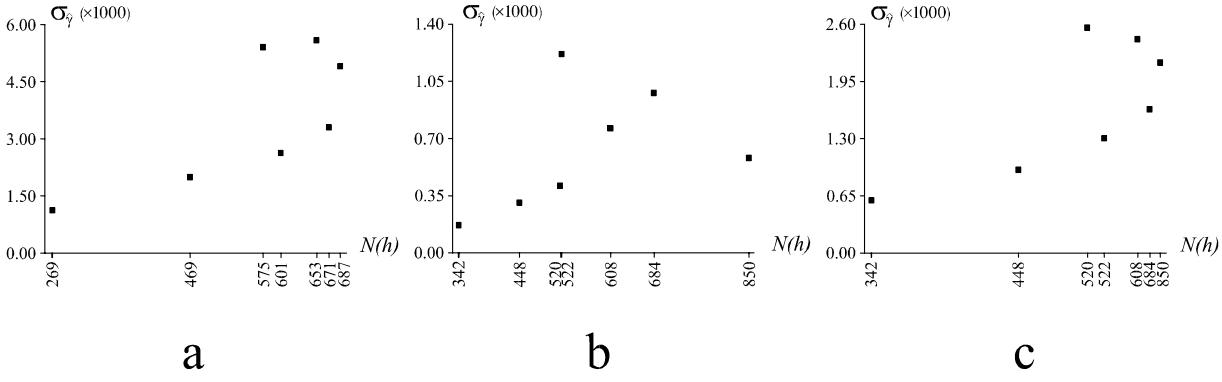


Figure 7.8: Relation entre $\sigma_{\widehat{\gamma}(h)}$ et $N(h)$. (a) Dispositif EAS. (b) Dispositif ES. (c) Dispositif STR.

7.3.2.2 Approche model-based

Dans le cadre de l'inférence *model-based*, nous proposons de recourir à la simulation d'une fonction aléatoire afin de calculer un intervalle d'estimation de $\gamma_D(h)$. Le modèle numérique de la superpopulation est construit en identifiant :

- la moyenne z_D à l'espérance $\mathrm{E}_\xi[Z_D]$,
- le variogramme local $\gamma_D(\cdot)$ au variogramme théorique $\gamma(\cdot) = \mathrm{E}_\xi[\gamma_D(\cdot)]$.

En pratique, z_D et $\gamma_D(\cdot)$ sont estimés respectivement par z_D^* et $\widehat{\gamma}_D(\cdot)$, et le modèle de variogramme $\widetilde{\gamma}(\cdot)$ utilisé dans les calculs dérive exclusivement de $\widehat{\gamma}_D(\cdot)$.

La procédure d'inférence *model-based* peut sembler circulaire dans la mesure où l'intervalle d'estimation de $\widehat{\gamma}_D(\cdot)$ est calculé grâce à un modèle $\widetilde{\gamma}(\cdot)$ construit à partir de $\widehat{\gamma}_D(\cdot)$ lui-même. D'une façon générale, l'inférence *model-based* de la précision d'une valeur estimée θ^* est circulaire toutes les fois où le modèle est construit en utilisant exclusivement θ^*, *i.e.* sans données ou hypothèses supplémentaires. Il est en effet théoriquement impossible de déduire des mêmes données à la fois une valeur estimée et la précision de cette valeur (Matheron 1965).

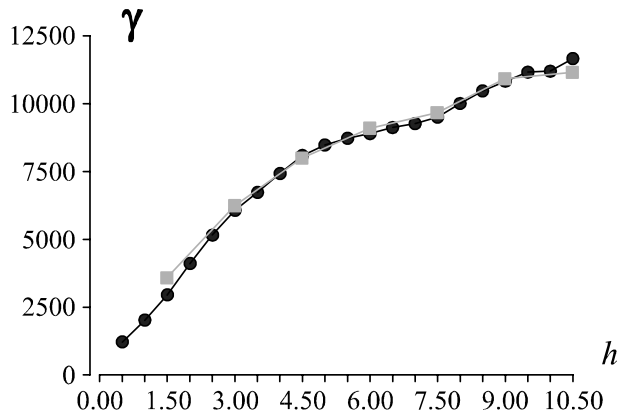
Dans le cas de l'intervalle d'estimation de la moyenne globale z_D, nous avons utilisé la simulation conditionnelle afin de calculer une variance tenant compte de $\widetilde{\gamma}(\cdot)$, ainsi que de toutes les valeurs de l'échantillon (Section 6.3.1.5).

Le conditionnement des simulations par les valeurs permettait à la fois :

- de calculer une variance conditionnelle,
- d'accroître la robustesse du modèle, notamment vis-à-vis de l'erreur de spécification du modèle de variogramme local.

Afin de sortir de la circularité de l'inférence *model-based* de l'intervalle d'estimation $\gamma_D(\cdot)$, nous conditionnons également les réalisations de $Z(\cdot)$ en utilisant toutes les valeurs de l'échantillon. Cependant, le conditionnement ne garantit pas que le variogramme moyen $\mathrm{E}_\xi[\gamma_D(\cdot)]$ calculé dans le cadre de la simulation conditionnelle de $Z(\cdot)$ estime mieux $\gamma_D(\cdot)$ que ne le fait $\widehat{\gamma}_D(\cdot)$ lui-même, car ce sont les mêmes valeurs qui servent à calculer $\widehat{\gamma}_D(\cdot)$ et à conditionner la simulation. La circularité qui apparaît à nouveau ne peut être rompue qu'en introduisant davantage d'information dans le modèle.

Etude de cas Considérons la population C utilisée dans l'étude de cas de l'inférence de la moyenne globale (Section 6.3.1.8, p. 171). La population C est définie sur une grille 30×30 de maille $\Delta = 0.5$. L'échantillonnage systématique selon une grille 10×10 de maille $\Delta = 1.5$ conduit à 9 échantillons possibles. Le variogramme local $\gamma_D(\cdot)$ est calculé pour $k = 21$ classes $(0.5, 0.25)$ tandis que les variogrammes expérimentaux sont calculés pour $k = 7$ classes $(1.5, 0.75)$. Bien que cette procédure ne soit pas optimale (Section 7.1.5, p. 197), la moyenne des 9 estimations $\widehat{\gamma}_D(\cdot)$ obtenues à partir des 9 échantillons systématiques se superpose pratiquement à $\gamma_D(\cdot)$, ce qui témoigne de ce que l'échantillonnage systématique est sans p-biais (Fig. 7.9).



Figure˜7.9: Superposition du variogramme expérimental moyen obtenu à partir des neuf échantillons systématiques (figuré en gris) et du variogramme local (figuré en noir).

Considérons les estimations extrêmes de $\gamma_D(\cdot)$ (non figurées) :

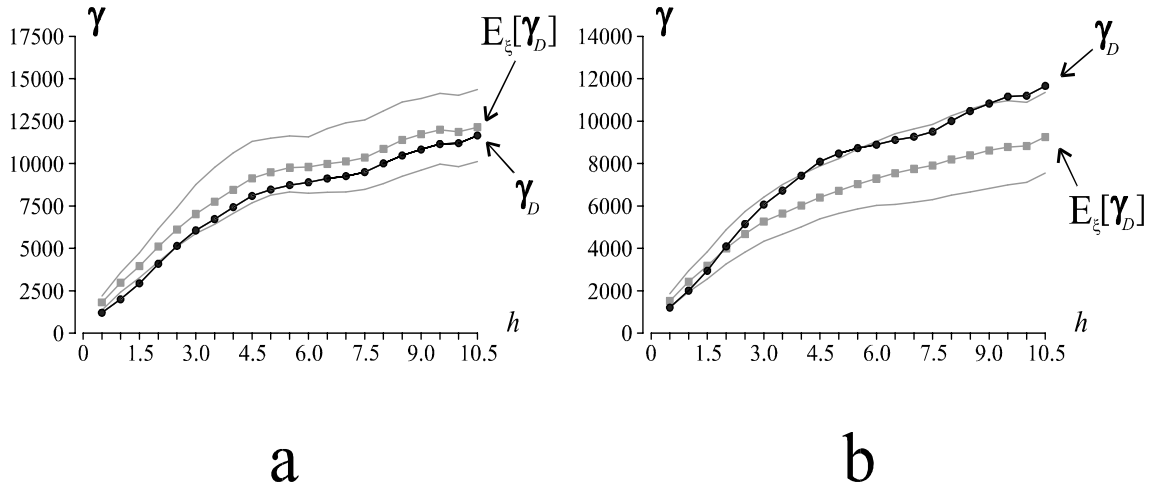
- la plus forte surestimation est due à l'échantillon s_5 d'origine 5 (Fig. 6.8, p. 174),
- la plus forte sous-estimation est due à l'échantillon s_8 d'origine 8.

Les échantillons s_5 et s_8 subissent une anamorphose gaussienne (Annexe E), puis les variogrammes $\widehat{\gamma}_D(\cdot)$ sont calculés sur les données transformées et des modèles de type exponentiel sont ajustés automatiquement par la méthode de Levenberg-Marquardt.

Chaque modèle est simulé 10^4 fois conditionnellement aux échantillons correspondants. Chaque réalisation subit une anamorphose réciproque et le variogramme local correspondant $\gamma_D(\cdot)$ est calculé. Trois variogrammes synthétiques peuvent résumer la distribution des 10^4 variogrammes locaux :

- le variogramme formé par les valeurs moyennes, $\mathrm{E}_\xi[\gamma_D(\cdot)]$,
- le variogramme formé par les valeurs les plus basses, $\gamma_{\min}(\cdot)$,
- le variogramme formé par les valeurs les plus élevées, $\gamma_{\max}(\cdot)$.

Qu'il s'agisse des échantillons s_5 ou s_8, le variogramme moyen $\mathrm{E}_\xi[\gamma_D(\cdot)]$ est plus bas que $\widehat{\gamma}_D(\cdot)$ (non figuré). Ceci compense la surestimation dans le cas de s_5, mais accentue la sous-estimation dans le cas de s_8 (Fig. 7.10). Ce phénomène est peut-être causé par le fait que l'anamorphose gaussienne est pratiquement impossible lorsque les données comportent une forte proportion de valeurs nulles (Rivoirard 1991, p. 41). Quoi qu'il en soit, dans les deux cas l'intervalle d'estimation de $\gamma_D(h)$ obtenu à partir des valeurs extrêmes $[\gamma_{\min}(h), \gamma_{\max}(h)]$ s'avère assez précis. Néanmois, $\gamma_D(\cdot)$ est légèrement en dehors de "l'enveloppe de prédiction", ce qui témoigne d'un certain manque de robustesse quant à la validité de ses limites (Fig. 7.10).



Figure˜7.10: Enveloppe de prédiction (figurée en gris) du variogramme local (figuré en noir) obtenue par simulation conditionnelle. (a) Cas de la plus forte surestimation (échantillon s_5). (b) Cas de la plus forte sous-estimation (échantillon s_8).

7.3.2.3 Jackknife

Afin de disposer d'une approximation de l'intervalle de confiance de chaque valeur du variogramme $\widehat{\gamma}(h)$, Shafer & Varljen (1990) suggèrent de recourir à la technique du jackknife. Cette technique, initialement proposée pour réduire le biais d'un estimateur $\widehat{\theta}$ (Quenouille 1956), est également utilisée pour calculer un intervalle de confiance lorsque la distribution d'échantillonnage de $\widehat{\theta}$ n'est pas connue *a priori* (Tukey 1958). Nous nous intéressons ici uniquement à l'estimation d'un intervalle de confiance pour chaque valeur $\widehat{\gamma}(h)$.

Considérons un échantillon s comportant n supports $s = \{s_i \mid i = 1, \ldots, n\}$, partitionné en g groupes disjoints G, chacun de taille $m = n/g$. Pour une classe de distances, écrivons $\widehat{\gamma}_{-j}(h)$ la valeur de $\widehat{\gamma}(h)$ calculée sans prendre en compte les supports du groupe G_j avec $j = 1, \ldots, g$. Des *pseudo-valeurs* sont définies comme $\widehat{\gamma}_j(h) = g\widehat{\gamma}(h) - (g-1)\widehat{\gamma}_{-j}(h)$ pour $j = 1, \ldots, g$. La moyenne et la variance du jackknife dans le cas général où m valeurs sont supprimées à la fois (m-jackknife) s'écrivent :

$$\bar{\gamma}_J(h) = \frac{1}{g}\sum_{j=1}^{g}\widehat{\gamma}_j(h) \tag{7.36}$$

$$s_J^2(h) = \frac{1}{g-1}\sum_{j=1}^{g}\left\{\widehat{\gamma}_j(h) - \bar{\gamma}_J(h)\right\}^2 \tag{7.37}$$

Tukey (1958) propose d'utiliser la variance de jackknife afin de calculer un intervalle de confiance. L'intervalle de confiance de $\widehat{\gamma}(h)$ est calculé autour de l'estimateur du jackknife $\bar{\gamma}_J(h)$ avec la variance :

$$s^2_{\widehat{\gamma}(h)} = \frac{s_J^2(h)}{g} \tag{7.38}$$

Le calcul de l'intervalle de confiance peut faire référence à la loi normale $\mathcal{N}(0,1)$. La loi normale est souvent remplacée par le t de Student à $g-1$ degrés de liberté, où g est le nombre de pseudo-valeurs, mais il n'existe cependant pas de support théorique à cette pratique (Hinkley 1983). Il est classique de considérer que les pseudo-valeurs sont indépendantes (*e.g.*, Tomassone *et al.* 1993, p. 75) mais Hinkley (1983) fait remarquer qu'en général elles sont corrélées. Ce qui rend l'utilisation du jackknife séduisante, c'est que les limites de l'intervalle de confiance sont assez robustes, au sens de leur validité (Hinkley 1983). Les limites sont valides lorsque le paramètre à estimer est approximable de façon précise par une moyenne (Hinkley 1983). Le jackknife est généralisé aux U-statistiques pourvu que les données soient indépendantes et identiquement distribuées (Miller 1974). L'estimateur $\widehat{\gamma}(h)$ de Matheron (7.1) a la forme d'une U-statistique mais le jackknife de $\widehat{\gamma}(h)$ (Shafer & Varljen 1990) soulève néanmoins quelques interrogations.

Premièrement, le jackknife suppose que les données sont indépendantes, ce qui est évidemment contredit par l'autocorrélation spatiale. A ce propos, Miller (1974) note que le jackknife a peu de succès dans le domaine des séries temporelles. *A fortiori*, le problème de la validité du jackknife dans un contexte spatial reste largement à explorer (Cressie 1991, pp. 491-492).

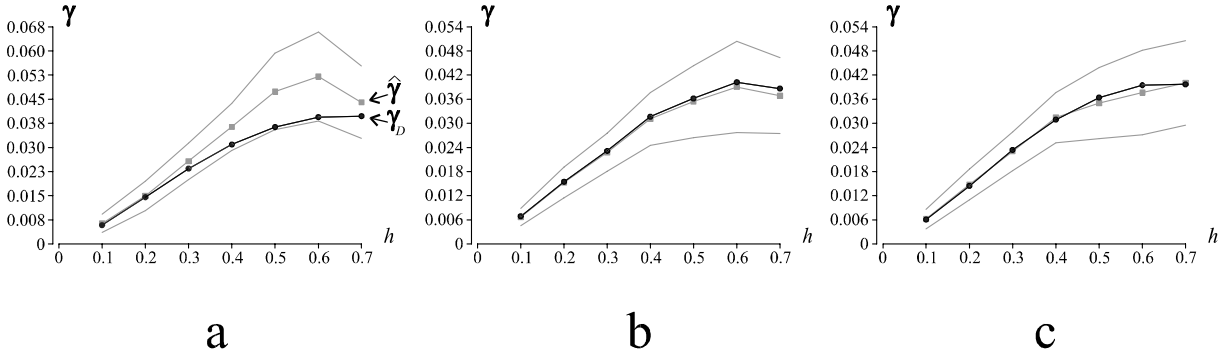
Deuxièmement, le variogramme a la forme d'une variance et de ce fait, certaines pseudo-valeurs peuvent être négatives comme c'est le cas pour le jackknife de σ^2, et de trop fréquentes pseudo-valeurs négatives peuvent introduire des estimations déformées (Miller 1974). En transformant la variance σ^2 en $\log(\sigma^2)$ une pseudo-valeur négative correspond à une petite variance (Miller 1974), et l'estimation devient plus précise (Matloff 1980).

Enfin, le choix de g présente l'inconvénient de donner une procédure qui n'est pas définie de façon unique (Hinkley 1983). Pour des raisons de temps de calcul, Shafer & Varljen (1990) cherchent le plus petit nombre de groupes correct ($g \ll n$). Les auteurs évaluent la stabilité de la variance (7.38) comme une fonction de g afin de déterminer le nombre de groupes pour un échantillon particulier, et posent le problème de l'indépendance des

pseudo-valeurs correspondant aux partitions envisagées. Pour g fixé, il est possible qu'une certaine partition optimale P^* permette de minimiser la corrélation des pseudo-valeurs, mais envisager cette possibilité d'un point de vue pratique nécessiterait de trouver une solution à un problème d'optimisation combinatoire certainement assez difficile, et de fixer arbitrairement g.

Le cas $g = n$ ($m = 1$) élimine quant à lui toute forme d'arbitraire dans la formation des groupes et constitue vraisemblablement la meilleure forme de jackknife, à utiliser dans n'importe quel problème (Miller 1974). L'utilisation de l'algorithme naïf consistant à supprimer une valeur des données puis à recalculer le variogramme en entier est particulièrement inefficace. Aussi, nous avons conçu un algorithme de jackknife optimisé pour lequel le cas $g = n$ est traité très rapidement. Nous ne traitons donc par la suite que du cas $g = n$ ou 1-jackknife.

Etudes de cas Il est possible d'apprécier la validité du jackknife du variogramme en considérant trois échantillons obtenus par échantillonnage EAS, ES et STR, du modèle de la demi-sphère centrée dans un domaine carré 1×1. Quel que soit le dispositif d'échantillonnage, l'intervalle de confiance à 95% calculé à partir de la variance (7.38) contient bien le variogramme local (Fig. 7.11).



Figure˜7.11: Intervalle de confiance à 95% (figuré en gris) calculé à partir de la variance de jackknife, et variogramme local (figuré en noir). (a) Dispositif EAS. (b) Dispositif ES. (c) Dispositif STR.

Comme nous connaissons les distributions d'échantillonnage — approximées par 10^4 réplications de chaque type de dispositif — il est également possible de comparer les écarts-types d'échantillonnage à leur estimation dans le cadre du jackknife : $s_{\hat{\gamma}(h)}$ surestime $\sigma_{\hat{\gamma}(h)}$ dans tous les cas (Tab. 7.5). Cette surestimation s'avère plus élevée pour l'ES que pour le STR, et plus élevée pour le STR que pour l'EAS. Dans le cas de l'EAS, l'estimation obtenue par le jackknife apparaît assez proche de $\sigma_{\hat{\gamma}(h)}$.

Shafer & Varljen (1990) évaluent la validité de leur approche grâce à une étude de Monte-Carlo et traitent explicitement de la variance d'estimation et non pas de la variance de fluctuation. En effet, ces auteurs considèrent avec raison qu'il n'est pas possible de comparer la variance obtenue par jackknife (7.38) à celle obtenue en simulant des réalisations non conditionnelles. Ils estiment le variogramme pour 100 échantillons aléatoires tirés à partir d'une grille dense 100×100, et comparent le variogramme moyen avec l'estimation obtenue en calculant (7.36). Ainsi, les auteurs se placent exclusivement dans le

cadre de l'EAS et concluent que la variance calculée selon l'expression (7.38) constitue une estimation adéquate de la vraie variance mais qu'elle tend à être conservative au sens où elle la surestime un peu.

	EAS		ES		STR	
Classe	$\sigma_{\widehat{\gamma}(h)}$	$s_{\widehat{\gamma}(h)}$	$\sigma_{\widehat{\gamma}(h)}$	$s_{\widehat{\gamma}(h)}$	$\sigma_{\widehat{\gamma}(h)}$	$s_{\widehat{\gamma}(h)}$
1	1.11	1.45	0.17	1.08	0.59	1.24
2	1.99	2.36	0.31	1.94	0.94	1.95
3	2.63	2.89	0.41	2.42	1.30	2.48
4	3.29	3.73	0.58	3.36	1.63	3.20
5	4.90	6.08	0.76	4.56	2.16	4.52
6	5.58	7.06	0.98	5.78	2.43	5.35
7	5.40	5.74	1.22	4.82	2.56	5.36

Tableau 7.5: Comparaison des écarts-types d'échantillonnage $\sigma_{\widehat{\gamma}(h)}$ obtenus dans le cas de l'ES, de l'EAS, et du STR, avec les écarts-types de jackknife $s_{\widehat{\gamma}(h)}$, pour 7 classes de distances (détails dans le texte).

Notre étude de cas et celle de Shafer & Varljen (1990) semblent montrer que la variance (7.38) constitue une bonne estimation de la variance d'échantillonnage dans le cas du dispositif aléatoire simple. L'importante surestimation de la variance d'échantillonnage, notamment dans le cas de l'échantillonnage systématique, implique une grande robustesse des limites de l'intervalle de confiance, au sens de leur validité. Cette robustesse s'accompagne nécessairement d'une perte de précision, comme le montre le cas des échantillons systématiques s_5 et s_8 utilisés pour illustrer l'inférence *model-based* (Fig. 7.12).

Enfin, en ce qui concerne l'ajustement automatique d'un modèle de variogramme à $\widehat{\gamma}(\cdot)$, il est possible d'envisager une pondération en $\left[s^2_{\widehat{\gamma}(h)}\right]^{-1}$ au lieu de celle en $N(h)$. L'intérêt pratique de cette pondération n'a cependant pas été évalué.

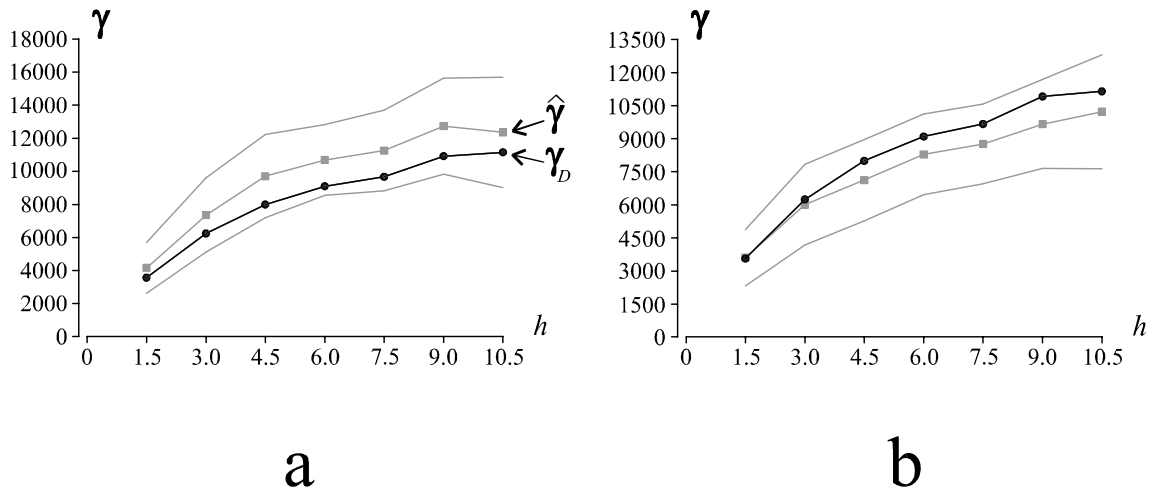


Figure 7.12: Enveloppe de confiance à 95 % (figurée en gris) du variogramme expérimental (carrés figurés en gris) obtenue par la méthode du jackknife et variogramme local (figuré en noir), pour les mêmes classes de distances que le variogramme expérimental. (a) Cas de la plus forte surestimation. (b) Cas de la plus forte sous-estimation.

Chapitre 8

Optimisation de l'échantillonnage

> *"Prior to any sampling design, one should establish the objective of the study in order to deal with the question of data collection"* (Rouhani 1985)
>
> *"Typically, the computation of an actual optimal design will require numerical optimization, and evaluation of any criterion functions for a specific design will often involve numerical integration or optimization, so obtaining optimal designs is by no means simple"* (Cox *et al.* 1997)

Tout échantillon devrait être prélevé en fonction d'objectifs clairement définis et la pratique de l'échantillonnage "pour voir" devrait être évitée. Dans cette optique, il est souhaitable que le motif d'échantillonnage à implémenter soit le mieux adapté possible aux objectifs envisagés et aux moyens disponibles. Ceci peut s'effectuer en exploitant une connaissance *a priori* :

- de la variable régionalisée étudiée (*e.g.*, forme de sa structure d'autocorrélation spatiale),
- des objectifs de l'étude (*e.g.*, nature du paramètre à estimer),
- des contraintes pratiques (*e.g.*, effort d'échantillonnage, coût des mesures, accessibilité sur le terrain).

Cette connaissance *a priori* peut être formalisée sous la forme d'une fonction-objectif $J(\cdot)$ qui mesure la qualité d'un échantillon s, ou sous la forme d'un ensemble de règles $\mathcal{R}$ à appliquer. Le problème de l'optimisation de l'échantillonnage consiste à identifier un échantillon optimal s^* obtenu par minimisation de $J(s)$ ou par un système à base de connaissances manipulant les règles de $\mathcal{R}$. Nous ne traitons dans ce chapitre que de l'optimisation par minimisation d'une fonction-objectif $J(\cdot)$ parce qu'il s'agit de l'approche la plus simple et la plus répandue.

L'optimisation de l'échantillonnage peut se poser, notamment, en termes de taille d'échantillon et de localisation des supports (*e.g.*, McBratney & Webster 1983a, Webster & Burgess 1984, Pardo-Igúzquiza 1998b). Le problème de la taille de l'échantillon nécessaire afin d'atteindre une précision d'estimation fixée *a priori* (*e.g.*, McBratney & Webster 1983a, Di *et al.* 1989, Atkinson 1996) nécessite cependant de définir dans quel cadre inférentiel on se place, *i.e. design-based* ou *model-based* (Brus & de Gruijter 1997).

Sans discuter pour le moment de la pertinence du critère adopté pour mesurer la qualité de l'échantillon qui est adopté, le choix de la taille d'échantillon optimale peut s'effectuer séparément de l'optimisation de la localisation des supports. Il suffit en effet de répéter l'optimisation de la localisation des supports pour différentes tailles d'échantillons, puis d'identifier la solution qui satisfait à la valeur limite requise pour $J(s)$. En conséquence, dans ce qui suit nous examinons uniquement l'optimisation du motif d'échantillonnage pour une taille d'échantillon fixée à n.

Le problème pratique à résoudre consiste à identifier un élément s^* de l'ensemble des échantillons $S^* \subset S$ optimaux vis-à-vis d'une fonction-objectif $J : S \rightarrow \mathbb{R}^+$, avec S l'ensemble de tous les échantillons possibles de taille n. Formellement, un échantillon optimal s^* est tel que :

$$J(s^*) = \min_{s \in S} J(s) \tag{8.1}$$

L'ensemble S est fini ou infini selon que la population $\mathcal{U}$ est elle-même finie ou infinie. La théorie classique de l'optimisation continue traite le cas où S est infini tandis que le cas fini est l'objet de l'optimisation combinatoire (Grötschel & Lovász 1995).

Il est raisonnable de considérer le cas d'une population finie consituée de N unités localisées dans D, parmi lesquelles chaque ensemble distinct de n unités constitue un échantillon s. L'algorithme naïf consiste à évaluer les $\binom{N}{n}$ éléments de S, ce qui n'est envisageable que pour des problèmes de petite taille (*e.g.*, $\binom{25}{9} \simeq 2.043 \times 10^6$). Dans les problèmes de taille plus importante (*e.g.*, $\binom{900}{100} \simeq 9.384 \times 10^{134}$), l'explosion combinatoire rend l'utilisation de l'algorithme naïf impossible (Goldberg 1994, Grötschel & Lovász 1995). Le nombre de solutions à envisager peut être réduit si certaines unités de $\mathcal{U}$ sont fixées *a priori*, par exemple lorsqu'elles sont déjà équipées en pièges, sondes, matériel de mesure, etc. Il convient donc de distinguer trois situations selon que :

1. aucune des n unités à échantillonner dans $\mathcal{U}$ n'est fixée *a priori* : $\binom{N}{n}$ possibilités,
2. il existe un ensemble M de m unités déjà équipées auquel il faut ajouter $n - m$ nouvelles unités choisies dans $\mathcal{U} - M$: $\binom{N-m}{n-m}$ possibilités,
3. il existe un ensemble M de m unités déjà équipées auquel il faut enlever $m - n$ unités choisies dans M : $\binom{m}{m-n}$ possibilités.

La situation (2) nécessite de tenir compte de l'ensemble des unités obligatoirement incluses dans s. Cette situation correspond en fait au problème le plus général dont la situation (1) ne constitue qu'un cas particulier, *i.e.* lorsque $m = 0$. Par ailleurs, dans la situation (3), pour se ramener à un cas formellement équivalent au (1), il suffit de considérer que les unités déjà équipées constituent la population et que les unités à supprimer constituent l'échantillon . En conséquence, pour traiter les trois situations envisagées, il suffit de savoir traiter le cas (2).

Même lorsque certaines unités de s sont fixées, la solution exacte du problème d'optimisation combinatoire considéré ici ne peut pas être obtenue en temps polynomial (en fonction de la taille du problème) sur une machine déterministe, *i.e.* sur un ordinateur existant. En théorie, le problème d'optimisation combinatoire peut être résolu en temps polynomial sur une machine non déterministe (problème *NP* ou *polynomial non déterministe*), autrement dit, une machine qui, confrontée à plusieurs possibilités, aurait la capacité de "deviner" quel est le bon choix (Sedgewick 1991, p. 657, Aho & Ullman 1993, p.

733). En général, les problèmes d'optimisation combinatoire figurent parmi les problèmes les plus difficiles de la classe *NP* et sont dits *NP-complets*[1] (Sakarovitch 1984, Grötschel & Lovász 1995). En conséquence, nous nous contentons de chercher une "bonne" solution (*i.e.*, proche de l'optimun absolu) en un temps raisonnable.

La recherche d'une solution approximative dans le cas d'un problème complexe s'effectue en utilisant une *heuristique* $\mathcal{H}$ définie comme un critère, un principe, ou une méthode qui utilise des informations facilement observables, tout en étant vaguement applicable, et qui permet de contrôler un processus de résolution de problème (Pearl 1990).

Dans la situation générale (2), l'optimisation combinatoire de l'échantillonnage nécessite donc essentiellement de choisir :

- une fonction-objectif $J(\cdot)$,
- une heuristique $\mathcal{H}$.

Nous traitons de façon relativement approfondie l'optimisation combinatoire de l'échantillonnage en vue de l'estimation de deux paramètres populationnels :

- la moyenne globale z_D,
- le variogramme local γ_D.

Pour ces deux paramètres, nous choisissons une fonction-objectif raisonnable et utilisons plusieurs heuristiques. L'optimisation de l'échantillonnage en vue de la cartographie par krigeage sera également abordée, mais uniquement en se référant à la littérature. Les questions concernant l'optimisation de l'échantillonnage pour le partitionnement du domaine D ne seront que très succinctement évoquées (*cf.* Aspie & Barnes 1990, Christakos & Killam 1993).

8.1 Fonctions-objectif

Il est possible de concevoir de nombreuses fonctions-objectif mais les fonctions les plus couramment retenues peuvent se classer essentiellement en quatre catégories, selon l'objectif de l'échantillonnage :

- estimation globale (moyenne arithmétique ou krigeage par bloc),
- estimation locale (cartographie par krigeage ponctuel ou par blocs),
- estimation du variogramme,
- partitionnement d'un domaine.

[1] Des compléments sur les classes de complexité algorithmique peuvent être trouvés dans Sakarovitch (1984, pp. 29-48), Tournassoud (1988, pp. 51-54), Sedgewick (1991, pp. 656-662), Aho & Ullman (1993, pp. 733-734), Prins (1994, pp. 32-37) et Rozenberg & Salomaa (1995).

8.1.1 Estimation globale

Dans le cas de l'estimation globale, que l'optimisation soit vue comme un problème continu (*e.g.*, Hughes & Lettenmaier 1981) ou discret (*e.g.*, Sacks & Schiller 1988), la fonction-objectif est généralement la variance d'erreur d'estimation (Sacks & Schiller 1988, Christakos 1992) :

$$J_{\bar{Z}}(s) = \mathrm{Var}_{\xi}\left[Z_D^* - Z_D\right] = \sigma_E^2 \tag{8.2}$$

avec Z_D la variable aléatoire modélisant la moyenne de population z_D :

$$z_D = \frac{1}{N}\sum_{u\in\mathcal{U}} z\left(u\right) \tag{8.3}$$

et Z_D^* la variable aléatoire modélisant une estimation z_D^* de z_D, *e.g.* la moyenne d'échantillon :

$$z_D^* = \frac{1}{n}\sum_{u\in s} z\left(u\right) \tag{8.4}$$

Dans le cas d'une VR isotrope, la minimisation de la fonction-objectif (8.2) tend à répartir régulièrement les supports d'échantillonnage dans le domaine d'étude D, en fonction de la forme du variogramme décrivant la structure d'autocorrélation spatiale.

Afin de répartir les supports d'échantillonnage de façon régulière dans D, van Groenigen & Stein (1998) n'utilisent pas la variance σ_E^2 mais préfèrent un critère purement géométrique de la forme :

$$J_d(s) = \mathrm{E}\left[d\left(x, s\right)\right] \tag{8.5}$$

avec $d\left(x, s\right)$ la distance euclidienne entre le support x et le plus proche voisin $x_i \in s$. L'optimisation de l'échantillonnage par minimisation du critère (8.5) suppose implicitement que la VR est isotrope, mais aucune information concernant sa structure d'autocorrélation n'est exploitée. Bien que l'objectif de l'étude ne soit pas précisé par van Groenigen & Stein (1998), nous pouvons légitimement supposer qu'il s'agit de l'estimation globale.

Domburg *et al.* (1997) proposent d'optimiser des dispositifs d'échantillonnage (au sens de l'échantillonnage probabiliste) soit dans un cadre doublement stochastique en considérant $\mathrm{E}_{\xi}\left[\mathrm{Var}_p\left[Z_D^*\right]\right]$, soit dans le cadre classique des fonctions de coût de la théorie de l'échantillonnage (Cochran 1977). Nous avons fait le choix de ne pas considérer l'optimisation des *dispositifs d'échantillonnage* mais plutôt l'optimisation des *motifs d'échantillonnage*.

8.1.2 Estimation locale

L'optimisation de l'échantillonnage pour l'estimation locale peut s'effectuer par optimisation de différentes fonctions-objectif, notamment :

- la minimisation de la variance de krigeage maximale (Sacks & Schiller 1988, Christakos & Olea 1992) :

$$J_{\max}(s) = \max_{u\in\mathcal{U}} \sigma_k^2\left(u \mid s\right) \tag{8.6}$$

- la minimisation de la somme (ou ce qui revient au même, la moyenne) des variances de krigeage locales (Christakos & Olea 1992, van Groenigen[2] *et al.* 1999) :

$$J_{\Sigma}(s) = \sum_{u \in \mathcal{U}} \sigma_k^2 (u \mid s) \tag{8.7}$$

- la minimisation du déterminant de la matrice de covariance[3] des prédictions locales (Sacks & Schiller 1988) :

$$J_{\det}(s) = \det G_s \tag{8.8}$$

avec G_s une matrice $(N-n) \times (N-n)$ et $G(t,u)$ la covariance des prédictions pour les supports $t, u \in \mathcal{U} - s$.

Dans le cadre de l'optimisation des dispositifs expérimentaux, le critère (8.6) correspond à l'*E-optimalité*, le critère (8.7) à l'*A-optimalité*, et le critère (8.8) à la *D-optimalité* (*cf.* Atkinson 1982, Nguyen & Miller 1992). En pratique, les résultats obtenus avec les critères (8.6) et (8.8) s'avèrent très différents (Sacks & Schiller 1988), et il en est vraisemblablement de même pour le critère (8.7).

8.1.3 Estimation du variogramme

Dans le cadre de l'optimisation de l'échantillonnage pour l'estimation du variogramme, une fonction-objectif encore récemment considérée (*e.g.*, van Groenigen & Stein 1998) est construite à partir de celle proposée par Warrick & Myers (1987). Le principe de ce type de fonction-objectif est détaillé et discuté dans la Section 8.5.1.

8.1.4 Partitionnement d'un domaine

Aspie & Barnes (1990) examinent le problème du partitionnement d'un domaine en zones de valeurs faibles et élevées, et considèrent la minimisation du coût (financier) de l'erreur de classification. A la différence de la variance d'estimation globale σ_E^2, ce critère incorpore la localisation des supports en même temps que les valeurs mesurées. Au contraire de la minimisation de σ_E^2 qui tend à combler les zones peu échantillonnées, le critère du coût de l'erreur de classification proposé par Aspie & Barnes (1990) tend à placer les nouveaux supports d'échantillonnage à l'interface entre les zones de valeurs faibles et élevées. Bien que les concepts soient développés dans le domaine minier, ils peuvent être appliqués à l'identification de zones hautement contaminées ou à la cartographie d'écotones (Aspie & Barnes 1990).

[2] van Groenigen *et al.* (1999) se placent dans le cadre de l'optimisation continue, et la fonction objectif est par conséquent une intégrale d'espace plutôt qu'une somme discrète.

[3] Le déterminant d'une matrice de covariance est connu sous le nom de *variance généralisée* (Kocherlakota & Kocherlakota 1983).

8.2 Heuristiques

Le recours à une heuristique pour identifier un échantillon $s^* \subset \mathcal{U}$ optimal au sens de $J(\cdot)$ revient à explorer une partie de l'espace des solutions. Dans le cadre de l'optimisation combinatoire, l'espace des solutions est discret et fini, et peut être modélisé par un graphe valué $G = (S, A, J)$, avec S l'ensemble des échantillons possibles (les sommets du graphe), A un ensemble d'arêtes, et $J(\cdot)$ la fonction de valuation des sommets, autrement dit, la fonction-objectif. Une arête dans G entre deux échantillons s_i et s_j représente une modification élémentaire permettant de passer d'un échantillon à l'autre, *i.e.* par changement d'une seule unité d'échantillonnage u. Ainsi, le nombre d'arêtes du chemin le plus court dans G entre deux sommets représente le nombre d'unités qu'il est nécessaire de changer pour passer d'un échantillon à un autre. Il est plus classique de visualiser l'espace des solutions comme un paysage muni de pentes, de cuvettes et de sommets, mais en toute rigueur, cette représentation n'est valide que dans le cas de l'optimisation continue.

Pour trouver une solution optimale s^* en temps polynomial, il faut exploiter la structure de l'espace des solutions (Grötschel & Lovász 1995). L'efficacité d'une heuristique — *i.e.* sa capacité à donner le plus souvent possible une solution la plus proche possible de la solution optimale — dépend donc de sa capacité à tenir compte de la structure de G.

L'heuristique la plus ancienne et la plus largement répandue est celle qui consiste à échantillonner au hasard un ensemble (Pearl 1990). Par définition, cette approche ne tient pas compte de la structure de cet ensemble. Dans le cas de l'optimisation combinatoire, l'inefficacité de cette heuristique se comprend aisément puisqu'il s'agit de rechercher au hasard, parmi un nombre fini mais très élevé de solutions possibles, une solution rare voire même unique. En d'autres termes, en construisant de nombreuses solutions au hasard, il est relativement facile d'identifier la solution moyenne $\bar{s}$ telle que :

$$J(\bar{s}) = \frac{1}{\mathrm{Card}(S)} \sum_{s \in S} J(s) \tag{8.9}$$

mais beaucoup plus difficile de connaître les queues de la distribution des $J(s)$, ce qui ne s'accorde pas avec l'objectif même de l'optimisation (minimisation ou maximisation). Nous considérons donc dans ce qui suit d'autres heuristiques :

- algorithme glouton,
- échange séquentiel,
- optimisation locale,
- recuit simulé,
- recuit simulé modifié,
- recherche taboue,
- algorithmes génétiques.

L'algorithme glouton construit la solution pas à pas tandis que toutes les autres méthodes améliorent de façon itérative une solution initiale. A la différence de l'échange séquentiel et de l'optimisation locale, le recuit simulé ou SA (*Simulated Annealing*), la

recherche taboue ou TS (*Tabu Search*) et les algorithmes génétiques ou GA (*Genetic Algorithms*), sont des méthodes qui ne restent pas bloquées par un optimum local au cours de l'exploration de l'espace des solutions. Ces méthodes, parfois qualifiées de *métaheuristiques* (Glover *et al.* 1993, Prins 1994), ont été identifiées par le CONDOR (*Committee on the Next Decade of Operations Research*) comme extrêmement prometteuses pour le traitement futur des problèmes d'optimisation (Glover *et al.* 1993).

Il est également envisageable d'hybrider plusieurs heuristiques afin de profiter de leurs avantages respectifs, par exemple le SA, la TS et les GA (Fox 1993, Lin *et al.* 1993, Kido *et al.* 1994, Glover 1994, Glover *et al.* 1995), mais il ne nous a pas semblé nécessaire d'explorer cette voie. Enfin, d'autres heuristiques telles que les GRASP (*Greedy Randomized Adaptative Search Procedures*) sont parfois jugées aussi prometteuses que le SA, la TS ou les GA (Feo & Resende 1995). Néanmoins les GRASP ne semblent pas avoir fait leurs preuves, aussi nous ne les avons pas considérées par la suite.

8.2.1 Algorithme glouton

Le principe de l'algorithme glouton est d'effectuer un certain nombre de choix, chacun étant le meilleur au moment où il est fait. La solution est donc construite pas à pas en choisissant de façon optimale le premier élément, puis le second, etc. (Grötschel & Lovász 1995). Le plus souvent, la séquence de décisions locales optimales ne conduit pas à une solution globalement optimale (Aho & Ullman 1993). L'algorithme glouton s'écrit :

1. A l'étape $k = 1$, choisir la meilleure unité u_1 parmi les N possibles. Faire $s = u_1$.
2. A l'étape $k + 1$, choisir une unité u_{k+1} parmi les $N - k$ unités de $\mathcal{U} - s$ telle que u_{k+1} donne la meilleure combinaison avec les unités de s. Faire $s \leftarrow s \cup u_{k+1}$.
3. Faire $k \leftarrow k + 1$ et retourner en 2 jusqu'à ce que $\mathrm{Card}(s) = n$.

Cet algorithme très simple n'est pas optimal mais s'avère utile pour donner très rapidement une première solution (Aspie & Barnes 1990, Christakos 1992). Cependant, la séquence de choix étant déterministe, l'algorithme glouton conduit à une solution unique s_0^+ et ne permet donc pas d'explorer l'espace des solutions. Nous préférons utiliser une version légèrement modifiée dans laquelle la première unité de l'échantillon n'est pas choisie de façon optimale. Deux versions sont envisageables selon que la première unité de l'échantillon est obtenue :

- par examen successif de toutes les unités de $\mathcal{U}$,
- par échantillonnage aléatoire simple de $\mathcal{U}$.

L'itération de l'algorithme avec différentes unités de départ permet d'explorer rapidement l'espace des solutions. La meilleure solution obtenue au cours de ces itérations est généralement meilleure que s_0^+. L'itération de l'algorithme glouton à départs multiples revient à examiner différents sommets de G (au maximum N sommets), mais contrairement à l'heuristique de l'échantillonnage aléatoire, les sommets ne sont pas examinés selon une distribution de probabilités uniforme.

8.2.2 Echange séquentiel

L'heuristique d'échange séquentiel consiste à améliorer une solution initiale en changeant chaque élément de façon optimale tout en laissant les autres éléments fixés. Cette amélioration est répétée jusqu'à ce que l'algorithme converge, *i.e.* lorsqu'il n'est plus possible d'améliorer sensiblement la solution courante (Aspie & Barnes 1990). En pratique, la convergence est atteinte lorsque :

$$\frac{|J(s_k) - J(s_{k+1})|}{J(s_k)} < \varepsilon \tag{8.10}$$

avec $s_k \prec s_{k+1}$ et $\varepsilon \ll 1$. L'algorithme peut s'écrire :

1. Déterminer un échantillon initial $s = s_0$ de n unités choisies parmi les N possibles.
2. Pour chaque unité $u_j \in s$ $(j = 1, \ldots, n)$ chercher une unité $u^* \in \mathcal{U} - s$ telle que sa substitution à l'unité courante u_j améliore l'échantillon s, et si u^* existe, alors l'échanger avec l'unité courante u_j.
3. Itérer sur l'étape 2 jusqu'à la convergence.

Conformément à la modélisation de l'espace des solutions, chaque échange de l'unité courante u_j avec une unité u^* conduisant à une meilleure solution revient à emprunter une arête dans le graphe G. L'exécution de cette heuristique conduit ainsi à cheminer de façon déterministe dans G jusqu'à une solution, souvent seulement localement optimale. L'échange séquentiel constitue ainsi une méthode d'amélioration itérative déterministe, *i.e.* pour laquelle aucun choix n'est effectué au hasard.

8.2.3 Optimisation locale

La méthode d'amélioration itérative stochastique la plus simple consiste à modifier une solution initiale en changeant une unité au hasard, du moins lorsque ce changement s'avère bénéfique. L'algorithme peut s'écrire :

1. Déterminer un échantillon initial $s = s_0$ de n unités choisies parmi les N possibles.
2. Choisir au hasard une unité $u_j \in s$ et une unité $u \in \mathcal{U} - s$. Si la substitution de u_j par u améliore l'échantillon s alors remplacer u_j par u.
3. Si la condition d'arrêt (à spécifier) est satisfaite alors FIN sinon retourner en 2.

Cet algorithme explore aléatoirement le voisinage $V(s)$ d'un sommet s dans le graphe G et se déplace vers la première solution améliorante. Une autre option consiste à chercher à chaque itération la meilleure solution du voisinage (Prins 1994). Tout comme l'algorithme d'échange séquentiel, l'exploration stochastique peut conduire à une solution seulement localement optimale et y rester bloquée.

La condition d'arrêt de l'algorithme peut être spécifiée de deux façons selon que l'échantillonnage aléatoire de $V(s)$ est réalisé avec ou sans remise. Avec remise, l'algorithme se termine lorsque η tentatives successives se sont avérées infructueuses, tandis que sans remise, l'algorithme s'achève lorsque le voisinage $V(s)$ a été totalement exploré sans aboutir à une meilleure solution que s. A efficacité approximativement égale, l'échantillonnage sans remise évite d'avoir à spécifier le paramètre η et s'avère plus rapide.

Au lieu d'échantillonner aléatoirement $V(s)$, il est également possible de l'examiner de façon exhaustive (Prins 1994). Cependant, cette approche est plus coûteuse en temps de calcul et ne s'avère pas nécessairement beaucoup plus performante. Dans ce qui suit, nous considérons donc deux versions selon que le voisinage $V(s)$ est examiné de façon exhaustive ou par échantillonnage aléatoire simple sans remise.

8.2.4 Recuit simulé

Les heuristiques d'amélioration itérative d'une solution initiale (échange séquentiel et optimisation locale) présentent l'inconvénient de rester bloquées lorsqu'elles rencontrent un optimum local. Afin d'explorer plus complètement l'espace des solutions, il est nécessaire de réaliser plusieurs exécutions à partir de différentes solutions initiales (Kirkpatrick *et al.* 1983, Siarry & Dreyfus 1988, Johnson *et al.* 1989). Pour éviter le blocage à une solution seulement localement optimale, une autre stratégie consiste à autoriser l'augmentation occasionnelle de $J(s)$ afin de poursuivre l'exploration de l'espace des solutions. Ce principe est à la base de la méthode du recuit simulé.

La méthode du recuit simulé (SA) — introduite indépendamment par Kirkpatrick *et al.* (1983) et Černý (1985) — exploite les ressemblances entre un problème d'optimisation combinatoire et la mécanique statistique, en établissant une analogie entre la minimisation de la fonction-objectif et celle de l'énergie d'un système physique. L'heuristique du SA est ainsi issue de la transposition au domaine de l'optimisation combinatoire de la technique du *recuit* utilisée en physique des matériaux. La technique du recuit consiste à abaisser lentement la température d'un matériau par paliers, par opposition à la technique de la *trempe* (Siarry & Dreyfus 1988). La méthode du SA fait donc intervenir un paramètre T analogue à la température, et d'autres paramètres contrôlant le nombre et la durée des paliers de température. L'ensemble des valeurs prises par ces paramètres définit le *schéma de recuit* (Kirkpatrick *et al.* 1983). L'algorithme général du SA peut s'écrire :

1. Déterminer un échantillon initial $s = s_0$ de n unités choisies parmi les N possibles.
2. La température T est fixée à une valeur initiale $T_0 > 0$.
3. Choisir au hasard deux unités $u_j \in s$ et $u \in \mathcal{U} - s$. Construire le voisin s' de s dans G en échangeant u_j et u.
4. Calculer $\Delta J = J(s') - J(s)$.
5. Si $\Delta J \leq 0$ alors accepter la solution s' sinon accepter s' avec la probabilité π.
6. Si l'équilibre est atteint aller en 7 sinon retourner en 3.
7. Abaisser T.
8. Si le système est figé alors FIN sinon retourner en 3.

L'algorithme précédent nécessite d'être explicité en choisissant la forme de la probabilité d'acceptation π, ainsi que le schéma de recuit spécifiant la valeur de la température initiale T_0 et décrivant les étapes 6 à 8. Dans la version originelle du SA (Kirkpatrick *et al.* 1983) la règle d'acceptation utilisée s'écrit (Metropolis *et al.* 1953) :

$$\Pr(s \leftarrow s') = \begin{cases} 1 & \text{si } \Delta J \leq 0 \\ \exp(-\Delta J / k_B T) & \text{sinon} \end{cases} \tag{8.11}$$

avec k_B une constante analogue à la constante de Boltzmann. Généralement la constante k_B est fixée à 1 et disparaît de la règle d'acceptation (Siarry & Dreyfus 1988, Bertsimas & Tsitsiklis 1993).

Bien que la description du SA par une suite de chaînes de Markov permette d'étudier la vitesse de convergence sur un plan théorique (Siarry & Dreyfus 1988), le schéma de recuit est généralement déterminé par essais/erreurs (Kirkpatrick *et al.* 1983). La température initiale T_0 peut être fixée en fonction de $J(s_0)$, avec s_0 la solution initiale à améliorer, ou bien résulter d'une procédure préalable (Siarry & Dreyfus 1988, Johnson *et al.* 1989). La durée des paliers de température est généralement réglée de façon empirique, notamment en fonction de la taille moyenne du voisinage d'une solution dans l'espace des solutions (Johnson *et al.* 1989). En général la température T décroît de façon géométrique selon un taux δ constant bien que le formalisme des chaînes de Markov suggère un taux variable[4] (Siarry & Dreyfus 1988). La méthode du SA offre ainsi une grande liberté quant à la spécification de l'algorithme puisqu'il est possible de choisir différentes règles d'acceptation (*e.g.*, Sacks & Schiller 1988), ainsi qu'une infinité de schémas de recuit.

La règle d'acceptation classique du type Boltzmann proposée par Metropolis *et al.* (1953) (8.11) est largement utilisée dans le SA pour l'optimisation de l'échantillonnage (*e.g.*, Ferri & Piccioni 1992, van Groenigen & Stein 1998, Pardo-Igúzquiza 1998b, van Groenigen *et al.* 1999). Il existe des raisons mathématiques pour utiliser la règle (8.11), mais elles concernent le comportement asymptotique de la méthode (Johnson *et al.* 1989). En outre, l'exécution du SA est particulièrement longue avec la technique classique du type Boltzmann (Ingber 1993). En pratique, il est donc légitime de considérer d'autres règles que (8.11), du moment qu'elles donnent de bons résultats tout en étant plus rapides (Johnson *et al.* 1989).

Dans le contexte de l'optimisation de l'échantillonnage, nous avons constaté que la règle du type Boltzmann ne présente aucun avantage par rapport à une règle plus simple. En conséquence, nous avons choisi de manipuler directement la probabilité d'acceptation π sans faire intervenir la température et (8.11), ce qui donne la règle d'acceptation :

$$\Pr(s \leftarrow s') = \begin{cases} 1 & \text{si } \Delta J \leq 0 \\ \pi & \text{sinon} \end{cases} \tag{8.12}$$

En outre, nous avons retenu :

- une durée de palier fixe limitée à un nombre d'itération η_0,
- la décroissance géométrique de la probabilité π selon un taux δ constant,
- l'arrêt de l'algorithme lorsque π devient inférieur ou égal à une valeur minimale $\pi_{\min}$ ou bien lorsqu'aucune diminution de $J(s)$ n'a été observée au cours du dernier palier de température.

Le schéma de recuit conditionne le temps d'exécution de l'algorithme et la qualité de la solution obtenue (Siarry & Dreyfus 1988, Bertsimas & Tsitsiklis 1993). Contrairement à ce qu'écrivent certains auteurs (*e.g.*, Pardo-Igúzquiza 1998b), l'algorithme du SA ne converge pas toujours. En effet, il peut arriver qu'un schéma de recuit médiocre ne permette pas d'aboutir à une solution finale meilleure que toutes les solutions précédentes. Il est donc judicieux de retenir comme solution de l'heuristique s^+ la meilleure solution $\min J(s)$ obtenue au cours du déroulement de l'algorithme (Prins 1994).

[4] Une méthode d'ajustement adaptatif de la température est proposée par Ferri & Piccioni (1992).

8.2.5 Recuit simulé modifié

L'algorithme général du recuit simulé peut être modifié afin de mieux exploiter la structure du problème d'optimisation combinatoire traité. Dans le cadre de l'optimisation de l'échantillonnage, Sacks & Schiller (1988) proposent un algorithme de SA modifié. En notant π_k et s_k respectivement la probabilité d'acceptation et l'échantillon solution à l'étape k, l'algorithme de SA modifié peut s'écrire (Sacks & Schiller 1988, Christakos 1992) :

1. Déterminer un échantillon initial s_0 de n unités choisies parmi les N possibles.
2. La probabilité d'acceptation est fixée à une valeur initiale π_0.
3. A l'étape $k+1$, choisir au hasard une unité $u' \in \mathcal{U} - s_k$ et une autre unité $u^* \in s_k$ qui satisfait :
$$J\left(s \cup u' - u^*\right) = \min_{u \in s_k} J\left(s \cup u' - u\right)$$
4. Soit $s' = s \cup u' - u^*$. Calculer $\Delta J = J(s') - J(s_k)$.
5. Si $\Delta J \leq 0$ alors prendre $s_{k+1} = s'$ sinon prendre $s_{k+1} = s'$ avec la probabilité π et $s_{k+1} = s_k$ avec la probabilité $1 - \pi$.
6. Si s' n'est pas accepté comme nouvelle solution (*i.e.,* $s_{k+1} = s_k$), prendre une autre unité $u'' \in \mathcal{U} - s_k - u'$ et retourner en 3 en utilisant u'' au lieu de u'.
7. Répéter l'étape 6 si nécessaire η_0 fois ($\eta_0 \leq N - n$). Si après η_0 tentatives aucune modification n'a été acceptée, alors prendre $s_{k+1} = s_k$, modifier π (à spécifier) et retourner en 3 sinon faire :
$$\pi_{k+1} = \begin{cases} \delta\pi_k & \text{si } J(s_{k+1}) \leq \alpha \min_{i \leq k} J(s_i) \\ \pi_k & \text{sinon} \end{cases} \tag{8.13}$$
8. Si la condition d'arrêt (à spécifier) est satisfaite alors FIN sinon retourner en 3.

Afin de spécifier complètement l'algorithme, il reste à préciser le mode de modification de π à l'étape 7 et la condition d'arrêt de l'étape 8. Afin de ne pas trop augmenter le nombre de paramètres à régler, nous avons choisi de modifier π à l'étape 7 simplement comme $\pi_{k+1} = \delta\pi_k$. Pour établir un compromis entre le temps de calcul et la qualité de la solution, nous avons choisi d'arrêter l'algorithme lorsque $\pi_{k+1} \leq \pi_{\min}$ ou lorsque η changements successifs n'ont pas conduit à améliorer la meilleure solution rencontrée $\min_{i \leq k} J(s_i)$.

L'algorithme de SA modifié apparaît beaucoup plus directif que dans sa version originelle, ce qui se traduit par un nombre plus élevé de paramètres et une structure plus compliquée.

8.2.6 Recherche taboue

Introduite plus récemment que le recuit simulé (Glover 1989, 1990a), la recherche taboue (TS) a également pour but d'éviter les optima locaux et de guider des heuristiques plus spécialisées (Glover 1990b). Comme le recuit simulé, la TS est parfois qualifiée de *méta-stratégie* (Glover 1989) ou de *métaheuristique* (Glover *et al.* 1993). La TS peut être vue

comme une technique basée sur des concepts issus de l'Intelligence Artificielle (Glover *et al.* 1993). A notre connaissance, la TS n'a pas encore été utilisée pour l'optimisation de l'échantillonnage.

Dans sa version la plus simple, la TS recherche la meilleure solution s^+ : (i) dans le voisinage $V(s)$ d'un sommet courant s, (ii) dans un échantillon de $V(s)$ si l'examen exhaustif du voisinage est trop coûteux (Glover *et al.* 1993). Le processus de recherche passe ensuite à s^+ même si cette solution est moins bonne que s. En autorisant d'emprunter des sommets qui n'améliorent pas la solution courante, la TS évite que la recherche reste bloquée par une solution qui serait seulement localement optimale, et constitue donc une amélioration évidente de l'algorithme d'optimisation locale (Glover *et al.* 1993).

Cependant, cette stratégie peut conduire à des cycles. Afin d'éviter de boucler indéfiniment sur les mêmes sommets, l'algorithme peut utiliser une liste $\mathcal{L}$ FIFO (*First-In First-Out*) de sommets tabous, *i.e.* déjà utilisés dans l'exploration de G. Cette liste correspond en quelque sorte à la mémoire de l'algorithme, et sa taille maximale ω constitue un paramètre clef de la TS. La solution de l'heuristique est la meilleure solution $\min J(s)$ rencontrée au cours de l'exploration de G (Glover *et al.* 1993, Prins 1994).

En pratique, il s'avère préférable d'explorer uniquement un échantillon aléatoire de $V(s)$. En effet, l'examen exhaustif de $V(s)$ est souvent trop coûteux en temps de calcul et ne s'avère pas indispensable pour obtenir une bonne solution. En considérant un échantillon aléatoire avec remise de $V(s)$, fixé à une taille $\eta_0 < \operatorname{Card} V(s)$, l'algorithme de la TS peut s'écrire :

1. Déterminer un échantillon initial $s = s_0$ de n unités choisies parmi les N possibles.
2. Initialiser la liste taboue $\mathcal{L} = s$.
3. Initialiser $J_{\min} = \infty$.
4. Choisir au hasard deux unités $u_j \in s$ et $u \in \mathcal{U} - s$. Construire le voisin s' de s dans G en échangeant u_j et u.
5. Si $s' \notin \mathcal{L}$ alors calculer $J(s')$ sinon aller en (6). Si $J(s') < J_{\min}$ alors $J_{\min} = J(s')$ et $s'' = s'$.
6. Retourner en 4 jusqu'à ce que η_0 voisins aient été examinés.
7. Faire $s = s''$ et $\mathcal{L} \leftarrow \mathcal{L} \cup s$.
8. Si la condition d'arrêt (à spécifier) est satisfaite alors FIN sinon retourner en 3.

Afin de spécifier complètement l'algorithme, il reste à préciser la condition d'arrêt de l'étape 8. Pour établir un compromis entre le temps de calcul et la qualité de la solution, nous avons choisi d'arrêter l'algorithme lorsque η itérations successives n'ont pas conduit à améliorer la meilleure solution rencontrée $\min J(s)$.

8.2.7 Algorithmes génétiques

Les algorithmes génétiques (GA) dérivent d'une théorie des systèmes adaptatifs inspirée de l'adaptation des populations d'organismes à leur environnement (Holland 1962). Les GA constituent actuellement une classe d'heuristiques qui permettent de traiter des problèmes d'optimisation continue discrétisés (*e.g.*, Riolo 1992) ou d'optimisation combinatoire proprement dite (*e.g.*, Holland 1992). A notre connaissance, les GA n'ont pas été évalués

dans le cadre de l'optimisation de l'échantillonnage bien qu'ils soient mentionnés parmi les algorithmes intéressants dans ce contexte (Christakos 1992).

Par analogie avec les populations d'organismes, les GA considèrent une population de solutions de taille q. L'heuristique des GA évalue plusieurs solutions à la fois et se différencie donc immédiatement des précédentes par son *parallélisme implicite* (Holland 1992, Goldberg 1994). Par analogie avec un chromosome unique représentant le génome d'un organisme (*i.e.*, une bactérie), chaque solution est codée sous la forme d'une chaîne d'éléments, le plus souvent une chaîne de bits. Toujours par analogie avec l'évolution d'une population d'organismes, les itérations des GA sont appelées *générations*. A chaque génération, des chaînes de la génération précédente sont sélectionnées selon leur *fitness*, *i.e.* selon la valeur de $J(s)$, puis elles subissent des mutations (*e.g.*, translocation réciproque). Les mutations sont définies par analogie avec la génétique (*e.g.*, Goldberg 1994), et la sélection peut éventuellement faire intervenir un mécanisme de *compétition* entre solutions (*e.g.*, Riolo 1992).

Au fil des générations, les mutations permettent d'explorer l'espace des solution, la sélection améliorant progressivement la population toute entière. L'intérêt des GA est que les bonnes solutions sont encouragées à échanger leurs caractéristiques par translocation réciproque et à engendrer des solutions encore meilleures (Prins 1994). La mutation par translocation joue donc un rôle fondamental dans cette classe d'heuristiques. La forme générale d'un GA peut donc s'écrire :

1. Former la génération $k = 0$ au hasard.
2. Former la génération $k + 1$ en sélectionnant selon leur *fitness* q chaînes parmi celles de la génération k.
3. Recombiner les chaînes par translocation réciproque.
4. Réaliser d'autres mutations (*e.g.*, ponctuelles, par inversion, par transposition).
5. Si η générations ont été construites alors FIN sinon retourner en 2.

La spécification d'un GA s'avère ainsi bien plus longue que celle des heuristiques précédentes et nécessite de définir notamment :

- le codage d'une solution sous la forme d'une chaîne,
- le mécanisme de reproduction différentielle (sélection),
- les types de mutations et leurs probabilités d'occurrence à chaque génération.

Nous considérons ici le cas de l'optimisation de l'échantillonnage, l'étude générale des propriétés et de la mise en oeuvre des GA étant exposée dans Goldberg (1994).

8.2.7.1 Codage

Dans le cadre de l'optimisation combinatoire d'un échantillon, le codage s'effectue naturellement sous la forme d'une chaîne de bits $C = (c_i \mid \ i = 1, \ldots, N)$ avec $c_i \in \{0, 1\}$. La position i dans C d'un bit c_i représente la position de l'unité u_i qui lui correspond dans l'ensemble ordonné $\mathcal{U}$. Les bits à 1 correspondent aux éléments de l'échantillon s tandis que les bits à 0 représentent les unités de $\mathcal{U} - s$. Comme la taille de l'échantillon est fixée

à n, le nombre de bits à 1 dans C est une constante, ce qui représente une contrainte peu compatible avec l'esprit des GA. En outre, afin de pouvoir traiter le problème sous sa forme la plus générale — *i.e.* tenant compte d'un ensemble d'unités obligatoirement présentes dans s — il faut concevoir C comme la concaténation d'une *amorce* invariable et d'une partie variable. Les mutations concerneront donc exclusivement la partie variable située après l'amorce.

8.2.7.2 Sélection

La sélection opère à chaque génération en favorisant la reproduction de bonnes solutions aux dépens de solutions moins favorables, tout en essayant de maintenir une certaine diversité au sein de la population. La reproduction différentielle consiste à attribuer aux chaînes une probabilité de contribuer à la génération suivante d'autant plus élevée que les solutions correspondantes sont bonnes (Goldberg 1994). Un premier mécanisme de sélection consiste à évaluer la contribution des chaînes à l'adaptation totale de la population (somme des *fitness* élémentaires), à l'exprimer sous la forme d'une proportion, et à l'utiliser comme une probabilité de sélection (Goldberg 1994). La probabilité de sélection $\Pr(C^{(i)})$ d'une chaîne $C^{(i)}$ dans la population de taille q s'écrit ainsi :

$$\Pr(C^{(i)}) = \frac{f(C^{(i)})}{\sum_{i=1}^{q} f(C^{(i)})} \tag{8.14}$$

et dans un problème de minimisation, la *fitness* peut se calculer selon

$$f(C_i) = \max_{s \in S} J(s) - J(s_i) \tag{8.15}$$

ce qui nécessite toutefois de connaître $\max_{s \in S} J(s)$, au moins de façon approximative. Une seconde approche consiste à itérer q fois une compétition entre deux chaînes choisies au hasard dans la population de la génération précédente. A l'issue de chaque compétition, la meilleure solution est sélectionnée avec une probabilité P_c. Les deux modes de sélection peuvent éventuellement être combinés.

8.2.7.3 Mutation

Conformément à la contrainte énoncée précédemment qui oblige à ne manipuler que des chaînes dont le nombre de bits à 1 est constant et égal à n, les opérateurs de mutation n'offrent pas autant de liberté que dans le cas général. Soit $C_{i \to j}$ une sous-chaîne de C définie comme la séquence connexe de bits $C_{i \to j} = (c_k \mid \ k = i, \ldots, j)$. Nous avons retenu quatre types de mutations :

- l'*inversion* d'une sous-chaîne $C_{i \to j}$ ou, à la limite, de toute la chaîne $C_{1 \to N}$,
- la *transposition par excision*, entre i et j inclus, d'une sous-chaîne $C_{i \to j}$ et son insertion à une nouvelle position dans la chaîne C,
- l'échange de deux sous-chaînes disjointes $C_{i \to j}$ et $C_{k \to l}$ d'une même chaîne C avec $k > j$ ou *substitution*,
- l'échange de deux sous-chaînes $C^{(1)}_{j \to N}$ et $C^{(2)}_{j \to N}$ entre deux chaînes $C^{(1)}$ et $C^{(2)}$ avec $j \leq N$ ou *translocation réciproque*.

Inversion Appliquée selon une probabilité P_i, l'inversion conserve le nombre de bits à 1 dans une chaîne. L'inversion de la sous-chaîne $C_{i\to j}$ consiste à inverser l'ordre des bits qui la compose. La mutation est neutre si $\mathcal{P}(C_{i\to j})$ est vrai, avec $\mathcal{P}$ un prédicat testant si $C_{i\to j}$ est un *palindrome*, *i.e.* si $c_k = c_{j-i+2-k}$ avec $k = i, \ldots, j$ (Arnold & Guessarian 1992). Le cas limite de l'élément neutre est celui de l'inversion d'un bit ($i = j$). Si $j > i$, la mutation est d'autant plus importante que $\Delta = N - \mathrm{Card}(C_{i\to j})$ est faible. Le cas limite $\Delta = 0$ correspond à l'inversion de la chaîne toute entière. L'algorithme de l'inversion s'écrit :

1. Choisir au hasard deux positions i et j telles que $i \neq j$.
2. Retourner en 1 tant que $\mathcal{P}(C_{i\to j})$ est vrai.
3. Inverser l'ordre des bits de $C_{i\to j}$.
4. Mettre à jour s et $J(s)$ pour C.

Transposition Appliquée selon une probabilité P_{tp}, la transposition conserve le nombre de bits à 1 dans une chaîne. La transposition d'une sous-chaîne $F = C_{i\to j}$ consiste à l'exciser de C, à reconcaténer C, puis à insérer F dans C entre deux bits successifs c_k et c_{k+1}. Nous ne considérons pas les cas neutres dégénérés pour lesquels $i = 1$ et $j = N$ ou $k = i - 1$. Soit $\mathcal{H}$ un prédicat vrai si C est une chaîne homogène pour un motif M, *i.e.* si $C = \bigcup_k M$ avec k le nombre de répétitions de M. La transposition est une mutation neutre lorsque $\mathcal{H}(M, C_{i\to j})$ est vrai, que $F = C_{i\to j}$ est issu d'une séquence C_0 telle que $\mathcal{H}(M, C_0)$ est vrai, et que l'insertion de F s'effectue entre deux bits de $C_0 - C_{i\to j}$ en redonnant une séquence homogène pour M. Pour simplifier, nous négligeons les cas où la transposition est neutre. L'algorithme de la transposition s'écrit alors :

1. Choisir au hasard i et j tels que $1 \leq i \leq j \leq N$.
2. Retourner en 1 tant que $(i, j) = (1, N)$.
3. Faire $F = C_{i\to j}$.
4. Exciser : si $i = 1$ alors $C \leftarrow C_{j+1\to N}$ sinon si $j = N$ alors $C \leftarrow C_{1\to i-1}$ sinon $C \leftarrow C_{1\to i-1} \cup C_{j+1\to N}$.
5. Choisir au hasard k tel que $k \in [0, \mathrm{Card}(C)]$.
6. Retourner en 5 tant que $k = i - 1$.
7. Insérer : si $k = 0$ alors $C \leftarrow F \cup C$ sinon si $k = \mathrm{Card}(C)$ alors $C \leftarrow C \cup F$ sinon $C \leftarrow C_{1\to k} \cup F \cup C_{k+1\to \mathrm{Card}(C)}$.
8. Mettre à jour s et $J(s)$ pour C.

Substitution Appliquée selon une probabilité P_s, la substitution conserve le nombre de bits à 1 dans une chaîne. Pour simplifier, nous envisageons uniquement la substitution de deux sous-chaînes $C_{i\to j}$ et $C_{k\to l}$ de mêmes tailles soit $\mathrm{Card}(C_{i\to j}) = \mathrm{Card}(C_{k\to k})$, ce qui équivaut à $j - i = l - k$. Il existe deux cas limites quant à la taille des sous-chaînes. Une mutation mineure est obtenue lorsque $\mathrm{Card}(C_{i\to j}) = 1$ tandis qu'une mutation majeure correspond à $\mathrm{Card}(C_{i\to j}) = \left[\frac{N}{2}\right]$ avec $[x]$ la partie entière de x. La mutation est neutre lorsque $C_{i\to j} = C_{k\to l}$. Nous nous contentons de considérer le cas limite de la substitution de deux bits, ce qui revient à effectuer deux mutations ponctuelles dans C avec la contrainte que le nombre de bits à 1 reste constant.

L'algorithme de la substitution ponctuelle s'écrit :

1. Choisir au hasard deux positions i et j telles que $i \neq j$.
2. Retourner en 1 tant que $c_i = c_j$.
3. Faire $c_i \leftarrow 1 - c_i$ et $c_j \leftarrow 1 - c_j$.
4. Mettre à jour s et $J(s)$ pour C.

Translocation Appliquée selon une probabilité P_{tl}, la translocation ne conserve pas nécessairement le nombre de bits à 1 dans une chaîne. Pour simplifier, nous envisageons uniquement la translocation à un point j qui consiste à échanger les sous-chaînes terminales respectives $C^{(1)}_{j\to N}$ et $C^{(2)}_{j\to N}$ de deux chaînes $C^{(1)}$ et $C^{(2)}$. La mutation est neutre pour $C^{(1)}_{j\to N} = C^{(2)}_{j\to N}$. Compte tenu de la contrainte du nombre de bits à 1 constant, une translocation au point j n'est possible que si $\eta\left(C^{(1)}_{j\to N}\right) = \eta\left(C^{(2)}_{j\to N}\right)$ avec $\eta\left(C_{i\to j}\right)$ le nombre de bits à 1 dans la sous-chaîne $C_{i\to j}$. Avec cette contrainte, l'algorithme de translocation s'écrit :

1. Soit $\mathcal{L}$ l'ensemble des positions $i < N$ telles que $\eta\left(C^{(1)}_{i+1\to N}\right) = \eta\left(C^{(2)}_{i+1\to N}\right)$.
2. Si $\mathcal{L} = \emptyset$ alors FIN.
3. Choisir au hasard une positions $i \in \mathcal{L}$.
4. Faire $C^{(1)} \leftarrow C^{(1)}_{1\to i} \cup C^{(2)}_{i+1\to N}$ et $C^{(2)} \leftarrow C^{(2)}_{1\to i} \cup C^{(1)}_{i+1\to N}$.
5. Mettre à jour s et $J(s)$ pour $C^{(1)}$ et $C^{(2)}$.

8.2.8 Choix de l'heuristique

Le choix d'une heuristique parmi toutes celles que nous avons décrites dépend, par ordre d'importance :

1. du problème traité,
2. du niveau de qualité de la solution recherchée,
3. du temps d'exécution maximal toléré.

En premier lieu, selon le problème traité et par conséquent selon la forme de la fonction-objectif utilisée, toutes les heuristiques ne sont pas utilisables. La nature du problème traité peut donc conduire à une restriction de l'ensemble des heuristiques à envisager.

Il est important que le niveau de qualité des solutions suboptimales obtenues en appliquant une heuristique soit garanti pour toutes les entrées de l'algorithme (Grötschel & Lovász 1995). Le niveau de qualité de la solution recherchée dépend quant à lui des performances relatives des heuristiques considérées. Il n'est pas de notre propos de traiter des performances relatives des heuristiques dans un cadre abstrait (*cf.* Sakarovitch 1984, pp. 236-249, Pearl 1990, pp. 139-169), mais plutôt dans un cadre empirique.

La comparaison empirique des performances des heuristiques déterministes ne pose pas de difficulté particulière parce que leur exécution ne dépend pas d'un ou plusieurs paramètres. En revanche, tous les algorithmes stochastiques — optimisation locale, SA, TS

et GA — font intervenir parfois de nombreux paramètres. Les performances de ces heuristiques sont donc évaluées conditionnellement aux valeurs prises par leurs paramètres. Dans ces conditions, il est difficile de garantir qu'une méthode paramétrée est utilisée au maximum de ses performances. Par exemple, dans le cas de la méthode du SA, un mauvais schéma de recuit peut conduire à des performances très médiocres. Il apparaît alors un second problème d'optimisation qui vient se superposer au premier : l'optimisation du réglage des heuristiques. Toutefois, pour la plupart des heuristiques, de nombreux essais permettent de fixer de façon raisonnable les valeurs des paramètres, de sorte qu'une comparaison des performances reste envisageable.

En outre, pour une structure d'autocorrélation spatiale donnée, et des tailles de population et d'échantillon fixées, la structure de l'espace des solutions dépend évidemment de la fonction-objectif : les performances des heuristiques sont donc également conditionnées par la fonction-objectif.

Enfin, le temps d'exécution peut éventuellement entrer en ligne de compte, mais les performances croissantes des ordinateurs font passer ce critère au dernier plan. De plus, il est souvent difficile de prévoir le temps d'exécution de la majorité des heuristiques mentionnées. Il est cependant évident que l'algorithme glouton est l'heuristique la plus rapide et que les GA demandent un effort de calcul important.

En conséquence, il s'avère impossible d'établir un choix définitif et universel de l'heuristique, *i.e.* quels que soient la fonction objetctif, la taille du problème et le paramétrage des algorithmes. En pratique, l'approche la plus judicieuce consiste à utiliser plusieurs heuristiques, éventuellement en les exécutant plusieurs fois à partir de différentes solutions initiales, et à retenir la meilleure solution parmi celles obtenues.

8.3 Optimisation pour l'estimation globale

L'estimateur linéaire z_D^* le plus simple pour la moyenne globale z_D est la moyenne arithmétique $\bar{z}$ des valeurs de l'échantillon $\{z(s_i) \mid i = 1, \ldots, n\}$. Néanmoins, dans le cadre d'un modèle géostatistique (fonction aléatoire), l'estimateur z_D^* peut être optimisé par krigeage : il s'agit alors d'un krigeage par bloc portant sur le domaine D. L'optimisation de la stratégie d'échantillonnage de la moyenne globale peut donc porter à la fois sur la disposition des supports $\{s_i \mid i = 1, \ldots, n\}$ et sur l'estimateur lui-même, ce qui revient à minimiser la variance de krigeage par bloc σ_K^2 (*e.g.*, Pardo-Igúzquiza 1998b). Comme nous traitons uniquement de l'optimisation de l'échantillonnage, nous considérons un estimateur z_D^* fixé *a priori*, ici la moyenne arithmétique $\bar{z}$.

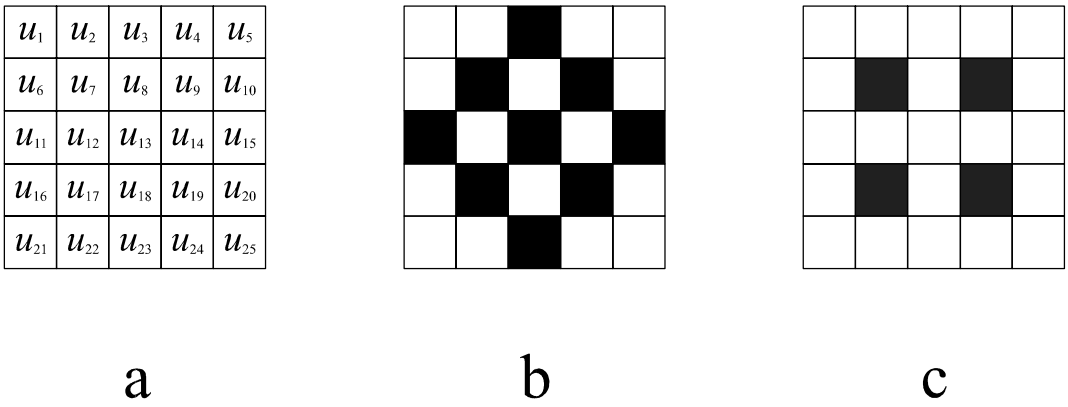
Nous avons choisi d'utiliser le critère $J_{\bar{Z}}$, *i.e.* la minimisation de la variance d'erreur d'estimation σ_E^2 de la moyenne globale z_D par la moyenne d'échantillon $\bar{z}$ dans le cadre d'un modèle de fonction aléatoire.

8.3.1 Performance des heuristiques

Il n'est généralement pas possible de quantifier la suboptimalité d'une solution obtenue pas une heuristique donnée (*e.g.*, le SA modifié), sauf dans le cas d'un exemple de petite taille pour lequel la solution optimale peut être obtenue par énumération exhaustive (Christakos

& Killam 1993). Dans cette optique, et comme une étude théorique des performances relatives de plusieurs heuristiques nous semble hors de portée, nous avons choisi de réaliser une étude empirique en appliquant les heuristiques considérées à un problème de petite taille, pour lequel nous pouvons déterminer de façon exacte la solution optimale s^*.

La population spatiale $\mathcal{U} = \{u_i \mid i = 1, \ldots, 25\}$ est consituée par une grille de 5×5 quadrats de maille $\Delta = 1$ (Fig. 8.1.a). La structure d'autocorrélation de la variable régionalisée $z(\cdot)$ est décrite arbitrairement par le modèle $\mathrm{Expo}(0, 500, 1)$. La taille de l'échantillon est fixée à $n = 9$ afin d'interdir l'existence de plusieurs solutions optimales, équivalentes par rotation d'un même motif au sein de la grille. La distribution exacte des $J(s)$ s'étend entre[5] $J_{\bar{Z}}(s^*) = \min J_{\bar{Z}}(s) = 43.98593$ et $\max J_{\bar{Z}}(s) = 51.54977$, avec une valeur moyenne $\mathrm{moy}\, J_{\bar{Z}}(s) = 46.79718$ (Fig. 8.2.a). La solution optimale correspond à l'échantillon $s^* = \{u_{3,} u_7, u_9, u_{11}, u_{13}, u_{15}, u_{17}, u_{19}, u_{23}\}$ (Fig. 8.1.b).



Figure˜8.1: Exemple de petite taille pour lequel toutes les solutions possibles peuvent être construites. (a) Population spatiale $\mathcal{U} = \{u_i \mid i = 1, \ldots, 25\}$ organisée selon une grille 5×5. (b) Echantillon optimal pour $n = 9$. (c) Quatre unités de départ de l'algorithme glouton modifié conduisant à l'échantillon optimal (détails dans le texte).

Soit s^+ la solution obtenue par une heuristique. Nous distinguons trois ensembles d'heuristiques selon que s^+ est obtenue par un mécanisme :

- de construction pas à pas (algorithme glouton),
- d'amélioration itérative d'une solution initiale (échange séquentiel, optimisation locale, SA, SA modifié, TS),
- de mutation/sélection d'une population de solutions (GA).

8.3.1.1 Algorithme glouton

Considérons l'algorithme glouton modifié, *i.e.* à départs multiples. L'utilisation successive de toutes les unités de $\mathcal{U}$ comme points de départ de l'algorithme conduit à la solution $s^+ = s^*$. Seulement quatre unités de départ conduisent à s^* : u_7, u_9, u_{17} et u_{19}. Ces quatre unités occupent une position particulière au sein de la grille et forment ensemble un motif régulier, centré, qui se révèle être la solution optimale pour le problème $n = 4$ (Fig. 8.1.c).

[5] Le nombre important de chiffres significatifs que nous utilisons n'a pas de valeur dans l'absolu mais nous permet de distinguer précisément les solutions entre elles.

Ainsi, la solution optimale est obtenue par l'examen d'un ensemble de solutions de taille N au lieu de $\binom{N}{n}$, soit un rapport d'échantillonnage de G de l'ordre de 1.22×10^{-5}, pour $N = 25$ et $n = 9$.

8.3.1.2 Méthodes d'amélioration itérative

Les performances des méthodes d'amélioration itérative peuvent être comparées en réalisant N_a applications de chacune des heuristiques, correctement paramétrées. Dans ces conditions, et sur la base de $N_a = 10^3$ exécutions, nous considérons qu'une heuristique est d'autant plus performante qu'elle trouve souvent s^*, et que la moyenne et le maximum des valeurs de $J_{\bar{Z}}(s)$ sont faibles. Toutes les heuristiques d'amélioration itérative modifient un échantillon initial s_0 : nous avons considéré deux modes selon que s_0 est déterminé au hasard ou par l'algorithme glouton à départ aléatoire. Les paramètres retenus pour les heuristiques sont les suivants :

- échange séquentiel

  $\varepsilon = 10^{-7}$

- recuit simulé

  $\pi_0 = 0.8$, $\pi_{\min} = 10^{-5}$, $\delta = 0.9$, $\eta_0 = 1000$

- recuit simulé modifié

  $\pi_0 = 0.59$, $\pi_{\min} = 10^{-4}$, $\delta = 0.8$, $\eta_0 = N - n$, $\alpha = 0.995$, $\eta = 1000$

- recherche taboue

  (i) dans $V(s)$: $\omega = 35$, $\eta = 35$,
  (ii) dans un EAS de $V(s)$, avec remise : $\omega = 35$, $\eta_0 = 50$, $\eta = 35$.

L'heuristique de l'optimisation locale avec examen de $V(s)$, exhaustif (i), ou par EAS sans remise (ii), ne comporte aucun paramètre. Lorsque la solution initiale est générée aléatoirement, le SA apparaît comme l'heuristique la plus performante (Tab. 8.1).

Heuristique	$\rho^{*/+}$	moy $J_{\bar{Z}}(s^+)$	max $J_{\bar{Z}}(s^+)$
Echange séquentiel	33.4	44.09918	44.50899
Optimisation locale (i)	32.3	44.11807	44.50899
Optimisation locale (ii)	29.5	44.13161	44.50899
SA	100.0	43.98593	43.98593
SA modifié	97.2	43.98935	44.10842
TS (i)	48.3	44.05948	44.18252
TS (ii)	50.0	44.04773	44.18252

Tableau 8.1: Performances des heuristiques pour la minimisation de $J_{\bar{Z}}(\cdot)$ évaluées d'après 10^3 exécutions, à partir d'une solution initiale aléatoire. $\rho^{*/+}$: pourcentage de solutions optimales. moy $J_{\bar{Z}}(s^+)$ et max $J_{\bar{Z}}(s^+)$: valeurs moyenne et maximale de $J_{\bar{Z}}(\cdot)$ pour les solutions de l'heuristique.

Le SA modifié est aussi performant que la version originelle mais s'avère par ailleurs beaucoup plus rapide. L'échange séquentiel et les deux versions de l'optimisation locale sont pratiquement équivalentes. La version de la TS qui échantillonne $V(s)$ par un EAS avec remise s'avère équivalente à celle qui explore $V(s)$ de façon exhaustive. Le recours à l'échantillonnage de $V(s)$ ne se traduit donc pas par un effondrement des performances de la TS. Ce résultat est important dans la mesure où, dans un problème de taille réelle, l'examen exhaustif de $V(s)$ représente une étape très coûteuse en temps de calcul. Les valeurs maximales de la fonction-objectif sont basses et correspondent à d'assez bonnes solutions. Deux solutions particulières reviennent plusieurs fois : $J_{\bar{Z}}(s^+) = 44.50899$ et $J_{\bar{Z}}(s^+) = 44.18252$ (Tab. 8.1). Ces solutions doivent correspondre à des optima locaux dont il est particulièrement difficile de s'échapper.

Soient $s_0 = s_{alea}$ une solution initiale aléatoire et $s_0 = s_{glou}$ une solution initiale obtenue par l'algorithme glouton à départ aléatoire. Soit $\rho^{*/+}$ le pourcentage de solutions optimales obtenu, soit formellement $N_0/N_a \times 100$, avec :

$$N_0 = \sum_{i=1}^{N_a} \delta\left(J(s^+), J(s^*)\right) \tag{8.16}$$

et

$$\delta(\alpha, \beta) = \begin{cases} 1 & \text{si } \alpha = \beta \\ 0 & \text{sinon} \end{cases} \tag{8.17}$$

Afin d'apprécier si les performances des heuristiques augmentent en utilisant s_{glou} à la place de s_{alea}, il faut distinguer :

- les heuristiques aux performances moyennes avec s_{alea}, *i.e.* pour lesquelles $\rho^{*/+} \leq 80$,
- les heuristiques déjà très performantes avec s_{alea}, *i.e.* pour lesquelles $\rho^{*/+} > 80$.

En effet, dans le cas du SA, pour apprécier l'impact de la solution initiale, il faut au préalable dégrader leurs performances, par exemple en diminuant de 1000 à 100 le paramètre η_0 du SA et le paramètre η du SA modifié. Il convient ensuite de distinguer les solutions optimales dues à $s_0 = s^*$ de celles dues à l'heuristique $\mathcal{H}$ elle-même. Pour cela, il suffit de faire la différence entre le pourcentage de solutions optimales obtenues dès l'initialisation ($\%\mathcal{H}_0$) et après l'exécution de l'heuristique ($\%\mathcal{H}_T$), afin d'obtenir la performance propre à l'exécution de l'heuristique ($\%\mathcal{H}$). Pour s_{alea}, nous avons toujours obtenu $\%\mathcal{H}_0 = 0$ tandis que pour s_{glou}, $\%\mathcal{H}_0 \simeq 15.8$. Il est possible de fixer la valeur de $\%\mathcal{H}_0$ en déterminant la solution initiale avec une séquence de nombres pseudo-aléatoires (Annexe B) différente de celle utilisée par les heuristiques stochastiques, mais cette option est sans véritable intérêt ici.

L'utilisation de s_{glou} au lieu de s_{alea} permet d'obtenir plus souvent s^* sauf dans le cas du SA. Excepté dans le cas de l'échange séquentiel, ce résultat est dû uniquement aux solutions optimales obtenues par l'algorithme glouton lui-même. En moyenne, les performances des heuristiques sont dégradées, sauf dans le cas de l'échange séquentiel pour lequel une légère amélioration est observée (Tab. 8.2). Ce résultat global ne signifie pas pour autant qu'une solution initiale particulière issue de l'algorithme glouton à départ aléatoire ne puisse par conduire à une meilleur résultat qu'une solution initiale aléatoire.

Heuristique	$\%\mathcal{H}$	$\%\mathcal{H}'_0$	$\%\mathcal{H}'_T$	$\%\mathcal{H}'$	sign Δ	$\lvert\Delta\rvert$
Echange séquentiel	33.4	15.3	51.8	36.5	+	3.1
Optimisation locale (i)	32.3	15.3	34.9	19.6	-	12.7
Optimisation locale (ii)	29.5	15.1	37.6	22.5	-	7.0
SA	64.6	17.2	68.8	51.6	-	13.0
SA modifié	57.1	14.9	33.4	18.5	-	38.6
TS (i)	48.3	15.3	61.7	46.4	-	1.9
TS (ii)	50.0	17.3	53.4	36.1	-	13.9

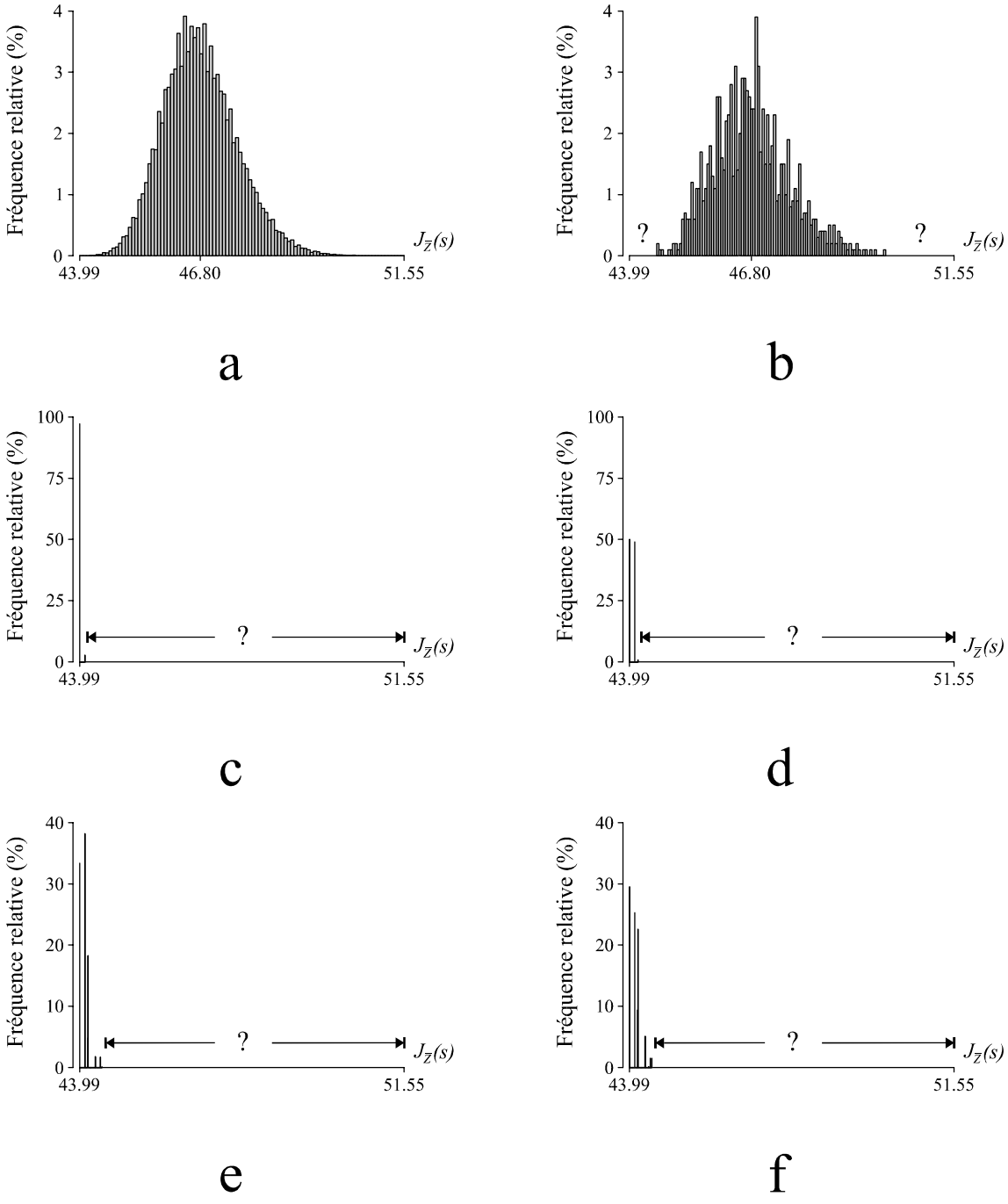
Tableau 8.2: Performances des heuristiques (détails dans le texte) selon le type de solution initiale, aléatoire (s_{alea}) ou obtenue par l'algorithme glouton à départ aléatoire (s_{glou}). $\%\mathcal{H}$: pourcentage de solutions optimales propre à l'heuristique pour s_{alea}. $\%\mathcal{H}'_0$: pourcentage de solutions optimales dues à la solution initiale s_{glou}. $\%\mathcal{H}'_T$: pourcentage de solutions optimales après l'exécution de l'heuristique initialisée par s_{glou}. $\%\mathcal{H}'$= $\%\mathcal{H}'_T$ - $\%\mathcal{H}'_0$: pourcentage de solutions optimales prorpre à l'heuristique pour s_{glou}. Δ : différence $\%\mathcal{H}'$-$\%\mathcal{H}$. Les heuristiques ont été exécutées 10^3 fois.

En ce qui concerne la seule méthode d'amélioration déterministe considérée ici — algorithme d'échange séquentiel de Aspie & Barnes (1990) — une solution initiale sub-optimale améliore légèrement les performances de l'heuristique tout en conduisant à une exécution plus rapide. En effet, pour $N_a = 10^3$ applications de l'algorithme, le nombre de passes k varie comme $2 \leq k \leq 13$ ($k = 6$ en moyenne) pour s_{alea} et comme $1 \leq k \leq 7$ ($k = 3$ en moyenne) pour s_{glou}. Pour une fonction-objectif différente, Aspie & Barnes (1990) mentionnent un nombre de passes inférieur à 5 et souvent de 1 ou 2 lorsque la solution initiale est raisonnable, ce qui semble en assez bon accord avec les résultats que nous obtenons pour $J_{\bar{Z}}$ et s_{glou}.

Cet exemple s'avère très instructif mais, comme pour toute étude empirique, de portée limitée. En effet, le problème traité est de petite taille, les paramétrages ont été choisis avec soin, et l'étude des performances concerne une population bien définie. En particulier, la taille de l'échantillon de $V(s)$ et les paramètres du SA dépendent très étroitement de la taille du problème. En outre, le nombre de passes de l'échange séquentiel dépend également de la taille du problème et de ε, et les considérations de Aspie & Barnes (1990) sur ce sujet ne peuvent pas avoir de valeur générale. Toutefois, certains résultats sont généralisables, notamment :

- l'inefficacité de la recherche aléatoire qui ne parvient pas à identifier les valeurs extrêmes mais uniquement la valeur moyenne (Fig. 8.2.b),
- l'intérêt des heuristiques étudiées qui se concentrent sur un ensemble limité de (bonnes) solutions, contrairement à la recherche aléatoire (Fig. 8.2.c, 8.2.d, 8.2.e & 8.2.f),
- la possibilité d'échantillonner $V(s)$ au lieu de l'examiner de façon exhaustive, dans le cas de l'optimisation locale ou de la TS.

Dans un cas de taille réelle, il devient pratiquement impossible d'optimiser les paramètres des heuristiques. Aussi est-il judicieux d'exécuter au moins une fois les différentes heuristiques d'amélioration itérative afin de retenir la meilleure solution s^+, vraisemblablement assez proche de la solution optimale s^*.



Figure~8.2: Histogrammes des valeurs de la fonction objectif $J_{\bar{Z}}(\cdot)$. (a) Histogramme exhaustif. (b) Histogramme pour la recherche aléatoire. (c) Histogramme pour le SA. (d) Histogramme pour la TS. (e) Histogramme pour l'algorithme d'échange. (f) Histogramme pour l'optimisation locale. Les solutions initiales des heuristiques d'amélioration itérative (c), (d), (e) et (f) sont aléatoires. Les points d'interrogation correspondent aux valeurs de $J_{\bar{Z}}(\cdot)$ pour lesquelles les heuristiques ne proposent pas de solutions. Les heuristiques ont été appliquées 10^3 fois.

8.3.1.3 Algorithmes génétiques

Contrairement aux méthodes d'amélioration itérative qui ne considèrent qu'une solution à la fois, les algorithmes génétiques manipulent une population de solutions. L'implémentation d'un GA implique de choisir le mécanisme de sélection, les types de mutations, et les valeurs de plusieurs paramètres. En pratique, en considérant $N_a = 10^3$ exécutions, et même pour un exemple de petite taille comme celui que nous considérons dans cette section, l'optimisation des valeurs des paramètres s'avère extrêmement coûteuse en temps de calcul. En effet, l'exécution d'un GA requiert un temps de calcul beaucoup plus élevé que celui des heuristiques considérées précédemment. D'après l'étude des performances moyennes pour $N_a = 10^3$, nous avons effectué les choix suivants :

- population de $q = 60$ solutions,
- évolution réalisée sur $\eta = 150$ générations,
- mode de sélection par compétition avec $P_c = 0.90$,
- mutations effectuées selon les probabilités $P_{tl} = 0.15$, $P_i = 0.10$, $P_s = 0.15$.

La mutation par transposition s'avère sans véritable intérêt, aussi nous l'avons supprimée ($P_{tp} = 0$). Les performances du GA calculées d'après $N_a = 10^3$ exécutions s'avèrent moins bonnes que celles du SA ou de la TS puisque $\rho^{*/+} = 40.5$, moy $J_{\bar{Z}}(s^+) = 44.07$ et $\max J_{\bar{Z}}(s^+) = 44.28$. Cette étude de cas confirme que les GA demandent un effort de calcul important tout en ayant des résultats inférieurs à ceux du SA ou de la TS, du moins dans le contexte de l'optimisation combinatoire[6] (Prins 1994). En conséquence, nous ne recommandons pas les GA pour l'optimisation combinatoire de l'échantillonnage, bien qu'ils aient été mentionnés parmi les algorithmes intéressants dans ce contexte (Christakos 1992).

8.3.2 Etude de cas

Afin de juger des performances des heuristiques et de la signification pratique de l'optimisation de l'échantillonnage pour l'estimation globale, nous considérons la population C étudiée dans la Section 6.3.1.8 (p. 171). La population spatiale $\mathcal{U}$ comporte $N = 900$ quadrats organisés selon une grille 30×30 de maille $\Delta = 0.5$ unité. Nous considérons un échantillon de taille $n = 100$. Un modèle exponentiel est ajusté au variogramme empirique. Parmi les neuf grilles 10×10 qu'il est possible d'implémenter dans la grille 30×30, nous savons que le dispositif central est optimal au sens de σ^2_E (Aubry & Debouzie 1999a). Nous conjecturons que ce dispositif est également optimal parmi tous les échantillons $s \in S$. Nous connaissons ainsi s^* bien qu'il soit impossible d'examiner les $\mathrm{Card}(S) > 9 \times 10^{134}$ échantillons possibles. Pour un échantillon s^+ issu de l'optimisation combinatoire selon l'heuristique $\mathcal{H}$, il est alors possible de calculer la performance relative (Prins 1994) :

$$R_{\mathcal{H}}\left(s^+\right) = \frac{J\left(s^+\right)}{J\left(s^*\right)} \tag{8.18}$$

Nous appliquons les heuristiques étudiées précédemment, sauf les GA et les versions de l'optimisation locale et de la TS qui examinent $V(s)$ de façon exhaustive. Afin

[6] La TS semble être moins performante que les GA dans le cadre de l'optimisation continue, notamment dans le cas de l'optimisation des modèles d'agrosystèmes (Mayer *et al.* 1998).

d'apprécier l'impact du modèle de variogramme $\widetilde{\gamma}(\cdot)$ sur le classement des résultats obtenus, nous considérons deux modèles alternatifs Expo $(0, 12313.834, 14.465)$ (modèle θ_1) et Expo $(0, 12316.199, 13.628)$ (modèle θ_2). En considérant l'échantillon systématique central s^*, les variances σ_E^2 sont, respectivement, $J_{\bar{Z}}^{(1)}(s^*) = 8.88104$ et $J_{\bar{Z}}^{(2)}(s^*) = 9.15999$.

8.3.2.1 Algorithme glouton

L'application de l'algorithme glouton à départ multiple examinant toutes les unités de $\mathcal{U}$ ne permet pas d'obtenir s^* contrairement au cas de petite taille traité précédemment. L'algorithme glouton à départ aléatoire fournit les solutions $J_{\bar{Z}}^{(1)}(s^+) = 11.65485$ et $J_{\bar{Z}}^{(2)}(s^+) = 11.60345$, ce qui correspond à une performance relative médiocre $R_{\mathcal{H}}(s^+) \simeq 1.30$.

8.3.2.2 Méthodes d'amélioration itérative

Chaque heuristique d'amélioration itérative est exécutée en considérant une solution initiale s_0 aléatoire (s_{alea}) ou obtenue par l'algorithme glouton à départ aléatoire (s_{glou}), ce qui conduit à deux valeurs J_{alea} et J_{glou}, pour chacun des deux modèles, θ_1 et θ_2. Les paramètres des heuristiques stochastiques ont été adaptés à la taille du problème :

- recuit simulé

  $\pi_0 = 0.99$, $\pi_{\min} = 10^{-10}$, $\delta = 0.99$, $\eta_0 = 1000$

- recuit simulé modifié

  $\pi_0 = 0.7$, $\pi_{\min} = 10^{-4}$, $\delta = 0.73$, $\eta_0 = N - n$, $\alpha = 0.98$, $\eta = 1000$

- recherche taboue (ii)

  $\omega = 35$, $\eta_0 = 10000$, $\eta = 35$

L'utilisation d'une solution initiale issue de l'algorithme glouton est défavorable pour le SA ainsi que pour l'optimisation locale, et favorable pour l'échange séquentiel. Le SA s'avère en fait incapable d'améliorer la solution s_{glou}. En ce qui concerne la TS, l'utilisation de s_{glou} améliore la solution uniquement dans le cas du modèle θ_2. La meilleure solution est obtenue par la TS avec s_{alea} dans le cas du modèle θ_1 et par échange séquentiel avec s_{glou} dans le cas du modèle θ_2 (Tab. 8.3).

	Modèle θ_1		Modèle θ_2	
Heuristique	J_{alea}	J_{glou}	J_{alea}	J_{glou}
Echange séquentiel	9.55167	9.44796	9.86508	*9.76416*
Optimisation locale (ii)	9.67233	9.82000	9.86535	9.98571
SA	9.70395	9.70493	10.00399	10.00833
SA modifié	9.98690	11.65485	10.02644	11.60345
TS (ii)	*9.54423*	9.61790	9.96713	9.95472

Tableau 8.3: Résultats des heuristiques pour les modèles θ_1 et θ_2. J_{alea} et J_{glou} : valeurs de $J_{\bar{Z}}(\cdot)$ pour la solution obtenue à partir d'une solution initiale aléatoire *vs.* issue de la méthode glouton à départ aléatoire (meilleures solutions en italique).

Ces résultats illustrent l'intérêt d'une utilisation conjointe de différentes stratégies, *i.e.* différentes heuristiques et solutions initiales. Les solutions retenues correspondent à une bonne performance relative $R_{\mathcal{H}}(s^+) \simeq 1.07$.

Dans l'ensemble, les différentes heuristiques se comportent assez bien, quel que soit le modèle de variogramme considéré. Les paramètres du SA, et surtout de la TS, sont certainement robustes vis-à-vis du modèle. Mais les résultats issus du SA modifié s'avèrent extrêmement sensibles aux paramètres α et δ, et cette sensibilité est elle même fonction du modèle, au moins dans le cas de δ. Ainsi, une modification de α d'une amplitude aussi faible que 0.01 peut faire passer d'une solution satisfaisante $J_{\bar{Z}} \simeq 10$ à une mauvaise $J_{\bar{Z}} \simeq 38$, qu'il s'agisse du modèle θ_1 ou du modèle θ_2. Pour une modification de δ de même ampleur, la variation est moindre et diffère en signe et en amplitude selon le modèle (Tab. 8.4). En conséquence, nous ne recommandons pas la méthode du SA modifié.

D'une façon générale, il est préférable d'utiliser des heuristiques simples, ne comportant pas trop de paramètres à régler, et suffisamment robustes. Nous retenons donc l'échange séquentiel, l'optimisation locale et la TS avec échantillonnage de $V(s)$, ainsi que le SA sous sa forme originelle.

		θ_1	θ_2
α	δ	J_{alea}	J_{alea}
0.98	0.80	10.10395	10.39443
0.97	0.80	37.79646	38.07729
0.98	*0.73*	9.98690	10.08976
0.98	*0.72*	14.32413	10.02201

Tableau 8.4: Etude de la sensibilité de l'algorithme du SA modifié aux valeurs des paramètres α et δ pour les modèles θ_1 et θ_2. J_{alea} : valeur de $J_{\bar{Z}}(\cdot)$ pour la solution obtenue à partir d'une solution initiale aléatoire (valeurs modifiées en italique).

8.3.2.3 Interprétation de l'optimisation

Dans l'exemple étudié — domaine carré, autocorrélation spatiale exponentielle — les heuristiques d'optimisation combinatoire utilisées tendent vers un motif d'échantillonnage systématique centré, sans toutefois l'obtenir exactement (Fig. 8.3). Ces résultats semblent confirmer que pour la superpopulation considérée, l'échantillon systématique centré est le plus précis parmi tous les échantillons possibles. Cependant, pour la population étudiée C, nous savons que cet échantillon qualifié d'optimal ne conduit pas à l'estimation de z_D la plus précise, bien que nous ignorions exactement quel est l'échantillon le plus précis — pour le savoir, il faudrait énumérer les $\binom{900}{100} \simeq 9.384 \times 10^{134}$ échantillons possibles. Du point de vue de l'écologiste, ce résultat peut apparaître déconcertant : dans ce cas, ce ne sont pas les heuristiques qu'il convient de remettre en question mais l'interprétation de la fonction-objectif qu'il faut préciser.

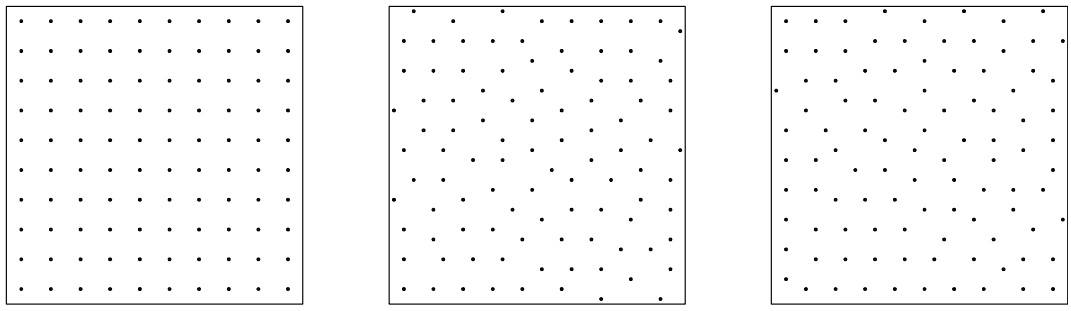
Pour un motif d'échantillonnage spatial particulier, l'efficacité de $\bar{z}$ est évaluée en moyenne pour l'infinité des populations issues de la superpopulation en calculant $\mathrm{Var}_\xi\left[\bar{Z} - Z_D\right]$. Nous avons considéré jusqu'à présent que seul le variogramme était stable dans le temps, et connu. Dans ces conditions, la variance $\mathrm{Var}_\xi\left[\bar{Z} - Z_D\right]$ est calculée pour des réalisations non conditionnelles de la fonction aléatoire $Z(\cdot)$, et il s'agit par conséquent

de la variance σ_E^2. La minimisation de σ_E^2 conduit à sélectionner les unités de l'échantillon de telle façon que l'estimation de z_D soit la plus précise possible en ignorant *a priori* quelle sera la distribution des valeurs de $z(\cdot)$ sur le domaine D, mais en tenant compte de la structure d'autocorrélation spatiale, de la répartition spatiale des unités d'échantillonnage et de la géométrie du domaine d'étude. En pratique, l'échantillon optimal pour la superpopulation ne fournit pas nécessairement l'estimation la plus précise pour une population particulière parce qu'il est impossible de définir l'échantillon le plus précis sans une connaissance exhaustive de la population. Même si nous connaissions complètement la population, nous serions néanmoins incapables de déterminer l'échantillon le plus précis à cause de l'explosion combinatoire du nombre d'échantillons possibles.

Pour optimiser l'échantillonnage dans le cas d'une population particulière ou d'une superpopulation de plus grande stabilité spatiale que précédemment, il faudrait nécessairement introduire davantage d'information dans la fonction-objectif, notamment en prenant en compte la distribution des valeurs dans D. Une telle approche nécessiterait de recourir à la variance d'erreur d'estimation conditionnée σ_C^2. D'un point de vue opératoire, il est possible de minimiser σ_E^2 soit en se référant à la formule de Matheron, soit en utilisant un ensemble de réalisations non conditionnelles conservé dans la mémoire de l'ordinateur (*e.g.*, 10^4 réalisations). En ce qui concerne la minimisation de σ_C^2, il suffirait de suivre la seconde option en remplaçant les réalisations non conditionnelles par des réalisations conditionnelles.

Enfin, la stabilité spatiale de la superpopulation est le plus souvent appréciée subjectivement. Une évalutation objective de cette stabilité nécessite de disposer de plusieurs échantillons répartis dans le temps, de façon à :

- apprécier la stabilité des variogrammes expérimentaux, notamment en vérifiant qu'ils se chevauchent, ou que leurs intervalles de confiance approximatifs se chevauchent (Shafer & Varljen 1990, Section 7.3),
- cartographier les populations à partir des échantillons puis comparer les cartes obtenues entre elles.



Figure˜8.3: Motifs d'échantillonnage de la population C. (a) : Motif optimal pour les modèles θ_1 et θ_2. (b) Motif suboptimal obtenu pour le modèle θ_1 par la TS, à partir d'une solution initiale aléatoire. (c) Motif suboptimal obtenu pour le modèle θ_2 par l'algorithme d'échange, à partir d'une solution initiale issue de la méthode glouton à départ aléatoire.

8.4 Optimisation pour l'estimation locale

Le principe de l'optimisation de l'échantillonnage spatial pour l'estimation locale par krigeage repose sur l'utilisation de la variance de krigeage vue comme une mesure de précision locale (*e.g.*, Chami 1984, Rouhani 1985, Pardo-Igúzquiza 1998b). Dans ce contexte, deux objectifs peuvent être envisagés selon qu'il s'agit :

- d'implémenter de toute pièce un motif d'échantillonnage,
- de modifier un motif d'échantillonnage déjà existant.

Les auteurs déploient beaucoup d'efforts afin d'obtenir des solutions quasi-optimales — au sens de la variance de krigeage — mais la question de l'interprétation de cette optimisation n'est pratiquement jamais abordée. En conséquence, nous ne détaillerons pas les aspects techniques mais nous traiterons de la pertinence de la variance de krigeage en tant que critère à optimiser.

8.4.1 Choix d'un motif d'échantillonnage

Trois types de dispositifs sont classiquement mentionnés pour l'échantillonnage spatial : aléatoire simple, systématique et stratifié (*cf.* Matérn 1960, Ripley 1981, Iachan 1985, Maling 1989, Haining 1990, Cressie 1991, Christakos 1992). Il convient d'abord d'identifier quel est le dispositif conduisant aux motifs d'échantillonnage les plus efficaces.

8.4.1.1 Choix du type d'échantillonnage

En considérant la cartographie par krigeage universel avec un variogramme linéaire et une dérive quadratique, Olea (1984) montre que le meilleur échantillonnage d'une VR isotrope — au sens de la minimisation de la variance de krigeage maximale — est obtenu par une tessellation régulière (échantillonnage systématique), de préférence de tuile hexagonale. L'auteur classe les types de motifs d'échantillonnage par ordre de précision décroissante comme suit : systématique > stratifié > aléatoire > agrégé. Ce résultat n'est pas surprenant compte tenu du critère d'optimalité retenu puisque la variance d'erreur d'estimation tend à répartir régulièrement les supports d'échantillonnage dans le domaine étudié. L'échantillonnage systématique étant jugé optimal, il est possible de préciser quelle doit être la géométrie de la maille élémentaire.

8.4.1.2 Choix de la géométrie de la maille

McBratney *et al.* (1981) et Burgess *et al.* (1981) considèrent qu'un motif d'échantillonnage est d'autant meilleur que la variance de krigeage ordinaire maximale est faible. Dans le cas d'une VR isotrope, les auteurs recommandent l'échantillonnage systématique avec une maille triangulaire équilatérale, tout en reconnaissant que l'utilisation d'une maille carrée est approximativement équivalente.

Dans le cadre du krigeage sous l'hypothèse de stationnarité d'ordre 2, et en faisant référence à un variogramme sphérique, Yfantis *et al.* (1987) comparent les trois tessellations régulières possibles (Section 2.1.2.3, p. 15) et classent les tuiles par ordre de préférence décroissante comme suit : triangle équilatéral > carré > hexagone.

Ces résultats sont conformes à ceux obtenus pour l'estimation globale par minimisation continue de la variance d'erreur d'estimation globale σ_E^2 (Section 5.5.1). En effet, il est généralement préférable d'utiliser un motif d'échantillonnage systématique. Cependant, il n'existe pas de géométrie de la maille qui soit optimale quel que soit le type de fonction aléatoire ou la forme de l'autocorrélation spatiale. Bien, qu'il s'agisse d'un problème intéressant sur le plan théorique, le problème de la forme optimale de la maille n'a pas de véritable conséquence pratique (Matérn 1960, p. 74). Même si la maille carrée n'est pas toujours optimale, la suboptimalité est négligeable et l'implémentation s'avère beaucoup plus facile qu'avec des mailles triangulaires ou hexagonales. *A priori*, pour une VR isotrope, il semble donc raisonnable de recommander un motif d'échantillonnage systématique de maille carrée.

8.4.2 Modification d'un motif d'échantillonnage

L'optimisation d'un motif d'échantillonnage s déjà implémenté peut consister :

- à supprimer des supports tout en minimisant la perte de précision,
- à ajouter des supports tout en maximisant le gain de précision.

Parce que s est nécessairement fini, la suppression de supports d'échantillonnage est un problème de nature combinatoire consistant à identifier quel sous-ensemble $s' \subset s$ peut être supprimé sans toutefois dépasser une certaine perte de précision.

Chen *et al.* (1995) considèrent une tolérance maximale pour la variance de krigeage et un variogramme connu *a priori*. Les auteurs étudient l'influence de la réduction de la densité d'échantillonnage sur l'efficacité de l'échantillon en examinant la corrélation entre les estimations par krigeage calculées à partir de l'échantillon s et celles calculées à partir d'un sous-ensemble $s - s'$. Au moins lorsque la VR est spatialement régulière, Chen *et al.* (1995) montrent qu'il est possible d'utiliser moins d'observations sans perte significative d'information, et concluent qu'une grille comportant seulement la moitié du nombre de supports initial n s'avère plus efficace qu'un dispositif irrégulier comportant n supports. Pour une VR spatialement peu régulière dont le variogramme est de type exponentiel, Gallichand *et al.* (1992) montrent également qu'il est possible de réduire considérablement l'espacement des supports situés sur une grille d'échantillonnage sans perte d'information importante. Ces résultats semblent indiquer que l'implantation régulière des supports d'échantillonnage dans D prime sur la taille de l'échantillon.

L'augmentation de s peut se traiter comme un problème d'optimisation combinatoire si le nombre de nouveaux supports possibles m est fini, ou comme un problème d'optimisation continue si m est infini. Cependant, on a rarement la flexibilité de localiser les supports de façon idéale (Ashraf *et al.* 1997), aussi nous trouvons à nouveau plus réaliste de considérer l'optimisation combinatoire plutôt que l'optimisation continue.

L'approche la plus élémentaire consiste à identifier les zones où la variance de krigeage est maximale, pour y implanter, à vue, de nouveaux supports de mesure (*e.g.*, Chami 1984). Cependant, cette méthode présente l'inconvénient d'ignorer l'impact global sur la précision des estimations sur D (Rouhani 1985). Une meilleure approche consiste à mettre à jour la variance de krigeage lorsqu'on ajoute un nouveau support d'échantillonnage dans s. Le gain "d'information" en n'importe quel point de D est alors apprécié

classiquement en termes de réduction de la variance de krigeage : ce type d'approche constitue par conséquent une *analyse de réduction de variance* (Rouhani 1985, Rouhani & Hall 1988).

Il est utile de pouvoir recalculer les variances de krigeage autrement qu'en résolvant à chaque fois un nouveau système de krigeage. Ainsi, des méthodes de mise-à-jour efficaces ont été proposées, notamment par Rouhani (1985) et Christakos & Olea (1992) dans le cas général du krigeage en FAI-k, par Barnes & Watson (1992) dans le cas du krigeage ordinaire, et par Gao *et al.* (1996) dans le cas du krigeage universel. Mais si l'on se place dans le cadre d'un échantillonnage séquentiel — en procédant successivement à de nouvelles mesures et pas simplement en plaçant de nouveaux supports — il conviendrait d'estimer à nouveau la structure d'autocorrélation spatiale (variogramme, covariance ou covariance généralisée) (Rouhani 1985). Dans ce contexte précis, les méthodes de mise-à-jour des variances de krigeage sont sans objet.

Enfin la modification de s peut concerner la position des supports sans en changer le nombre. A nouveau, la modification du motif d'échantillonnage initial peut être traitée comme un problème d'optimisation combinatoire ou continue. Ainsi, dans un contexte d'optimisation combinatoire, Sacks & Schiller (1988) cherchent à minimiser la variance de krigeage maximale au moyen de l'algorithme de SA modifié tandis que van Groenigen *et al.* (1999) utilisent la technique du SA originelle afin de minimiser l'intégrale d'espace des variances de krigeage locales, *i.e.* dans le cadre de l'optimisation continue.

8.4.3 Interprétation de l'optimisation

L'utilisation de la variance de krigeage pour optimiser l'échantillonnage en vue de la cartographie par krigeage peut se justifier :

- en considérant qu'il est souhaitable de répartir de façon homogène les supports d'échantillonnage au sein de D,
- en considérant que la variance de krigeage est une mesure de précision des estimations locales.

Si la première option est *a priori* raisonnable en absence d'une connaissance approfondie de la VR étudiée, en revanche la pertinence de la seconde option — qui n'est pas nécessairement une conséquence logique de la première — mérite d'être discutée à la lumière de l'interprétation du modèle de superpopulation sous-jacent (fonction aléatoire), de la stabilité spatiale du phénomène modélisé, et de la quantité d'information disponible au préalable.

Il existe plusieurs interprétations du concept de superpopulation (Cassel *et al.* 1977, p. 81), mais l'utilisation des fonctions aléatoires est essentiellement justifiée comme un moyen de modéliser un phénomène naturel se déployant dans un domaine fixé D. Ce point étant acquis, il convient de poser au moins trois questions :

- le phénomène est-il unique ou bien susceptible de se répéter dans le temps ?
- si le phénomène n'est pas unique, quelle est la stabilité de sa structure spatiale ?
- de quelles données objectives dispose-t-on au préalable ?

A la lumière des réponses à ces trois questions, l'utilisation de la variance de krigeage peut apparaître illusoire, pertinente, ou mal adaptée. En effet, si le phénomène est unique (*e.g.*, la géologie, la pédologie, la topographie), la variance de krigeage ne peut pas être considérée comme une mesure de précision au sens strict, et il n'y a aucune garantie que le motif d'échantillonnage optimisé conduise effectivement à une cartographie plus précise qu'un autre motif. La variance de krigeage ne représente alors qu'un moyen parmi d'autres de répartir régulièrement les sites d'échantillonnage dans D. En revanche, si le phénomène se répète dans le temps (*e.g.*, les précipitations, la chute des fruits sous un arbre), alors la variance de krigeage constitue une mesure de précision, mais uniquement en moyenne, pour un grand nombre d'occurrences dudit phénomène. Dans ce contexte, il est raisonnable d'utiliser la variance de krigeage comme critère à optimiser, bien que l'optimisation ne signifie nullement que la cartographie sera la plus précise possible à chaque occurrence du phénomène.

La stabilité de la structure spatiale du phénomène joue un rôle primordial dans l'utilisation de la variance de krigeage. Si le variogramme de la VR associée au phénomène n'est pas stable au cours du temps, le recours à la variance de krigeage n'est pas fondé. Lorsque la stabilité est importante et ne concerne pas uniquement la structure d'autocorrélation spatiale, la variance de krigeage constitue un critère mal adapté, et il faudrait conditionner par les données disponibles, par exemple en utilisant la simulation conditionnelle (*e.g.*, Easley *et al.* 1991, van Groenigen *et al.* 1997). Dans cette optique, Cox *et al.* (1997) mentionnent une procédure d'échantillonnage séquentiel du type suivant :

1. Ajustement d'un modèle spatial aux données disponibles.
2. Simulation conditionnelle en utilisant le modèle.
3. Calcul de la fonction-objectif.
4. Sélection des unités d'échantillonnage supplémentaires.
5. Nouvelles mesures et mise à jour du modèle spatial, retourner en 1 jusqu'à ce qu'un compromis soit trouvé entre le coût d'acquisition des mesures et le coût des erreurs par manque de données.

8.5 Optimisation pour l'estimation du variogramme

L'optimisation de l'échantillonnage pour l'estimation de la moyenne globale supposait *a minima* la connaissance de la forme du variogramme. Il est possible que l'objectif de l'échantillonnage consiste justement à estimer le variogramme lui-même. D'une façon générale, il est légitime de poser le problème de l'optimisation de l'échantillonnage pour l'estimation du variogramme local, qu'il s'agisse de l'utiliser à des fins strictement descriptives (*e.g.*, Schotzko & O'Keeffe 1990), ou plus généralement, afin de mener les calculs géostatistiques.

Le motif d'échantillonnage utilisé en vue de l'estimation du variogramme peut être produit selon les dispositifs d'échantillonnage habituels tels que l'échantillonnage aléatoire simple (EAS), l'échantillonnage systématique (ES), ou l'échantillonnage aléatoire stratifié par une grille, à un élément par strate (STR). En dehors de ces dispositifs classiques, il est parfois recommandé d'utiliser un échantillonnage emboîté (EE) dont l'analyse de variance hiérarchique associée permet de construire des estimateurs du variogramme (Oliver &

Webster 1986, 1991, Oliver 1992, Pettitt & McBratney 1993, Corsten & Stein 1994). De tels dispositifs sont notamment proposés en géographie physique (Oliver *et al.* 1989b) et en nématologie (Webster & Boag 1992a, 1992b).

L'intérêt de l'EE réside essentiellement dans sa capacité à explorer tout un intervalle de distances, ce qui permet de connaître le comportement à l'origine, et garantit que la portée du variogramme sera effectivement détectée. Au contraire, avec l'ES par exemple, il est possible que le pas d'échantillonnage Δ soit trop grand vis-à-vis de la portée du variogramme a, ce qui introduit inévitablement un fort effet de pépite si Δ est légèrement inférieur à a, voire même un effet de pépite pur si $\Delta \geq a$. Toutefois, l'estimation du variogramme par l'intermédiaire de l'analyse de variance hiérarchique associée à l'EE est assez grossière de sorte qu'un second échantillon demeure nécessaire afin d'estimer précisément le variogramme (Oliver 1992). En outre, Corsten & Stein (1994) concluent que les dispositifs habituels peuvent s'avérer supérieurs aux dispositifs emboîtés. En conséquence, dans ce qui suit nous n'examinons pas l'échantillonnage emboîté en vue de l'estimation du variogramme.

Dans le cadre de l'optimisation continue, Russo (1984) propose une méthode d'amélioration itérative à partir d'un échantillon initial, minimisant la variabilité des distances h entre les classes — idéalement, $N(h) = cste$ — et à l'intérieur des classes. La méthode présente cependant l'inconvénient d'être sensible au motif d'échantillonnage initial, ce qui revient à dire qu'elle aboutit à une solution seulement localement optimale. Bien que l'auteur ne distingue pas les erreurs d'estimation et de fluctuation (Section 4.2.1.4), la méthode proposée est vraisemblablement censée améliorer l'estimation du variogramme théorique. En suivant le même type d'approche que Russo (1984), Russo & Jury (1988) considèrent que $N(h)$ doit être constant dans le cas d'un échantillonnage dense, mais peut être variable lorsque la densité d'échantillonnage est moindre.

De façon similaire, Warrick & Myers (1987) considèrent que les nombres de vecteurs inter-supports $\mathbf{h}$ figurant dans les classes de distances et d'angles sont fixés *a priori*, et que le problème est d'approximer le mieux possible cette distribution de reférence. Ces auteurs proposent notamment d'utiliser un algorithme d'optimisation continue procédant par substitution de points tirés au hasard. Récemment, la fonction-objectif proposée par Warrick & Myers (1987) a été reprise par van Groenigen & Stein (1998), et utilisée dans un contexte d'optimisation continue exploitant cependant la technique du SA.

Contrairement aux approches précédentes, Bogaert & Russo (1999) ne considèrent plus la distribution des vecteurs inter-supports $\mathbf{h}$ dans des classes de distances et d'angles fixées *a priori*, mais l'estimation des paramètres d'un modèle de variogramme théorique par ajustement GLS du variogramme expérimental (Section 7.2.3.3). Bogaert & Russo (1999) se placent dans le contexte de la D-optimalité et cherchent à minimiser le déterminant de la matrice de covariance des paramètres du modèle de variogramme. Ces auteurs proposent un algorithme d'optimisation continue procédant par amélioration itérative à partir d'une solution initiale aléatoire. Afin d'éviter les optima locaux trop éloignés de l'optimum absolu, Bogaert & Russo (1999) explorent l'espace des solutions en exécutant leur algorithme à partir de plusieurs solutions initiales générées aléatoirement. Cette approche est très intéressante dans le contexte de l'estimation du variogramme théorique car elle prend en compte explicitement le type de modèle et la procédure d'ajustement du variogramme expérimental. Néanmoins, la méthode proposée suppose une FAST-2 gaussienne, ce qui peut éventuellement apparaître comme une hypothèse trop restrictive.

Dans le cadre de l'échantillonnage écologique, et sans discuter pour le moment la pertinence des fonctions-objectif utilisées, il nous semble nécessaire de reformuler le problème.

En premier lieu, considérer l'optimisation de la fonction-objectif en déplaçant les sites initiaux, que ce soit de façon déterministe (Russo 1984) ou stochastique (van Groenigen & Stein 1998), ou en substituant des sites choisis au hasard (Warrick & Myers 1987, Bogaert & Russo 1999), nous semble peu réaliste. En effet, comment être assuré que ces sites seront accessibles sur le terrain et pourront effectivement faire l'objet de mesures ou d'observations ?

En second lieu, il nous semble que considérer l'optimisation de façon continue constitue une inutile surspécification du problème. En conséquence, nous considérons à nouveau une population finie $\mathcal{U}$ de N unités dont il sera effectivement possible de prélever un échantillon de taille n. Dans ce cadre, l'optimisation de l'échantillonnage pour l'estimation du variogramme nécessite de définir une fonction-objectif $J_\gamma(\cdot)$ à minimiser, ainsi que les heuristiques utilisables. Enfin, contrairement aux méthodes issues de la littérature dont l'intérêt n'est généralement pas démontré (*e.g.*, Russo 1984, Warrick & Myers 1987, van Groenigen & Stein 1998), nous proposons d'évaluer l'optimisation de l'échantillonnage pour l'estimation du variogramme local au moyen d'une étude de Monte-Carlo.

8.5.1 Fonctions-objectif

Warrick & Myers (1987) considèrent une fonction-objectif de la forme :

$$SS(s) = a \sum_{i=1}^{k} \omega_i \left(n_i^{th} - n_i^{obs}\right)^2 + b \sum_{i=1}^{k} m_{1i} + c \sum_{i=1}^{k} m_{2i} \tag{8.19}$$

avec (a, b, c) un triplet de pondérateurs spécifié *a priori*, n_i^{th} l'effectif théorique et n_i^{obs} l'effectif observé pour la classe i, ω_i le poids de l'écart-quadratique pour la classe i, m_{1i} la dispersion dans la classe de distances i et m_{2i} la dispersion dans la classe d'angles i. Bien que Warrick & Myers (1987) estiment nécessaire que les distances et les angles moyens soient proches des centres des classes, ces critères n'apparaissent pas explicitement dans la fonction-objectif (8.19), mais doivent être pris en compte par les deux derniers termes.

Ces auteurs considèrent que le nombre $N(\mathbf{h})$ doit être le plus élevé possible pour chaque classe, et particulièrement pour les classes à faible distance. Finalement, les auteurs n'envisagent que le cas isotrope et différents jeux de paramètres afin d'obtenir une répartition uniforme des distances dans les classes. Il convient d'examiner ce qui peut justifier les choix de Warrick & Myers (1987).

En premier lieu, il y aurait un intérêt pratique à essayer d'obtenir l'uniformité de $N(h)$ pour toutes les classes de distances, mais aucune argumentation précise ne supporte cette assertion, et les avis sur ce sujet sont totalement contradictoires, vraisemblablement à cause d'objectifs différents. Ainsi, pour Legendre & Fortin (1989), l'uniformité permet de calculer des estimations valides pour les distances élevées, tandis que pour Webster & Oliver (1992a), les classes les plus importantes correspondent aux petites distances, ce qui suggère qu'il y a peu d'intérêt pratique à essayer d'obtenir un nombre égal de valeurs à toutes les classes.

En second lieu, l'estimation préférentielle du variogramme aux premières classes de distances fait implicitement référence à l'erreur de fluctuation $\gamma_D(h) - \gamma(h)$. En effet, on considère classiquement que l'erreur de fluctuation est très forte au-delà des faibles distances de sorte qu'il serait illusoire de chercher à estimer précisément le variogramme théorique pour toutes les distances (Matheron 1965, Journel & Huijbregts 1978). Dans ce contexte, Morris (1991) fait remarquer que la précision de $\widehat{\gamma}(h)$ est affectée, non seulement par $N(h)$, mais aussi par les corrélations entre les valeurs du variogramme. Morris (1991) propose une expression pour la variance de $\widehat{\gamma}(h)$ qui inclut à la fois la variance de fluctuation et la variance d'estimation, et considère qu'il s'agit d'une mesure d'incertitude de $\widehat{\gamma}(h)$ vu comme un estimateur de $\gamma(h)$ plutôt que de $\gamma_D(h)$. Cet auteur suggère un indice tenant compte de la corrélation entre valeurs du variogramme et l'utilise à la place du nombre de paires de distances dans les procédures de Russo (1984) ou de Warrick & Myers (1987).

Mais, si le phénomène étudié est unique dans un domaine donné D (*e.g.*, la pédologie), l'erreur de fluctuation n'a aucun caractère objectif, et la modélisation de la VR correspondante $z(\cdot)$ par une fonction aléatoire $Z(\cdot)$ n'implique pas l'immanence d'un variogramme théorique inconnu que l'on chercherait à estimer. Dans ce cas, nous considérons que l'objet d'étude est $z(\cdot)$ tandis que $Z(\cdot)$ joue un rôle strictement opératoire. Aussi, nous construisons notre modèle en identifiant $\gamma_D(h)$ à l'espérance $\mathrm{E}_\xi[\gamma_D(h)]$, comme nous l'avons fait dans les Chapitres 6 & 7. Ce n'est donc pas l'erreur de fluctuation qui nous intéresse mais bien l'erreur d'estimation $\widehat{\gamma}_D(h) - \gamma_D(h)$. Comme cela a été illustré dans la Section 7.3.2.1, la largeur de l'intervalle de confiance de $\widehat{\gamma}_D$ augmente avec la distance, de sorte qu'il faudrait plutôt augmenter le nombre de couples $N(h)$ pour les fortes distances plutôt que pour les faibles. Nous ne cherchons donc pas nécessairement à favoriser les premières classes de distances.

Finalement, nous posons le problème en termes d'erreur d'estimation de l'intégrale d'espace $\gamma_D(h)$ par l'approximation discrète $\widehat{\gamma}(h)$, et cela pour une série de k distances $h \in]0, L_{\max}/2]$ avec $L_{\max}$ plus grande distance au sein du domaine D. En conséquence, nous ne faisons pas du tout référence aux propriétés d'une FA comme le font, par exemple, Bogaert & Russo (1999). Il reste à définir une fonction-objectif cohérente avec le problème posé. En absence de toute connaissance *a priori*, seuls des critères géométriques peuvent intervenir dans le choix des unités d'échantillonnage. Quel que soit le critère choisi, son intérêt peut être déterminé *a posteriori*, par exemple grâce à une étude de Monte-Carlo. Il est possible de considérer deux types de critères selon que la fonction-objectif manipule directement :

- la répartition des unités échantillonnées dans D,
- la répartition des distances dans les classes utilisées pour calculer $\widehat{\gamma}(\cdot)$.

Les deux répartitions sont évidemment liées mais elles ne sont généralement pas considérées ensemble dans une même fonction-objectif. Les unités peuvent être réparties régulièrement (*e.g.*, selon une grille) afin d'explorer de façon homogène le domaine, aléatoirement, afin d'échantillonner la distribution des distances dans D, ou résulter d'un compromis obtenu par échantillonnage aléatoire stratifié par une grille. La répartition des distances dans les classes peut être ajustée afin de suivre une certaine distribution de référence.

Nous choisissons d'optimiser le second type de répartition en utilisant une fonction de coût de la forme :

$$J_\gamma(s) = \sum_{i=1}^{k} \left(n_i^{th} - n_i^{obs}\right)^2 + \alpha \sum_{i=1}^{k} \left(c_i - \bar{d}_i\right)^2 \tag{8.20}$$

avec α une constante multiplicative. Pour une classe de distances C_i de centre c_i, nous notons n_i^{th} l'effectif théorique, n_i^{obs} l'effectif observé correspondant à la distribution de distances induite par l'échantillon s et $\bar{d}_i$ la moyenne des n_i^{obs} distances :

$$\bar{d}_i = \frac{1}{n_i^{obs}} \sum_{h \in C_i} h \tag{8.21}$$

Nous ne considérons que le cas isotrope et ne tenons pas compte de la dispersion intra-classe, l'expérience nous ayant montré que ce critère n'était pas utile (voir aussi van Groenigen & Stein 1998).

Soit G l'ensemble des distances $h \in]0, L_{\max}/2]$ induites dans D par s. Soit $f(h)$ la fonction densité de probabilité de référence. L'effectif théorique n_i^{th} de la classe $C_i = \{h \mid h \in [a, b]\}$ est calculé comme :

$$n_i^{th} = \Pr\left(a < h < b \mid h \in]0, L_{\max}/2]\right) \cdot \mathrm{Card}(G) \tag{8.22}$$

avec

$$\Pr\left(a < h < b \mid h \in]0, L_{\max}/2]\right) = \int_a^b f(h)\, dh \;/\; \int_0^{L_{\max}/2} f(h)\, dh \tag{8.23}$$

Différentes fonctions de densité de probabilité peuvent être retenues, notamment une fonction trapézoïdale de la forme :

$$f(h) = \begin{cases} y_a + (y_0 - y_a)\,(h-a)\,/\,(b-a) & \text{si } h \in [a, b[\\ y_0 & \text{si } h \in [b, c] \\ y_0 + (y_d - y_0)\,(h-c)/\,(d-c) & \text{si } h \in]c, d] \\ 0 & \text{sinon} \end{cases} \tag{8.24}$$

avec

$$y_0 = \frac{1 - \frac{1}{2} y_a\,(b-a) - \frac{1}{2} y_d\,(d-c)}{(c-b) + \frac{1}{2}\,(b-a) + \frac{1}{2}\,(d-c)} \tag{8.25}$$

où $a \leq b \leq c \leq d$ sont des points dans $\mathbb{R}$, $y_a = f(a)$ et $y_d = f(d)$ (Fig. 8.4). Le cas particulier de la loi uniforme (loi rectangulaire) est obtenu pour $y_a = y_d = 0$, $a \equiv b$ et $c \equiv d$.

Une autre fonction intéressante est la densité de probabilité $f_D(h)$ des distances h au sein d'un domaine D. Lorsque D est un polygone convexe à β sommets, la fonction de répartition $F(h)$ est donnée par un résultat issu de la *géométrie probabiliste*[7] (Borel 1958)[8] :

$$\Pr(x \leq h) = \frac{1}{S^2} \left\{ \pi S h^2 - \frac{2}{3} P h^3 + \frac{h^4}{8} \sum_{i=1}^{\beta} \left[(\pi - A_i)\right] \cot A_i + 1 \right\} \tag{8.26}$$

[7] Pour une revue informelle de la *géométrie probabiliste cf.* Baddeley (1982).

[8] Cette formule — imprimée avec des coquilles — n'est valide que pour D convexe et dans le cas où h reste petit par rapport aux côtés de D (Borel 1958).

avec S la surface de D, P son périmètre et A_i l'angle au sommet i. Dans le cas particulier où D est carré, l'expression se simplifie en :

$$\Pr(x \leq h) = \frac{1}{S^2}\left\{\pi S h^2 - \frac{2}{3}Ph^3 + \frac{h^4}{2}\right\} \tag{8.27}$$

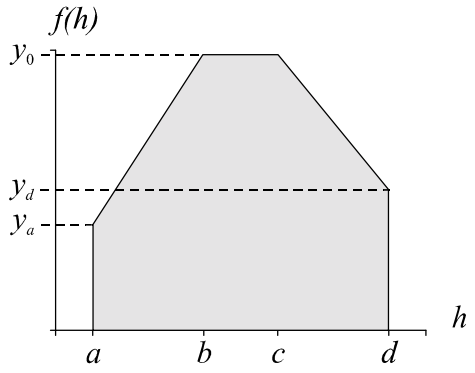
La fonction densité de probabilité correspondante $f_D(h)$ peut être approximée par discrétisation puisque :

$$f_D(h) = \lim_{\Delta h \to 0} F(h + \Delta h) - F(h) \tag{8.28}$$

Si D est un polygone concave, il devient nécessaire d'approximer $f_D(h)$ par l'histogramme des distances inter-points calculées pour un ensemble de points aléatoires discrétisant D. Cette approximation de Monte-Carlo nécessite plusieurs opérateurs géomatiques afin :

1. de calculer le rectangle minimum englobant R de D (Section 2.2.2.3),
2. d'implanter dans R un échantillon aléatoire très dense, de préférence stratifié par une grille,
3. de restreindre l'échantillon aux points inclus dans D (Section 2.2.2.1),
4. de contrôler que les segments supports des vecteurs inter-points $\mathbf{h}_{i,j}$ utilisés dans l'approximation de $f_D(h)$ sont bien inclus dans D, *i.e.* ne traversent pas la frontière ∂D.

L'étape (4) constitue la phase la plus compliquée sur le plan algorithmique, et corrélativement, la plus coûteuse en temps de calcul.



Figure˜8.4: Fonction de densité de probabilité de type trapézoïdale.

8.5.2 Heuristiques

Parmi les heuristiques étudiées, l'algorithme glouton n'est pas utilisable. En effet, la minimisation de $J_\gamma(\cdot)$ constitue un exemple typique de problème pour lequel une suite de décisions locales optimales conduit à un très mauvais résultat global. En partant d'une unité u_1, l'algorithme glouton cherche une deuxième unité u_2 minimisant $J_\gamma(\cdot)$, *i.e.* très proche de u_1 de façon à figurer dans la même classe. Lors du déroulement de l'algorithme,

les premières classes de distances se remplissent, puis l'heuristique tente finalement de compenser les décisions désastreuses effectuées en début d'exécution. En conséquence, nous ne retenons ici que les heuristiques d'amélioration itérative suivantes :

- échange séquentiel,
- optimisation locale,
- recuit simulé,
- recherche taboue.

Pour chaque minimisation, le résultat retenu est la meilleure solution obtenue après exécution de chacune des heuristiques, à partir d'une même solution initiale aléatoire.

8.5.3 Etude de Monte-Carlo

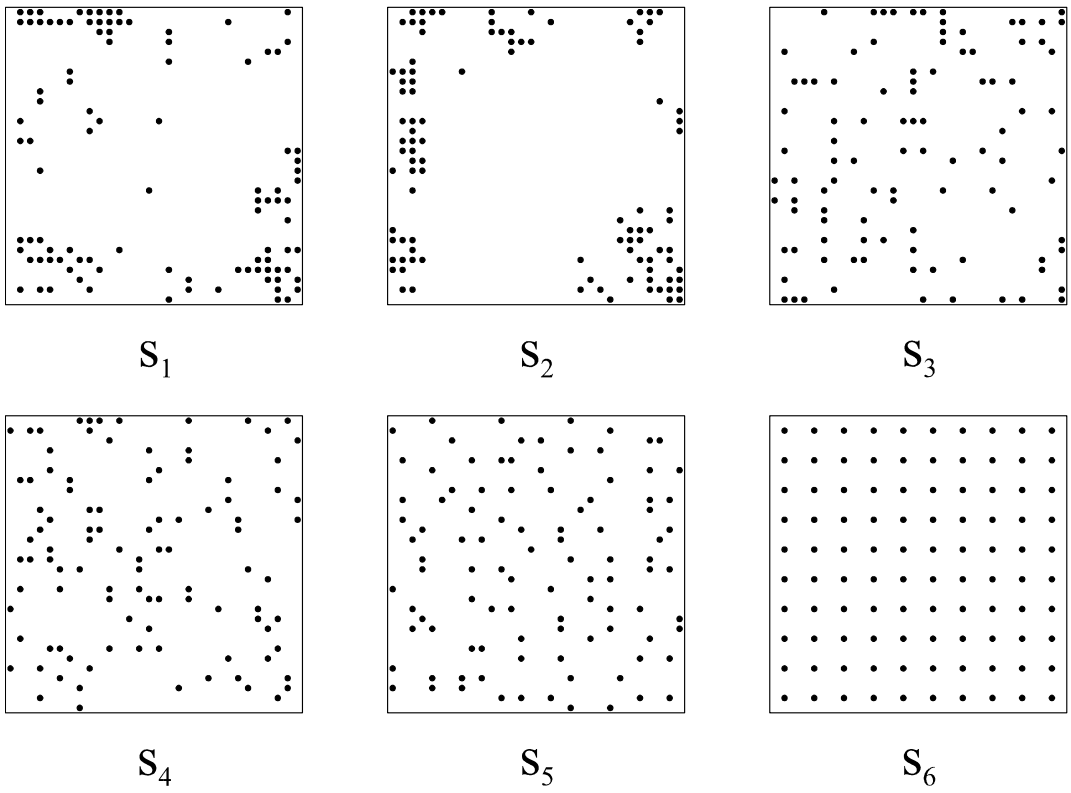
Nous considérons une population spatiale $\mathcal{U}$ constituée par les noeuds d'une grille 30×30 centrée dans un domaine carré D de 30 unités de côté. En premier lieu, nous examinons le résultat de l'optimisation de trois fonctions-objectif de la forme (8.20), correspondant :

1. à la fonction (8.24) avec $y_a = 0$, $y_d = 0$, $a = b = 0$ et $c = d = L_{\max}/2$, *i.e.* la loi uniforme ($N(h) = cste$),
2. à la fonction (8.24) avec $y_a = \frac{4}{3(d-a)}$, $y_d = \frac{1}{2} y_a$, $a = b = c = 0$ et $d = L_{\max}/2$, qui favorise les classes à faible distance,
3. à la fonction $f_D(h)$, *i.e.* la densité de probabilité des distances dans D.

La constante $\alpha = 150$ permet une mise à l'échelle des deux termes de (8.20). La meilleure minimisation de $J_\gamma(\cdot)$ est obtenue par SA pour les fonctions 1 et 3, et par optimisation locale pour la fonction 2. Les semis obtenus avec les fonctions 1 et 2 qui suivent les recommandations de Warrick & Myers (1987) s'avèrent d'emblée assez défavorables pour estimer $\gamma_D(\cdot)$ parce que les points ne se répartissent pas dans le domaine échantillonné mais ont tendance à se restreindre aux bords de D (Fig. 8.5, S_1 & S_2).

Sans doute à cause de ce mauvais comportement de leur fonction-objectif, Warrick & Myers (1987) suggèrent de fixer la moitié des sites sur une grille et de placer l'autre moitié de façon à minimiser (8.19). Mais ces auteurs semblent surtout s'intéresser au problème de la minimisation de (8.19), et n'en démontrent pas l'intérêt pratique. Comme il nous semble *a priori* impossible de justifier qu'une certaine fraction de l'échantillon suive un dispositif donné, et que l'autre fraction soit placée selon une fonction du type 1 voire 2, nous ne considérons aucune fonction-objectif utilisant (8.24).

En revanche, la minimisation de la fonction-objectif obtenue avec $f_D(h)$ conduit à un semis raisonnable (OPT), qui couvre bien l'ensemble de D (Fig. 8.5, S_3). Afin de juger l'intérêt de l'optimisation proposée, le motif OPT peut être comparé à des motifs classiques obtenus par EAS, STR ou ES (Fig. 8.5, S_4, S_5 & S_6).



Figure~8.5: Différents motifs d'échantillonnage d'une grille 30×30. S_1, S_2, S_3 : Motifs obtenus par optimisation des fonctions 1, 2 et 3. S_4, S_5, S_6 : Motifs produits par EAS, STR et ES.

D'un point de vue pratique, l'estimation du variogramme s'effectue pour un ensemble de classes de distances et se poursuit par l'ajustement d'un modèle. Aussi, dans l'évaluation de l'optimisation de l'échantillonnage, nous proposons d'intégrer à la fois :

- l'estimation de $\widehat{\gamma}(\cdot)$ au sens d'une certaine définition des classes de distances (pas et tolérance),
- l'ajustement automatique de $\widehat{\gamma}(\cdot)$ par un modèle $\widetilde{\gamma}(\cdot)$ au moyen de la méthode de Levenberg-Marquardt, à partir d'un paramétrage par défaut (Section 7.2.3.3).

Afin que les résultats de l'étude soient de portée suffisamment générale, nous considérons en outre :

- trois types de modèles d'autocorrélation spatiale très différents (exponentiel, gaussien et périodique) de pépite $c_0 = 1$ et de seuil $c_0 + c = 8000$,
- quatre portées pour chaque type de modèle (5, 10, 15 et 20),
- N_p populations pour chaque type de modèle et chaque portée.

L'étude de Monte-Carlo que nous proposons repose sur la simulation non conditionnelle d'un modèle de FAST-2, dans des conditions d'isotropie.

Les performances d'un ensemble $\mathcal{M}$ de quatre motifs d'échantillonnage m sont comparées :

- OPT obtenu par minimisation de $J_\gamma(\cdot)$,
- EAS obtenu par échantillonnage aléatoire simple,
- ES correspondant à une grille 10×10 centrée dans D,
- STR obtenu par échantillonnage aléatoire stratifié par une grille 10×10, à un élément par maille.

Chaque itération de l'étude de Monte-Carlo est constituée par la séquence des traitements suivants :

1. Simulation non conditionnelle d'une réalisation $z(\cdot)$ sur la grille 30×30.
2. Calcul du variogramme local $\gamma_D(\cdot)$ sur la totalité des 900 points, pour 19 classes $(1.0, 0.5)$.
3. Echantillonnage de $z(\cdot)$ selon OPT, EAS, ES et STR.
4. Calcul du variogramme expérimental $\widehat{\gamma}(\cdot)$ pour les quatre échantillons, avec la même définition des classes que pour le calcul de $\gamma_D(\cdot)$.
5. Paramétrage par défaut et ajustement automatique d'un modèle $\widetilde{\gamma}(\cdot)$ à chaque variogramme expérimental $\widehat{\gamma}(\cdot)$. Si l'ajustement automatique échoue, retourner en 1.
6. Evaluation de la qualité de l'ajustement de chaque modèle $\widetilde{\gamma}(\cdot)$ au variogramme local $\gamma_D(\cdot)$.
7. Retour à l'étape 1 jusqu'à obtenir N_p évaluations.

La définition des classes de distances pour le calcul de $\gamma_D(\cdot)$ coïncide avec la résolution de la grille définissant la population (grille 30×30 de maille $\Delta = 1$). En ce qui concerne la comparaison entre les motifs d'échantillonnage, il faut veiller à ne pas introduire un biais en favorisant un motif particulier. Par exemple, calculer le variogramme expérimental pour 6 classes $(3.0, 1.5)$ pourrait favoriser le motif ES (grille 10×10 de maille $\Delta = 3$). Aussi, tous les variogrammes $\widehat{\gamma}(\cdot)$ sont calculés avec la même définition des classes de distances que pour le calcul de $\gamma_D(\cdot)$.

Pour un échantillon s obtenu selon un certain motif d'échantillonnage m, la qualité de l'ajustement du modèle obtenu $\widetilde{\gamma}(\cdot)$ au variogramme local $\gamma_D(\cdot)$ est mesurée au moyen de l'erreur quadratique moyenne :

$$\mathrm{MSE}(m) = \frac{1}{k} \sum_{i=1}^{k} \left\{ \gamma_D(c_i) - \widetilde{\gamma}(c_i) \right\}^2 \tag{8.29}$$

Nous définissons le *score* d'un motif d'échantillonnage m_0 comme le pourcentage du nombre de fois N_0 où l'ajustement de son modèle $\widetilde{\gamma}(\cdot)$ à $\gamma_D(\cdot)$ s'avère être le meilleur, soit formellement $N_0/N_p \times 100$ avec :

$$N_0 = \sum_{i=1}^{N_p} \delta \left(\mathrm{MSE}(m_0), \min_{m \in \mathcal{M}} \mathrm{MSE}(m) \right) \tag{8.30}$$

et

$$\delta(\alpha, \beta) = \begin{cases} 1 & \text{si } \alpha = \beta \\ 0 & \text{sinon} \end{cases} \tag{8.31}$$

Enfin, pour l'ensemble des évaluations $\mathcal{B}$ où m_0 s'avère être le meilleur motif dans $\mathcal{M}$, nous définissons les *performances relatives moyennes* (PRM) de chacun des motifs concurrents $m \in \mathcal{M} - m_0$ comme :

$$\text{PRM}(m) = \frac{1}{N_0} \sum_{\mathcal{B}} \text{MSE}(m) \ / \ \text{MSE}(m_0) \tag{8.32}$$

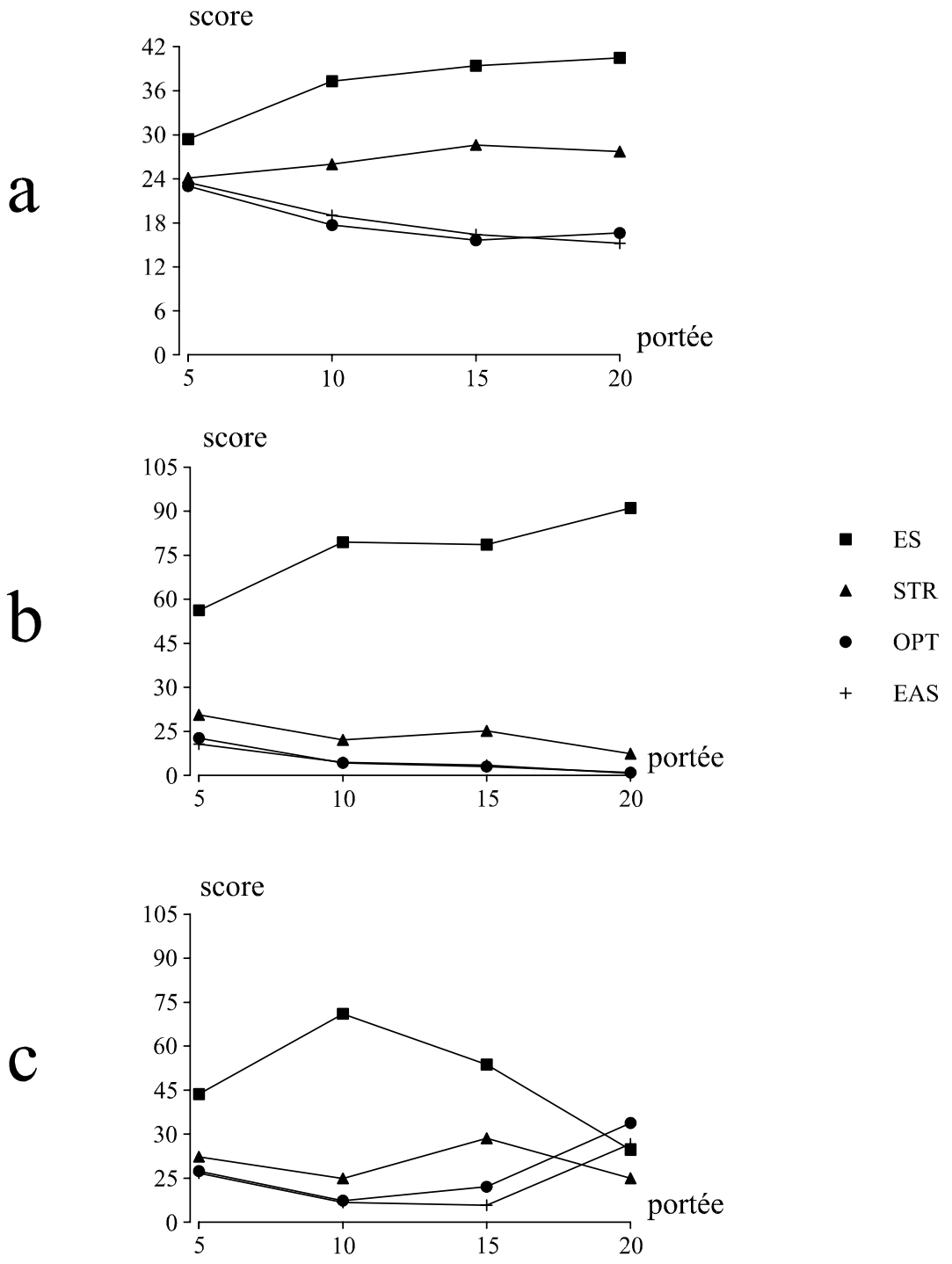
Les scores obtenus à l'issue de l'étude de Monte-Carlo montrent que, quels que soient le modèle et la portée, aucun motif ne s'avère systématiquement meilleur qu'un autre (Fig. 8.6). Cependant, ES est souvent le meilleur motif tandis que OPT et EAS sont moins fréquemment les meilleurs, STR se situant dans une position intermédiaire (Fig. 8.6). La supériorité de ES est plus fréquente lorsque la VR est spatialement très régulière (modèle gaussien), que lorsqu'elle l'est peu (modèle exponentiel), et d'autant plus que la portée est grande (Fig. 8.6.a & 8.6.b). En revanche, dans le cas du modèle périodique et de la portée $a = 20$, ES est moins fréquemment le meilleur motif que ne l'est OPT (Fig. 8.6.c).

Ces premiers résultats traduisent bien la difficulté d'une optimisation *a priori* de l'échantillonnage spatial pour l'estimation du variogramme. L'examen des PRM permet une analyse plus approfondie. En effet, la mesure $\text{PRM}(m)$ permet de savoir à quel point le meilleur motif m_0 s'avère supérieur à ses concurrents, vis-à-vis du critère (8.29). Examinons successivement chaque modèle d'autocorrélation.

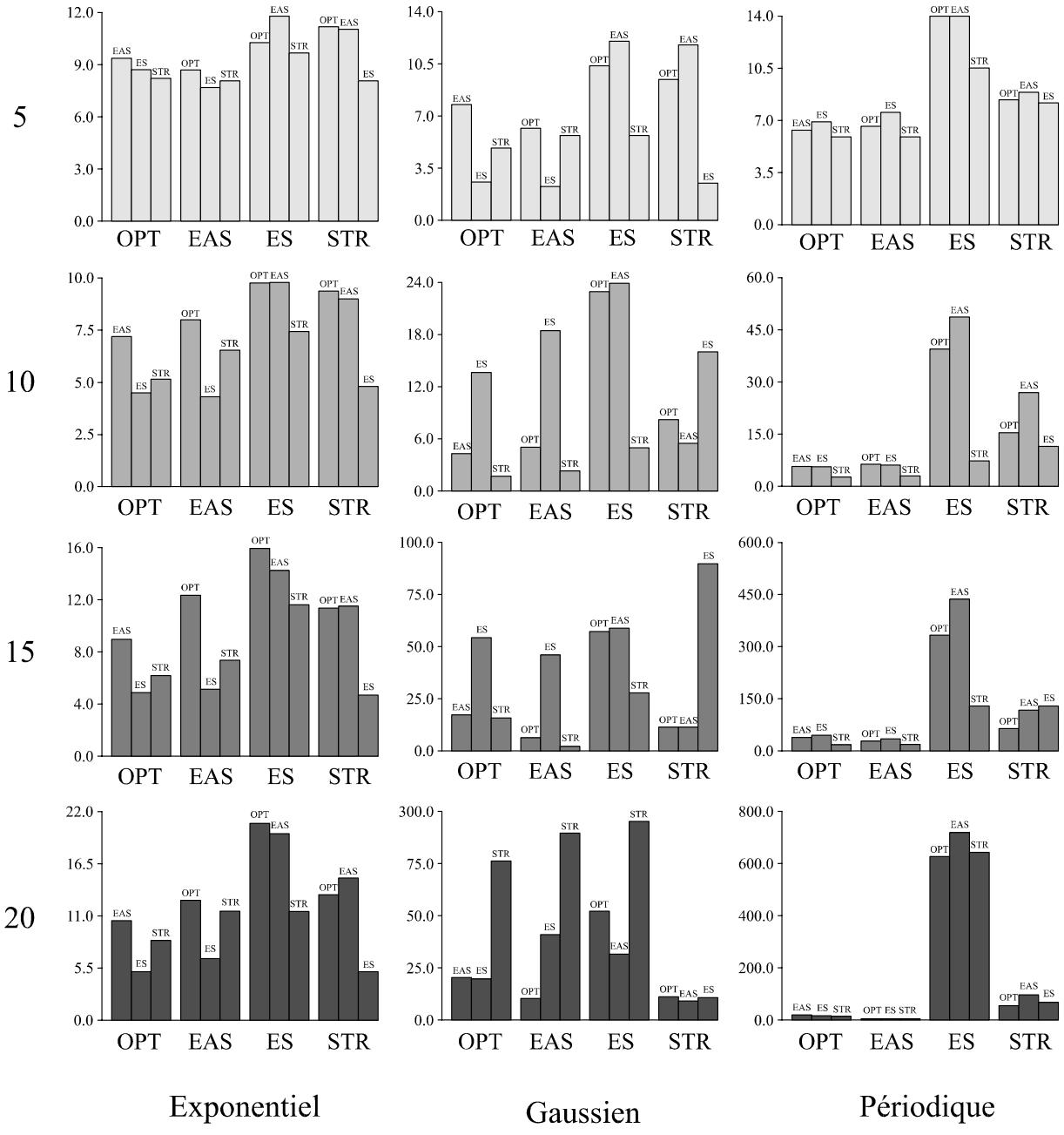
8.5.3.1 Modèle exponentiel

Les profils des PRM sont similaires quelle que soit la portée (Fig. 8.7). Cependant, lorsque la portée augmente, les écarts entre les PRM ont également tendance à augmenter. Ceci montre que lorsque la portée est faible vis-à-vis du domaine, les différents motifs se valent à peu près. L'interprétation de la Figure 8.7 peut se résumer ainsi :

- ES est plus souvent meilleur que ses concurrents.
- Lorsqu'il est le meilleur motif, ES est aussi le motif qui surclasse le plus ses concurrents.
- Lorsqu'il n'est pas le meilleur motif, ES est cependant le motif le moins surclassé par ses concurrents.
- EAS et OPT se comportent de façon similaire.
- STR se comporte de façon intermédiaire entre OPT / EAS et ES.



Figure~8.6: Scores des motifs OPT, EAS, STR et ES, en fonction de la portée du variogramme. (a) Modèle exponentiel. (b) Modèle gaussien. (c) Modèle périodique.



Figure~8.7: Performances relatives moyennes des motifs OPT, EAS, STR et ES, selon la portée du variogramme (5, 10, 15 et 20), et selon le modèle (exponentiel, gaussien et périodique). Détails dans le texte.

8.5.3.2 Modèle gaussien

Contrairement au modèle exponentiel, les profils des performances relatives diffèrent selon la portée (Fig. 8.7). L'interprétation de la Figure 8.7 peut se résumer ainsi :

- ES est très souvent le meilleur motif, et d'autant plus souvent que la portée augmente.
- Pour $a = 5$, les performances relatives permettent de classer les motifs du pire au meilleur selon EAS < OPT < STR < ES.
- Pour $a = 10$ et $a = 15$, OPT et EAS présentent des performances relatives similaires, tandis que STR présente un bon comportement. Lorsqu'il n'est pas le meilleur, ES est le pire des motifs.
- Pour $a = 20$, il est rare que ES ne soit pas le meilleur, et STR montre les plus mauvaises performances relatives.

8.5.3.3 Modèle périodique

Quelle que soit la portée, les PRM permettent de classer les motifs du pire au meilleur selon EAS < OPT < STR < ES (Fig. 8.7). Le motif ES s'avère souvent meilleur que les autres, et les surclasse d'autant plus que la portée augmente. Pour $a = 20$, ES n'est pas aussi fréquemment le meilleur motif que pour des portées inférieures, mais lorsqu'il est le meilleur, ES surclasse énormément ses concurrents.

8.5.3.4 Interprétation de l'optimisation

L'étude de Monte-Carlo montre que EAS est presque équivalent à OPT, ce qui s'explique par le fait qu'un échantillon aléatoire simple permet d'approximer assez correctement la densité de probabilité des distances dans D. L'échantillon stratifié STR présente des performances intermédiaires entre EAS et ES, ce qui s'explique par le fait qu'il s'agit d'un compromis entre l'échantillonnage aléatoire simple et l'échantillonnage systématique. Enfin, l'échantillon systématique selon une grille centrée ES apparaît souvent comme un bon motif.

Dans le cas d'une VR relativement homogène, notre étude de Monte-Carlo semble montrer qu'en absence de toute connaissance *a priori*, une bonne estimation de $\gamma_D(\cdot)$ requiert davantage une répartition des unités échantillonnées régulière plutôt qu'une bonne approximation de la densité de probabilité des distances dans D, ce qui milite en faveur de l'utilisation d'un échantillon systématique. Cependant, le problème essentiel posé par l'échantillonnage systématique est celui de la définition de la maille. Si la maille n'est pas suffisamment petite, il devient impossible de connaître le comportement à l'origine de $\gamma_D(\cdot)$, et difficile, voire même impossible, d'identifier la portée de l'autocorrélation spatiale. Face au risque que présente l'utilisation d'un motif d'échantillonnage sytématique inadapté à la structure spatiale de la VR, il conviendrait de lui préférer l'échantillonnage stratifié à un élément par strate, qui, comme l'échantillonnage systématique, présente l'avantage de répartir régulièrement les unités d'échantillonnage dans D, tout en diversifiant les distances inter-unités.

Chapitre 9

Test de la corrélation

> *"The simplest way to compare two maps* [...] *consists of computing the correlation coefficient with no consideration of the sample locations at all"* (Davis 1986)
>
> *"This seems to demonstrate conclusively the dangers of inferring from the product moment correlation coefficient when the observations may be spatially autocorrelated."* (Bivand 1980)
>
> *"If significance tests are to be useful, then they should have validity independent of the values of identifiable nuisance factors"* (Hinkley 1987)

La mesure de l'association (au sens large) entre deux variables régionalisées $x(\cdot)$ et $y(\cdot)$ constitue un traitement essentiel dans le cadre de l'écologie statistique. En effet, il ne suffit pas de savoir échantillonner, analyser et estimer les VR d'intérêt écologique indépendamment les unes des autres, il faut également pouvoir étudier leurs relations mutuelles. La façon la plus simple d'étudier l'association entre deux VR consiste à :

- quantifier l'association globale entre $x(\cdot)$ et $y(\cdot)$ au moyen d'une statistique classique,
- tester le caractère statistiquement significatif de la valeur observée de cette statistique.

Cette approche présente cependant deux inconvénients majeurs. En premier lieu, l'utilisation d'une statistique classique revient à négliger la nature régionalisée de $x(\cdot)$ et de $y(\cdot)$, ce qui représente une perte d'information préjudiciable à l'analyse : cet aspect sera traité spécifiquement dans le Chapitre 10. En second lieu, dans le cadre de la statistique classique, l'autocorrélation spatiale peut être vue comme un *paramètre de nuisance*, *i.e.* un paramètre dont on doit tenir compte pour la validité de la procédure statistique, mais qui n'est pas de premier intérêt pour le scientifique (Lindsay 1985). L'objet de ce chapitre est précisément d'étudier l'impact de l'autocorrélation spatiale dans le test statistique de l'association entre deux VR.

Nous considérons que les deux VR prennent leurs valeurs sur un ensemble de supports $s = \{s_i \mid i = 1, \ldots, n\}$, ce qui conduit aux données $\{(x(s_i), y(s_i)) \mid i = 1, \ldots, n\}$. La corrélation étant définie dans un contexte statistique classique, autrement dit, *a-spatial*,

nous modélisons les données comme des réalisations de deux variables aléatoires X et Y et nous les écrivons plus simplement $\{(x_i, y_i) \mid i = 1, \ldots, n\}$.

Il existe différentes statistiques mesurant la corrélation entre deux séries de données selon que la structure algébrique des variables est quantitative, ordinale, nominale ou binaire. Cependant, quelle que soit la statistique utilisée, les tests de la corrélation supposent l'indépendance des données $\{(x_i, y_i) \mid i = 1, \ldots, n\}$, ce qui revient à considérer que les couples (x_i, y_i) sont des réalisations indépendantes du couple de VA (X, Y), *i.e.* que les valeurs sont obtenues par échantillonnage aléatoire de la distribution bivariée de (X, Y). Lorsque les deux VR $x(\cdot)$ et $y(\cdot)$ présentent une structure d'autocorrélation spatiale, les données sont généralement elles-mêmes autocorrélées et l'hypothèse fondamentale d'indépendance n'est pas respectée. En conséquence, les données spatialement dépendantes ne peuvent pas être vues comme des réalisations indépendantes des VA X et Y. En général, les variables sont autocorrélées positivement, ce qui conduit à sous-estimer la *p-value* associée au test de la corrélation. Ainsi, dans la plupart des cas, l'application des procédures classiques pour tester la corrélation peut conduire à un résultat faussement significatif désigné sous le terme de *spurious correlation* ou *non-sense correlation* selon les auteurs (revue dans Aldrich 1995).

Ce problème est connu de longue date dans le cas des séries temporelles ou spatiales (*e.g.*, Student 1914). Pour y remédier, deux types d'approches sont généralement envisagés selon que l'on agit directement sur les données, ou que l'on change le mode de calcul de la *p-value*. Dutilleul (1993) distingue les méthodes qui consistent :

1. à filtrer la structure spatiale des données pour se ramener à des données spatialement indépendantes,

2. à développer des tests de permutation qui ne détruisent pas les structures spatiales, ou à modifier les tests paramétriques classiques afin de tenir compte de l'autocorrélation spatiale.

Dans le cadre de la seconde approche, il convient également d'ajouter les tests de Monte-Carlo utilisant un modèle statistique spatial pour une des deux VR. Mais l'approche primordiale consiste à aborder le problème de l'autocorrélation spatiale dès l'échantillonnage afin d'obtenir des données spatialement indépendantes. Dans ce qui suit, nous examinons la faisabilité des différentes approches dans le cas de la corrélation entre deux VR quantitatives, ainsi que la modification du test paramétrique dans le cas de l'association entre deux VR binaires.

9.1 Corrélation de Pearson

Nous supposons tout d'abord que l'hypothèse d'indépendance est satisfaite, *i.e.* que les VR $x(\cdot)$ et $y(\cdot)$ ne présentent pas de structure d'autocorrélation spatiale. La corrélation ρ entre les deux VA X et Y :

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} \tag{9.1}$$

est estimée par la corrélation d'échantillon :

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})}{\left[\sum_{i=1}^{n} (x_i - \bar{x}) \sum_{i=1}^{n} (y_i - \bar{y})\right]^{1/2}} \tag{9.2}$$

Lorsque l'échantillon $\{(x_i, y_i) \mid i = 1, \ldots, n\}$ est un échantillon aléatoire provenant d'une distribution bivariée normale, le test paramétrique de l'hypothèse nulle $\rho = 0$ peut s'effectuer selon deux procédures (Rodriguez 1982, Sokal & Rohlf 1995) :

- en normalisant la distribution de r grâce à la transformation de Fisher $z = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right)$, et en standardisant z par sa variance $\sigma_z^2 = \frac{1}{n-3}$ (pour $n \geq 50$),
- en standardisant r par sa variance sous H_0 estimée par $s_r = \left(\frac{1-r^2}{n-2}\right)^{1/2}$, et en testant r/s_r comme un t de Student à $n-2$ degrés de liberté.

Le test de la corrélation peut également s'effectuer conditionnellement aux données, *i.e.* sans faire référence à une distribution bivariée normale, ni à la notion d'échantillonnage aléatoire. La distribution de r sous H_0 est alors calculée en considérant les $n!$ échantillons bivariés obtenus en permutant les valeurs d'une des deux variables. Le test de permutation est envisageable pour $n \leq 10$. Pour $n > 10$ il faut avoir recours à un test de randomisation utilisant de nombreuses permutations aléatoires (*e.g.*, 10^5 permutations pour $n = 100$). Enfin, le test de Monte-Carlo consiste à calculer la distribution empirique de r sous H_0 en générant — au hasard et indépendamment les uns des autres — un grand nombre d'échantillons à partir d'un modèle de distribution pour une des deux variables. Le test de Monte-Carlo n'est pas nécessairement paramétrique puisqu'il peut s'effectuer directement à partir de la fonction de répartition empirique, *i.e.* sans modèle statistique (*e.g.*, loi de Gauss), et par conséquent sans avoir à estimer les paramètres d'un modèle (*cf.* Annexe E).

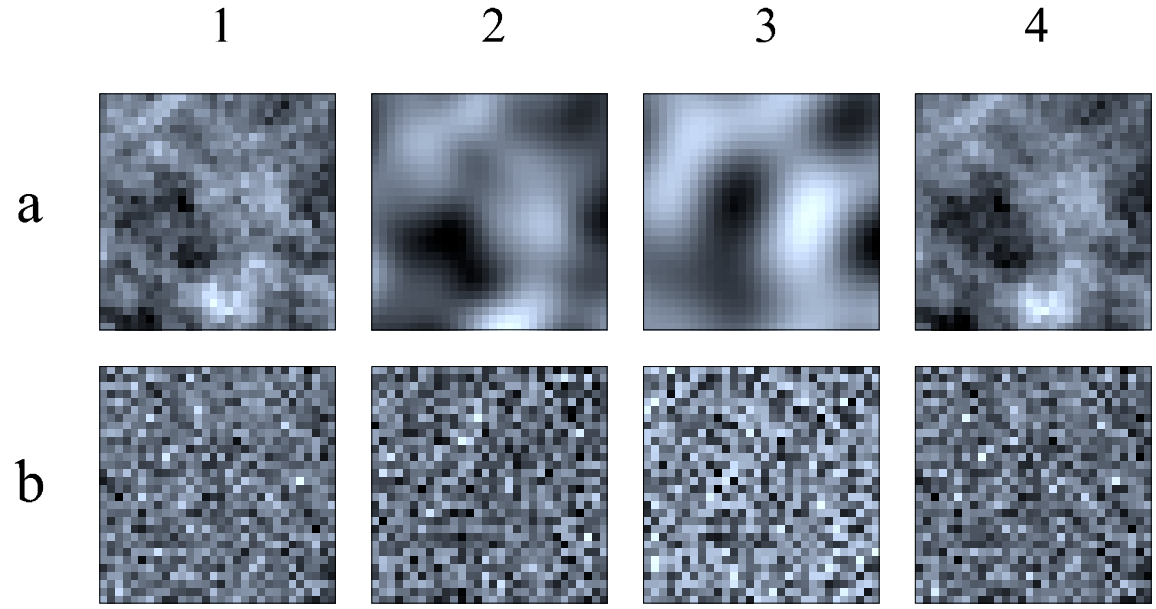
9.1.1 Etude de Monte-Carlo

Dans une étude de Monte-Carlo utilisant des modèles autorégressifs, Bivand (1980) explore les conséquences de la violation de l'hypothèse d'indépendance dans le test de la corrélation calculée sur des données spatialement autocorrélées. L'auteur conclut qu'en cas d'autocorrélation positive, l'utilisation du test paramétrique classique conduit à une *p-value* biaisée parce que :

- le degré de liberté utilisé est trop élevé,
- la variance de r est sous-estimée.

Bivand (1980) note également que le biais de la variance de r varie avec la répartition des supports et se révèle particulièrement élevé dans le cas d'une répartition irrégulière. Cependant, cet effet est certainement dû à la façon dont l'auteur a mené son étude de Monte-Carlo.

Afin d'illustrer l'erreur commise en testant la corrélation comme si les données étaient spatialement indépendantes, nous avons également procédé à une étude de Monte-Carlo, mais en utilisant des modèles de fonctions aléatoires plutôt que des modèles autorégressifs. La simulation gaussienne des modèles géostatistiques est effectuée par la méthode utilisant la décomposition de Cholesky de la matrice de covariance (Section 4.3.3.2). Quatre modèles de variogrammes théoriques sont simulés sur une grille 30×30 de pas $\Delta = 1$: exponentiel, gaussien, périodique et sphérique. Le paramétrage de tous les modèles est arbitrairement fixé à $\theta = (1, 7999, 10)$. Une première réalisation de chaque modèle est simulée en fixant la même valeur à la graine du générateur de nombres pseudo-aléatoires (Annexe B), de sorte que toutes les réalisations sont corrélées positivement deux à deux : ces réalisations jouent le rôle des populations finies sous étude (Fig. 9.1).



Figure˜9.1: Populations de référence pour l'étude de Monte-Carlo. (a) Populations simulées d'après des modèles de variogrammes : (1) exponentiel, (2) gaussien, (3) périodique, (4) sphérique. (b) Populations obtenues par randomisation des valeurs des populations simulées (1) à (4).

Les données sont obtenues par échantillonnage systématique selon une grille 10×10 centrée dans la grille 30×30. Pour les populations simulées d'après les modèles exponentiel, gaussien, périodique et sphérique, les échantillons systématiques sont désignés respectivement par es_1, es_2, es_3 et es_4. La corrélation observée est celle qui est calculée pour chaque couple d'échantillons systématiques. La *p-value* associée au test unilatéral de la corrélation observée est estimée en simulant 10^5 réalisations du modèle d'une des deux variables. Une légère différence est attendue selon que la distribution de r sous H_0 est obtenue en simulant le modèle spatial de $x(\cdot)$ ou celui de $y(\cdot)$. Afin d'apprécier cette différence, les simulations sont effectuées pour les deux variables. Par convention, la VR dont on simule le modèle est désignée par $y(\cdot)$. De nouvelles populations sont produites en randomisant en parallèle 10^6 fois les populations autocorrélées afin de détruire l'autocorrélation spatiale tout en maintenant intacte la distribution statistique des valeurs (Fig. 9.1). Les données sont obtenues par un échantillonnage identique au précédent et les échantillons sont désignés par Res_1, Res_2, Res_3 et Res_4.

Les échantillons sans autocorrélation sont croisés avec les échantillons autocorrélés, ce qui conduit à divers degrés de corrélation entre échantillons. La *p-value* est déterminée par simulation du modèle de la VR spatialement autocorrélée, désignée par $y(\cdot)$. Le générateur de nombres pseudo-aléatoires étant toujours initialisé avec la même graine, la première réalisation redonne la population sous étude, et la *p-value* est calculée comme :

$$p = \frac{\text{Card}\left(\{r \mid r \in \Omega,\ r \geq r_{obs}\}\right)}{\text{Card}\left(\Omega\right)} \tag{9.3}$$

et pour la corrélation négative comme :

$$p = \frac{\text{Card}\left(\{r \mid r \in \Omega,\ r \leq r_{obs}\}\right)}{\text{Card}\left(\Omega\right)} \tag{9.4}$$

avec Ω l'ensemble des valeurs de r et $\text{Card}\left(\Omega\right) = 10^5$. La *p-value* de référence (p_0) peut être comparée à celles obtenues par le test paramétrique de Student (p_1), le test de Monte-Carlo non spatial (p_2) et le test de randomisation (p_3). Les simulations de Monte-Carlo et les randomisations sont répétées 10^5 fois et concernent la variable Y. La distribution de r sous H_0 est résumée par la moyenne $\bar{r}$ et l'écart-type σ_r. L'écart-type peut être confronté à son estimation classique par $s_r = \left[\left(1 - r^2\right) / \left(n - 2\right)\right]^{1/2}$.

Lorsque les deux VR $x(\cdot)$ et $y(\cdot)$ sont spatialement autocorrélées, l'estimateur classique s_r sous-estime systématiquement l'écart-type de la distribution de r sous H_0 (Tab. 9.1). En outre, les *p-values* calculées par les procédures paramétrique (p_1) et non paramétriques (p_2 et p_3) sous-estiment la *p-value* de référence (p_0) sauf dans le cas des couples (es_1,es_4) et (es_2,es_3) pour lesquels la corrélation est élevée. Dans cette étude, la sous-estimation des *p-values* varie d'un facteur 10 à un facteur 10^3. En revanche, lorsqu'une seule des deux variables présente de l'autocorrélation spatiale, les *p-values* données par les tests se révèlent similaires aux *p-values* de référence (Tab. 9.2).

Ces résultats confirment qu'il est nécessaire de tenir compte de la dépendance spatiale uniquement lorsque les deux VR sont autocorrélées (Cliff & Ord 1981, Dutilleul 1993). Dans ce cas, il peut être envisagé d'éviter l'autocorrélation spatiale dès la collecte des données, et lorsque ce n'est pas possible, de filtrer la structure spatiale ou de modifier les procédures de test. Les *p-values* p_0 données dans le Tableau 9.1 peuvent servir de référence pour évaluer les solutions envisagées dans les sections suivantes.

9.1.2 Echantillonnage et sous-échantillonnage

Il est possible d'envisager de régler le problème posé par le test de la corrélation entre deux VR $x(\cdot)$ et $y(\cdot)$ autocorrélées dès l'échantillonnage en faisant en sorte que les données $\{(x_i, y_i) \mid i = 1, \ldots, n\}$ soient spatialement indépendantes. On sait bien qu'un échantillonnage aléatoire simple ou systématique ne garantit pas l'indépendance spatiale des données (Legendre & Fortin 1989, Fortin *et al.* 1989). Assurer l'indépendance spatiale dès l'échantillonnage nécessite en effet une connaissance *a priori* de la forme de la structure d'autocorrélation spatiale des deux VR $x(\cdot)$ et $y(\cdot)$.

$x(\cdot)$	$y(\cdot)$	σ_r	s_r	r_{obs}	p_0	p_1	p_2	p_3
es_2	es_1	0.16494	0.09358	0.37663	0.00939	0.00006	0.00005	0.00008
es_1	es_2	0.15355	0.09358	0.37663	0.00456	0.00006	0.00005	0.00005
es_3	es_1	0.16151	0.09698	0.27965	0.04195	0.00242	0.00246	0.00379
es_1	es_3	0.16015	0.09698	0.27965	0.03996	0.00242	0.00236	0.00301
es_4	es_1	0.14546	0.03373	0.94259	0.00001	0.00000	0.00001	0.00001
es_1	es_4	0.14009	0.03373	0.94259	0.00001	0.00000	0.00001	0.00001
es_3	es_2	0.21334	0.06848	0.73513	0.00002	0.00000	0.00001	0.00002
es_2	es_3	0.20415	0.06848	0.73513	0.00001	0.00000	0.00001	0.00002
es_4	es_2	0.18650	0.08003	0.61017	0.00005	0.00000	0.00001	0.00001
es_2	es_4	0.18955	0.08003	0.61017	0.00014	0.00000	0.00001	0.00001
es_4	es_3	0.19134	0.09196	0.41381	0.01252	0.00001	0.00001	0.00007
es_3	es_4	0.18885	0.09196	0.41381	0.01247	0.00001	0.00003	0.00006

Tableau 9.1: Résultats de l'étude de Monte-Carlo pour les couples d'échantillons formés par es_1, es_2, es_3 et es_4. σ_r : écart-type de la distribution de r sous H_0. s_r : estimateur classique de σ_r. r_{obs} : valeur observée de r. p_0 : *p-value* de référence. p_1 : *p-value* du test de Student. p_2 : *p-value* du test de Monte-Carlo non spatial. p_3 : *p-value* du test de randomisation. Les moyennes $\bar{r}$ sont comprises entre -1.49×10^{-3} et 6.8×10^{-4}.

$x(\cdot)$	$y(\cdot)$	σ_r	s_r	r_{obs}	p_0	p_1	p_2	p_3
Res_2	es_1	0.09301	0.10073	-0.07564	0.20986	0.22723	0.22568	0.23049
Res_3	es_1	0.09660	0.09975	-0.15787	0.05188	0.05836	0.05837	0.05493
Res_4	es_1	0.09090	0.10102	-0.00041	0.49895	0.49839	0.49856	0.48526
Res_1	es_2	0.08990	0.10056	0.09488	0.14893	0.17387	0.17277	0.18155
Res_3	es_2	0.09416	0.10008	0.13582	0.07639	0.08894	0.08937	0.09816
Res_4	es_2	0.08059	0.10061	0.08978	0.13394	0.18720	0.18658	0.19687
Res_1	es_3	0.08988	0.09995	0.14477	0.05451	0.07535	0.07520	0.07974
Res_2	es_3	0.07869	0.10099	0.02116	0.39728	0.41725	0.41972	0.41564
Res_4	es_3	0.08392	0.09990	0.14804	0.03807	0.07079	0.07030	0.07247
Res_1	es_4	0.09288	0.10083	0.06075	0.25843	0.27413	0.27471	0.28723
Res_2	es_4	0.08985	0.10096	-0.03374	0.35700	0.36948	0.36681	0.37083
Res_3	es_4	0.09594	0.10079	-0.06736	0.24570	0.25275	0.25089	0.24108

Tableau 9.2: Résultats de l'étude de Monte-Carlo pour les couples formés par les échantillons Res_1, Res_2, Res_3 et Res_4 avec les échantillons es_1, es_2, es_3 et es_4. σ_r : écart-type de la distribution de r sous H_0. s_r : estimateur classique de σ_r. r_{obs} : valeur observée de r. p_0 : *p-value* de référence. p_1 : *p-value* du test de Student. p_2 : *p-value* du test de Monte-Carlo non spatial. p_3 : *p-value* du test de randomisation. Les moyennes $\bar{r}$ sont comprises entre -1.22×10^{-3} et 8.6×10^{-4}.

Cependant, l'étude de Monte-Carlo (Section 9.1.1) a montré que les tests usuels de r sont valides dès que les données concernant l'une des deux VR ne présentent pas d'autocorrélation spatiale. En conséquence, nous considérons que l'échantillon spatial $s = \{s_i \mid i = 1, \ldots, n\}$ est défini en fonction de la structure d'autocorrélation spatiale d'une des deux VR, que nous désignons arbitrairement par $y(\cdot)$.

Supposons dans un premier temps que l'écologiste connaisse uniquement les caractéristiques essentielles du variogramme de $y(\cdot)$. Si le variogramme de $y(\cdot)$ est borné, *i.e.* si la VR correspond à un phénomène de transition (Section 4.2.2, p. 74), il est possible d'identifier la portée de l'autocorrélation spatiale a. Ainsi, il suffit d'échantillonner $y(\cdot)$ de sorte que la distance entre deux supports s_i et s_j soit contrainte par :

$$d(s_i, s_j) > a \quad \forall s_i \neq s_j \tag{9.5}$$

Si la structure d'autocorrélation de $y(\cdot)$ est isotrope, il suffit d'utiliser une grille de maille carrée de côté $\Delta > a$ pour garantir l'indépendance des données $\{y_i \mid i = 1, \ldots, n\}$. Dans le cas d'une anisotropie géométrique mettant en jeu uniquement deux directions perpendiculaires et des portées a_1 et a_2, il suffit d'utiliser une grille de maille rectangulaire de côtés $\Delta_1 > a_1$ et $\Delta_2 > a_2$. Une structure anisotropique complexe mettant en jeu m variogrammes directionnels définissant un ensemble de portées $A = \{a_i \mid i = 1, \ldots, m\}$ rend cependant ce type d'approche assez difficile à concevoir en pratique, à moins d'utiliser une grille de maille carrée de côté $\Delta > a_0$ avec :

$$a_0 = \max_{a \in A}(a) \tag{9.6}$$

Même dans une situation favorable d'isotropie, il existe plusieurs situations qui limitent le champ d'application de cette méthode. En effet, si la portée a est grande vis-à-vis du domaine à échantillonner D, il peut s'avérer impossible d'obtenir un échantillon qui respecte la contrainte (9.5) tout en étant de taille suffisante. Par ailleurs, si aucune des deux VR ne présente un variogramme borné, il est tout simplement impossible de définir une portée pour l'une d'entre elles. Enfin, cette approche ne peut s'appliquer qu'en amont de la collecte des données, ce qui ne règle évidemment pas le problème du test de la corrélation pour des données déjà acquises.

Toutefois, dans le cas d'un échantillonnage déjà effectué, il est envisageable de retirer des données jusqu'à ce que la dépendance spatiale soit fortement réduite, *i.e.* en procédant à un sous-échantillonnage, l'échantillon originel étant vu comme une population finie $\mathcal{U}$. Le problème consiste alors à identifier un échantillon $s \subset \mathcal{U}$ optimal vis-à-vis d'une fonction-objectif de la forme :

$$J(s) = \sum_{i=1}^{k} w_i \left| \widetilde{\gamma}_i - \widehat{\gamma}_i \right| \tag{9.7}$$

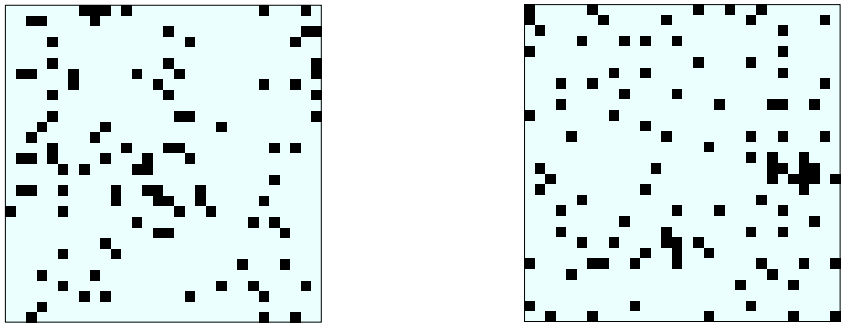
avec, pour la classe i, $\widehat{\gamma}_i$ la valeur du variogramme expérimental, $\widetilde{\gamma}_i$ la valeur du modèle de variogramme du type effet de pépite pur et w_i un pondérateur donnant davantage de poids aux premières classes parmi les k classes de distances (*e.g.*, $w_i = i^{-1}$). Le problème d'optimisation combinatoire associé à la minimisation de (9.7) peut être traité avec une heuristique telle que la recherche taboue (Section 8.2.6). Cette solution n'est généralement pas recommandée parce qu'elle entraîne une perte d'information. En outre, en considérant un échantillon originel de taille $n = 100$ comme dans la Section 9.1.1, le sous-échantillon

risque d'être de petite taille, ce qui se traduit par une perte de puissance statistique pour le test de la corrélation. En revanche, une situation toute différente est celle des données issues de la télédétection. En effet, ces données sont généralement très abondantes, de sorte qu'il est de toute façon nécessaire de procéder à un sous-échantillonnage avant de les analyser à l'aide d'un logiciel statistique standard. Dans ces conditions, il convient de procéder en trois étapes :

1. sous-échantillonnage représentatif (Section 5.3.3), par exemple par échantillonnage aléatoire simple, de façon à obtenir un premier échantillon s d'une taille compatible avec les moyens de calcul en vigueur (*e.g.*, $n = 1000$),
2. sous-échantillonnage de s minimisant la dépendance spatiale, conduisant à un échantillon $s^{'}$ de taille raisonnable (*e.g.*, $n^{'} = 500$),
3. test usuel de la corrélation en considérant les données pour les supports de $s^{'}$.

Ce type d'approche pose cependant un problème inférentiel fondamental au sens où l'échantillon minimisant la dépendance spatiale est, par définition, un échantillon qui n'est plus représentatif de la population dont il est extrait.

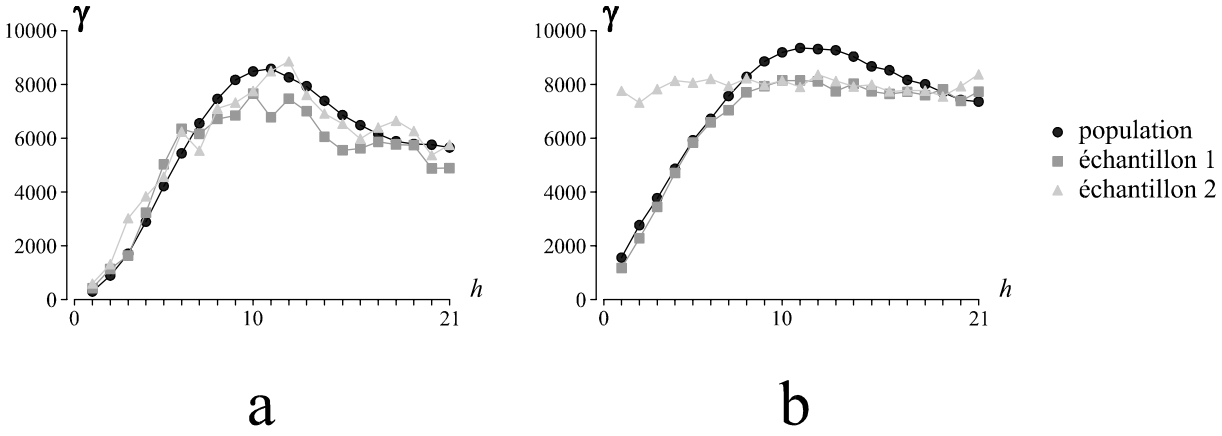
Considérons par exemple deux populations $x(\cdot)$ et $y(\cdot)$ obtenues par simulation sur une grille 30×30 de pas $\Delta = 1$ de deux FAST-2 de variogrammes périodique et sphérique (Section 9.1.1, Fig. 9.1.3a & 9.1.4a). Implémentons un motif d'échantillonnage comportant $n = 100$ supports (Fig. 9.2.a).



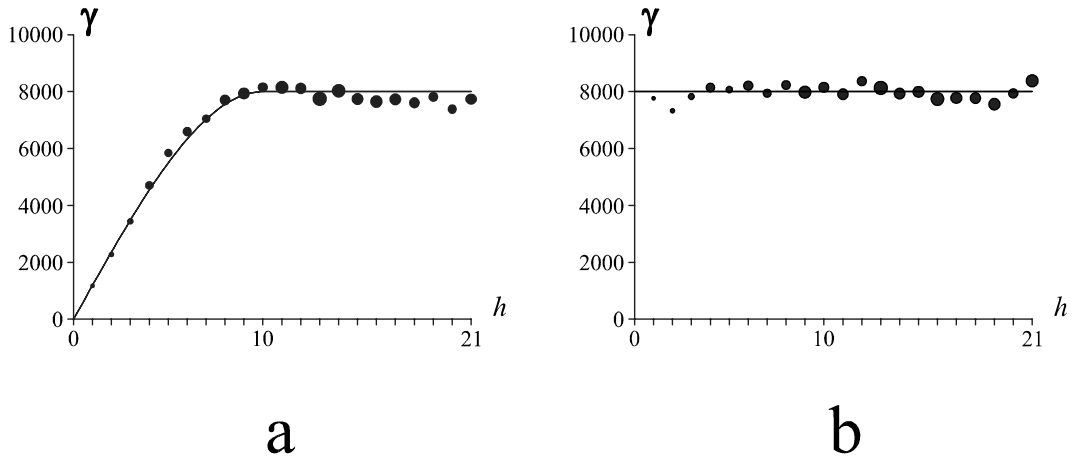
Figure˜9.2: Motifs d'échantillonnage des populations $x(\cdot)$ et $y(\cdot)$ définies sur une grille 30×30. (a) Motif d'échantillonnage conduisant aux échantillons $e_x^{(1)}$ et $e_y^{(1)}$. (b) Motif d'échantillonnage conduisant aux échantillons $e_x^{(2)}$ et $e_y^{(2)}$.

Les deux échantillons obtenus, $e_x^{(1)}$ et $e_y^{(1)}$, exhibent des variogrammes expérimentaux qui révèlent de fortes structures d'autocorrélation spatiale (Fig. 9.3). La valeur de r calculée pour les deux échantillons $e_x^{(1)}$ et $e_y^{(1)}$ est $r^{(1)} \simeq 0.43655$. La *p-value* associée à $r^{(1)}$ par un test de randomisation (10^5 permutations aléatoires) est $p_r^{(1)} \simeq 0.00006$, tandis que le test de Student donne $p_S^{(1)} \leq 0.00001$. Le variogramme expérimental de $e_y^{(1)}$ peut être correctement modélisé par le modèle $\mathrm{Sphe}(0, 8000, 10)$ (Fig. 9.4.a). En simulant 10^5

réalisations d'une FAST-2 ayant pour variogramme le modèle Sphe $(0, 8000, 10)$, il est possible de déterminer la *p-value* correcte associée à $r^{(1)}$, soit $p^{(1)} \simeq 0.01645$ (Section 9.1.1). L'ordre de grandeur de $p^{(1)}$ diffère d'un facteur 1000 de celui de $p_r^{(1)}$ ou $p_S^{(1)}$, ce qui illustre une nouvelle fois l'impact de l'autocorrélation spatiale sur le test de la corrélation.



Figure˜9.3: Variogrammes locaux et expérimentaux. (a) Variogrammes de la population $x(\cdot)$ et des échantillons $e_x^{(1)}$ et $e_x^{(2)}$. (b) Variogrammes de la population $y(\cdot)$ et des échantillons $e_y^{(1)}$ et $e_y^{(2)}$.



Figure˜9.4: Variogrammes expérimentaux et leurs modèles pour les échantillons extraits de la population $y(\cdot)$. (a) Modèle sphérique pour le variogramme de l'échantillon $e_y^{(1)}$. (b) Modèle d'effet de pépite pur pour le variogramme de l'échantillon $e_y^{(2)}$.

Implémentons un second motif d'échantillonnage de $n = 100$ supports (Fig. 9.2.b). Ce motif est conçu de façon à minimiser l'autocorrélation spatiale pour la variable Y tout en maintenant le même niveau de corrélation entre les nouveaux échantillons $e_x^{(2)}$ et $e_y^{(2)}$ qu'entre les échantillons $e_x^{(1)}$ et $e_y^{(1)}$. Effectivement, la valeur de r pour les échantillons $e_x^{(2)}$ et $e_y^{(2)}$ est $r^{(2)} \simeq 0.43696$, *i.e.* pratiquement identique à $r^{(1)} \simeq 0.43655$. Le variogramme expérimental de $e_y^{(2)}$ peut être modélisé par un effet de pépite pur situé à $c_0 = 8000$ (Fig. 9.4.b). A présent que les données concernant la variable Y sont spatialement indépendantes, la *p-value* associée à $r^{(2)}$ peut être calculée de façon valide par un

test de randomisation, d'où $p^{(2)} \simeq 0.00003$. Cette *p-value* est pratiquement identique à $p_r^{(1)}$ puisque $r^{(2)} \simeq r^{(1)}$ et $n = 100$ dans les deux cas. Il peut sembler paradoxal que $p^{(2)}$ constitue une *p-value* correcte alors que $p_r^{(1)} \simeq p^{(2)}$ est incorrecte. En fait, le variogramme de $e_y^{(2)}$ laisse supposer que la population $y(\cdot)$ n'est pas autocorrélée, et par conséquent, que le modèle de superpopulation qui convient est également sans autocorrélation spatiale. L'écart très important observé entre $p^{(1)}$ et $p^{(2)}$ ne fait que refléter la différence de modèle sous-jacent : spatialement autocorrélé dans le cas de $p^{(1)}$, et sans autocorrélation spatiale dans le cas de $p^{(2)}$. Le fait que le second modèle ne soit absolument pas conforme à la réalité résulte du manque de représentativité de l'échantillon $e_y^{(2)}$ vis-à-vis de la population $y(\cdot)$.

En effet, comme le montre la superposition des variogrammes locaux et expérimentaux, du point de vue de l'autocorrélation spatiale, les échantillons $e_x^{(1)}$ et $e_x^{(2)}$ sont assez représentatifs de la population $x(\cdot)$ (Fig. 9.3.a) alors que, par construction, l'échantillon $e_y^{(2)}$ est manifestement beaucoup moins représentatif de la population $y(\cdot)$ que ne l'est $e_y^{(1)}$ (Fig. 9.3.b). En conséquence, bien que la valeur observée de r soit identique dans les deux cas, et relativement proche de la corrélation entre les populations $x(\cdot)$ et $y(\cdot)$ ($r \simeq 0.41099$), la simulation du modèle de superpopulation de $y(\cdot)$ (10^5 réalisations) montre que la *p-value* associée à la corrélation entre populations est $p \simeq 0.01089$, ce qui est évidemment bien plus proche de $p^{(1)} \simeq 0.01645$ que de $p^{(2)} \simeq 0.00003$.

Finalement, nous pouvons conclure que l'échantillonnage en vue de minimiser la dépendance spatiale des données assure la validité des tests effectués **conditionnellement aux données**, mais ne permet en aucune façon d'inférer correctement la corrélation entre les populations d'origine.

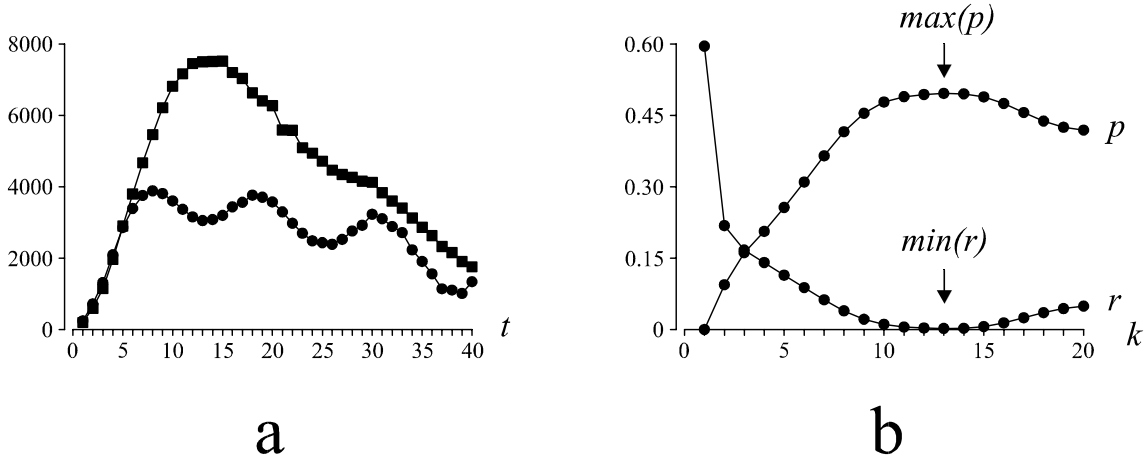
9.1.3 Filtrage des données

Le filtrage des données par différentiations successives représente la plus ancienne méthode pour résoudre le problème posé par l'autocorrélation dans le test de la corrélation entre deux variables (Student 1914). Student (1914) envisage aussi bien le cas des séries spatiales que celui des séries temporelles, mais ne considère en pratique que le domaine temporel. Cet auteur décompose les séries temporelles en une tendance polynomiale et un résidu indépendant du temps, et propose de filtrer la composante liée au temps en calculant la corrélation entre les différences entre valeurs successives à l'ordre $1, 2, \ldots, k$ de $x(\cdot)$ et les différences de mêmes ordres pour $y(\cdot)$. L'auteur considère que le filtrage est réalisé lorsque la corrélation ne varie plus d'une différentiation à l'autre.

Soit r_0 la corrélation entre les variables d'origine ; l'algorithme des différentiations successives peut s'écrire :

1. $k \leftarrow 1$.
2. Pour $i = 1$ à $n - 1$ faire $x_i \leftarrow x_i - x_{i+1}$ et $y_i \leftarrow y_i - y_{i+1}$.
3. $n \leftarrow n - 1$.
4. Calculer le coefficient de corrélation r_k pour les couples $\{(x_i, y_i) \mid i = 1, \ldots, n\}$.
5. Si $r_k \simeq r_{k-1}$ alors FIN sinon $k \leftarrow k + 1$, aller en 2.

Le critère d'arrêt tel qu'il est proposé et illustré dans Student (1914) n'est pas défini de façon très précise, aussi vaut-il mieux examiner la courbe des valeurs de r (ou des *p-values* associées aux valeurs de r), jusqu'à une valeur limite pour k. Si cette courbe présente un extremum encadré par deux valeurs de même ordre de grandeur, alors, d'après Student (1914), on peut considérer que la valeur extrême correspond à la corrélation corrigée des effets de la dépendance temporelle. Considérons par exemple les deux séries de la Figure 9.5. Faisons varier k entre 1 et 20 et calculons r et la *p-value* associée p. L'extremun local de la courbe de r ou de p en fonction de k est situé en $k = 13$ (Fig. 9.5).



Figure˜9.5: Illustration de la méthode des différentiations successives. (a) Deux séries temporelles fictives fortement autocorrélées. (b) Corrélation et *p-value* associée exprimées en fonction de l'ordre de différentiation k.

Dans le domaine temporel, la différentiation peut se calculer à tous les ordres entre valeurs successives le long de l'axe du temps. En revanche, dans le domaine spatial la différentiation d'ordre élevé n'est pas une opération bien définie puisqu'au contraire du temps, il n'existe pas d'ordre naturel (Haining 1991, Ripley 1988, p. 6). Plusieurs opérateurs de différentiation spatiale peuvent être construits, le plus communément utilisé étant exposé dans Cliff & Ord (1981, pp. 192-194).

D'autres approches de filtrage sont envisageables, notamment la soustraction d'une surface de tendance ajustée au données. Cette technique nécessite de choisir la forme de la tendance, ainsi qu'une méthode d'ajustement. Le résidu résultant de la soustraction de la tendance est censé ne plus présenter d'autocorrélation spatiale. Bien que ce type de filtrage soit souvent mentionné (Haining 1991, Legendre 1993, Dutilleul 1993), nous ne pouvons pas le recommander parce qu' il est très discutable d'interpréter la corrélation mesurée sur les résidus en termes de corrélation des variables d'origine (Aldrich 1995). Il est en effet difficile d'apprécier l'effet du filtrage — et plus généralement de toute transformation des données — sur le résultat du calcul de la corrélation à partir des données transformées (Rodriguez 1982).

Considérons par exemple deux réalisations de FAST-2 autocorrélées qui, par définition, présentent une dérive constante. Que peut-on bien filtrer comme tendance dans ce cas? Si la tendance ajustée correspond à la dérive, le résidu présente toujours une structure d'autocorrélation spatiale. Afin d'obtenir un résidu non autocorrélé, il serait certainement nécessaire d'ajuster une tendance de degré très élevé, mais dans ce cas,

quelle sera la signification du résidu? Le résidu ne représentera vraisemblablement plus qu'une fluctuation aléatoire sans structure spatiale, mais également sans aucune signification phénoménologique. Afin d'être en mesure de sélectionner le degré de la tendance de façon objective, il est possible d'envisager de recourir aux tests F de la régression multiple. Cependant, cette approche n'est pas valide car les données étant spatialement autocorrélées positivement, la *p-value* associée à chaque test F est sous-estimée, ce qui conduit finalement à ajuster des surfaces polynomiales de degrés trop élevés (Ripley 1981, p. 34).

Il semble donc préférable de ne pas altérer les données pour les faire rentrer de force dans un certain cadre d'hypothèses, mais plutôt de changer les procédures statistiques lorsqu'elles s'avèrent inadaptées aux données.

9.1.4 Test de Monte-Carlo spatial

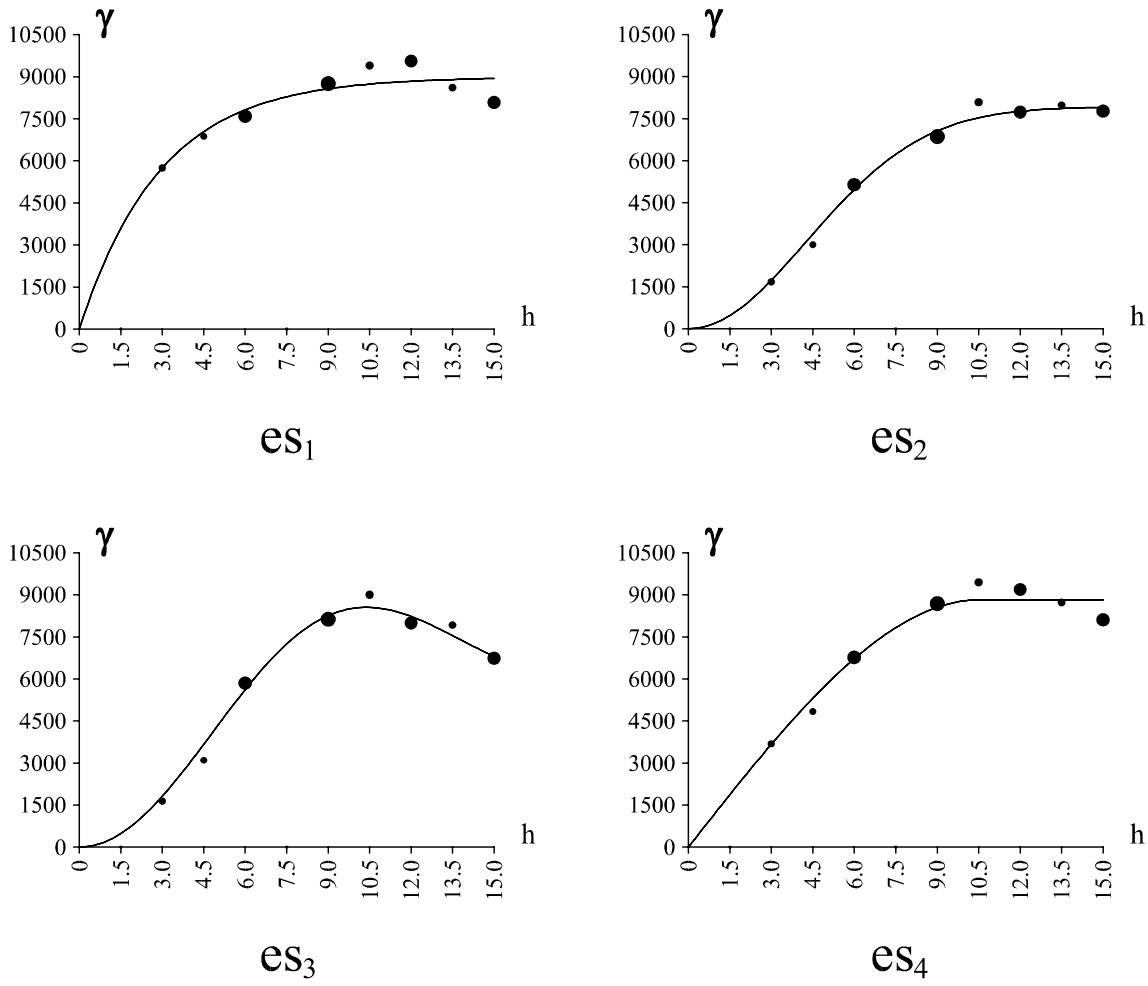
Il est possible d'estimer la *p-value* associée à la corrélation entre deux VR en réalisant un test de Monte-Carlo. Considérons qu'une des deux VR, par exemple $y(\cdot)$, est une réalisation d'une fonction aléatoire $Y(\cdot)$. Le test de Monte-Carlo spatial consiste à approximer la distribution de r sous H_0 en tirant au hasard un grand nombre de réalisations de $Y(\cdot)$. Pour chaque réalisation de $Y(\cdot)$ sur les supports $\{s_i \mid i = 1, \ldots, n\}$, la corrélation est calculée avec la VR $x(\cdot)$. La procédure est répétée de nombreuses fois (*e.g.*, 10^5 fois) et la *p-value* associée au test unilatéral est calculée comme en (9.3) ou en (9.4) selon que la corrélation est positive ou négative.

Ce mode d'inférence étant entièrement fondé sur un modèle, le choix de la classe de la fonction aléatoire et la modélisation de sa fonction structurale doivent être considérés avec le plus grand soin. Dans le cadre des exemples issus de notre étude de Monte-Carlo, il est attendu que le test de Monte-Carlo donne de très bons résultats. En effet, les conditions d'application sont idéales puisque les données résultent de l'échantillonnage de réalisations de simples FAST-2.

Nous considérons évidemment qu'en pratique le variogramme doit être modélisé à partir des données, mais pour apprécier l'origine de l'écart entre la *p-value* estimée et la *p-value* de référence, nous donnons également les résultats obtenus avec les modèles des variogrammes théoriques utilisés pour simuler les populations finies. Le pas de la grille 10×10 de l'échantillonnage systématique est $\Delta = 3$, mais nous avons calculé les variogrammes empiriques pour 10 classes de pas $\Delta = 1.5$ et de tolérance $\varepsilon = 0.75$ afin d'encadrer assez finement la portée de l'autocorrélation spatiale. Les modèles de variogrammes ont été ajustés par moindres carrés pondérés (Section 7.2.3.3) (Tab. 9.3, Fig. 9.6).

Echantillon	Modèle	c_0	c	a
es_1	exponentiel	0	8993.50	8.84
es_2	gaussien	1	7912.30	10.45
es_3	périodique	1	7026.74	10.39
es_4	sphérique	0	8822.50	10.54

Tableau 9.3: Types de modèles ajustés et paramètres estimés par moindres carrés pondérés pour les variogrammes empiriques des échantillons es_1, es_2, es_3 et es_4.



Figure˜9.6: Variogrammes expérimentaux et modèles ajustés pour les quatre échantillons systématiques es_1, es_2, es_3 et es_4.

Les résultats du Tableau 9.4 montrent un très bon accord entre les *p-values* de référence et celles des tests de Monte-Carlo spatiaux. La comparaison entre les *p-values* obtenues avec les modèles des variogrammes théorique et empirique montre la robustesse de la méthode vis-à-vis de la procédure d'inférence que nous avons utilisée.

Afin de mener une étude de sensibilité plus complète, il conviendrait d'évaluer l'impact :

- de l'erreur d'identification du modèle,
- d'une mauvaise estimation du variogramme empirique (*e.g.*, mauvais découpage en classes de distances),
- d'un mauvais ajustement du modèle (*e.g.,* ajustement manuel grossier).

Cependant, la validité du test tient davantage au choix du modèle général (stationnarité *vs.* non-stationnarité, isotropie *vs.* anisotropie), qu'au choix du modèle de fonction structurale, pour autant que les données autorisent une estimation correcte et que l'ajustement soit conduit avec soin.

$x(\cdot)$	$y(\cdot)$	r_{obs}	p_0	p_1	p_2
es_2	es_1	0.37663	0.00939	0.00911	0.00691
es_1	es_2	0.37663	0.00456	0.00447	0.00495
es_3	es_1	0.27965	0.04195	0.04175	0.03626
es_1	es_3	0.27965	0.03996	0.04120	0.04520
es_4	es_1	0.94259	0.00001	0.00001	0.00001
es_1	es_4	0.94259	0.00001	0.00001	0.00001
es_3	es_2	0.73513	0.00002	0.00001	0.00002
es_2	es_3	0.73513	0.00001	0.00001	0.00001
es_4	es_2	0.61017	0.00005	0.00006	0.00008
es_2	es_4	0.61017	0.00014	0.00012	0.00014
es_4	es_3	0.41381	0.01252	0.01287	0.01566
es_3	es_4	0.41381	0.01247	0.01202	0.01352

Tableau 9.4: Résultats des tests de Monte-Carlo spatiaux pour les couples d'échantillons formés par es_1, es_2, es_3 et es_4. r_{obs} : valeur observée du coefficient de corrélation de Pearson. p_0 : *p-value* de référence. p_1 : *p-value* du test utilisant le modèle du variogramme théorique. p_2 : *p-value* du test utilisant le modèle du variogramme empirique. Le nombre de valeurs constituant la distribution empirique de r sous H_0 s'élève à 10^5.

9.1.5 Test paramétrique modifié

Richardson & Hémon (1981) dérivent une expression asymptotique de la variance de r sous H_0 pour deux FA stationnaires gaussiennes définies sur une grille, sans en dériver toutefois d'application pour le test de r. Par la suite, Clifford & Richardson (1985) et Clifford *et al.* (1989) proposent de modifier le test t de la corrélation en utilisant une estimation de la variance de r et une *taille d'échantillon efficace* tenant compte de l'autocorrélation des deux variables, notées classiquement X et Y.

Soient $\mathbf{\Sigma}_X$ la matrice de covariance de X et $\mathbf{\Sigma}_Y$ celle de Y ; sous l'hypothèse que les variances et covariances sont spatialement homogènes, on peut estimer la covariance de X et de Y en utilisant l'estimateur classique :

$$\widehat{C}_Y(h) = \frac{1}{N(h)} \sum_{(i,j)|h_{ij}=h} (y_i - \bar{y})(y_j - \bar{y}) \tag{9.8}$$

Selon Clifford *et al.* (1989), la variance de r peut être estimée par :

$$\widehat{\sigma}_r^2 = \frac{\sum N(h) \widehat{C}_X(h) \widehat{C}_Y(h)}{n^2 s_X^2 s_Y^2} \tag{9.9}$$

et la taille d'échantillon efficace définie comme $m = 1 + \widehat{\sigma}_r^{-2}$. On constate que cette correction prend bien en compte l'autocorrélation spatiale des variables X et Y. Si les deux variables sont positivement autocorrélées alors $m < n$. Si au moins une des deux variables présente de l'autocorrélation négative, il est possible d'avoir $m > n$. Pour des variables non autocorrélées, il est logiquement attendu que $m = n$. La méthode est évaluée au moyen d'une étude de Monte-Carlo utilisant des processus autorégressifs gaussiens dans le cas de supports répartis sur une grille, et la méthode utilisant la décomposition de Cholesky de la matrice de covariance (Section 4.3.3.2) pour le cas des répartitions irrégulières (Clifford *et al.* 1989).

Dutilleul (1993) fait remarquer que l'estimateur (9.9) est en réalité une approximation qu'il est possible d'écrire sous la forme :

$$\widehat{\sigma}_r^2 \simeq \frac{\mathrm{tr}(\widehat{\mathbf{\Sigma}}_X \widehat{\mathbf{\Sigma}}_Y)}{\mathrm{tr}(\widehat{\mathbf{\Sigma}}_X)\,\mathrm{tr}(\widehat{\mathbf{\Sigma}}_Y)} \tag{9.10}$$

et qu'il compare à l'estimateur complet :

$$\widehat{\sigma}_r^2 = \frac{\mathrm{tr}(\mathbf{B}\widehat{\mathbf{\Sigma}}_X \mathbf{B}\widehat{\mathbf{\Sigma}}_Y)}{\mathrm{tr}(\mathbf{B}\widehat{\mathbf{\Sigma}}_X)\,\mathrm{tr}(\mathbf{B}\widehat{\mathbf{\Sigma}}_Y)} \tag{9.11}$$

avec $\mathbf{B} = n^{-1}(\mathbf{I}_n - n^{-1}\mathbf{J}_n)$ où $\mathbf{I}_n$ est une matrice identité $n \times n$ et $\mathbf{J}_n$ est une matrice $n \times n$ remplie de 1. Soient m_1 et m_2 les tailles d'échantillons efficaces calculées respectivement avec (9.10) et (9.11). Au moyen d'une étude de Monte-Carlo utilisant des processus autorégressifs, Dutilleul (1993) montre qu'en règle générale la différence entre m_1 et m_2 est sans grande conséquence en pratique sauf lorsque n est petit et que l'autocorrélation est élevée. Dutilleul (1993) recommande d'utiliser par défaut l'estimateur (9.11). De même que pour l'estimateur (9.9) (Clifford *et al.* 1989), les performances du test modifié utilisant (9.11) ont été évaluées dans des études de Monte-Carlo et semblent assez bonnes (Dutilleul 1993).

Cependant, l'utilisation pratique de ce test modifié nécessite de considérer les trois choix suivants (Haining 1991) :

- le choix de l'estimateur pour calculer les éléments de $\widehat{\mathbf{\Sigma}}_X$ et de $\widehat{\mathbf{\Sigma}}_Y$,
- le choix du nombre d'éléments à estimer,
- le choix entre des estimations individuelles et des valeurs issues d'un modèle.

Il est possible d'estimer la covariance spatiale de nombreuses façons, ce qui renvoit au Chapitre 7. Haining (1991) fait remarquer que l'estimateur (9.8) est sans biais uniquement si l'on connaît la moyenne (supposée homogène). Si tel n'est pas le cas, l'estimation de la moyenne introduit un biais, et il devient préférable d'estimer la covariance en passant par le variogramme. Mais le calcul de la covariance à partir du variogramme suppose que ce dernier est borné par un seuil qu'il convient de déterminer. Enfin, le découpage en classes de distances introduit une dose d'arbitraire supplémentaire, de même que le nombre de classes à considérer dans les calculs. Dans le cas de supports répartis sur une grille régulière, Clifford & Richardson (1985) utilisent par exemple quatre classes pour la covariance et considèrent au-delà que la covariance est nulle, tandis que dans un cas similaire, Clifford *et al.* (1989) utilisent toutes les classes de distances. Il est également possible d'ajuster un modèle au covariogramme expérimental ou au variogramme expérimental et d'en dériver les covariances pour tous les couples de supports.

Afin de retenir une stratégie pour calculer $\widehat{\mathbf{\Sigma}}_X$ et $\widehat{\mathbf{\Sigma}}_Y$, nous avons comparé les *p-values* obtenues en appliquant six approches différentes (p_1 à p_6) à une *p-value* de référence (p_0). Le test étant symétrique en X et en Y, pour simplifier nous avons déterminé une seule *p-value* de référence en arrondissant la moyenne des deux *p-values* obtenues pour chaque couple (Section 9.1.1, Tab. 9.1).

Les stratégies testées consistent à calculer chaque matrice de covariance à partir :

- du modèle de variogramme théorique utilisé dans l'étude de Monte-Carlo (p_1),
- du modèle de variogramme empirique (p_2),
- du variogramme empirique pour lequel on a déterminé la portée et le seuil, en fixant la covariance à zéro au-delà de la portée (p_3),
- de la même procédure que pour p_3 mais en prenant pour seuil la variance s^2_{n-1} (p_4),
- de la covariance empirique pour laquelle on a déterminé la portée, en fixant la diagonale principale à s^2_{n-1} et la covariance à zéro au-delà de la portée (p_5),
- de la covariance empirique, en considérant toutes les classes et en fixant la diagonale principale à $s^2_{n-1}(p_6)$.

Les variogramme et covariogramme empiriques ont été calculés pour 25 classes avec un pas constant $\Delta = 1.5$ et une tolérance $\varepsilon = 0.75$. Les modèles de variogrammes sont ceux de la section précédente (Tab. 9.3). Les portées et les seuils sont déterminés en examinant les p-grammes (Section 3.5.5), chacun étant calculé à partir de 10^5 valeurs (Tab. 9.5).

Echantillon	$c_0 + c$	s^2_{n-1}	$a_{\widehat{\gamma}}$	$a_{\widehat{C}}$
es_1	7590.55	7696.67	4.5	9.0
es_2	6860.90	6165.44	9.0	9.0
es_3	5848.96	6168.07	6.0	9.0
es_4	6772.19	7228.40	6.0	9.0

Tableau 9.5: Seuils des variogrammes ($c_0 + c$), portées des variogrammes ($a_{\widehat{\gamma}}$) et des corrélogrammes ($a_{\widehat{C}}$) déterminées par examen de p-grammes, et variances (s^2_{n-1}) pour les échantillons es_1, es_2, es_3 et es_4.

Les résultats du Tableau 9.6 montrent de façon évidente l'impact du choix de la stratégie de calcul des matrices de covariance sur les *p-values* obtenues par le test modifié. Aucune stratégie ne s'avère systématiquement meilleure que les autres. Il ne semble pas y avoir d'avantage à utiliser un modèle de variogramme plutôt que le variogramme empirique. Le pire résultat est obtenu avec le covariogramme lorsque toutes les classes sont utilisées (p_6). En revanche, lorsque la covariance est fixée à zéro après la portée, des résultats assez satisfaisants sont obtenus, aussi bien avec le variogramme empirique dont le seuil est estimé par s^2_{n-1} qu'avec le covariogramme (p_4 et p_5). Au moins dans les cas étudiés, il semble n'y avoir aucun avantage à utiliser le variogramme, et l'utilisation du covariogramme dont on fixe les valeurs à zéro après la portée se révèle très satisfaisante. Nous recommandons cependant de déterminer les portées en examinant les p-grammes.

Le choix de la stratégie à adopter pour calculer les matrices de covariance étant arrêté, Haining (1991) pose également le problème de la non-homogénéité des moyennes et suggère de filtrer une éventuelle tendance au préalable, ce qui revient à agir à la fois au niveau des données et au niveau du test. Clifford *et al.* (1989) considèrent que la formulation de la modification du test est suffisamment générale pour pouvoir traiter le cas des variances non homogènes ou d'autres types de "non-stationnarités".

X, Y	r_{obs}	p_0	p_1	p_2	p_3	p_4	p_5	p_6
es_1,es_2	0.37663	0.00700	0.01723	0.01428	0.00373	0.00386	0.00656	0.01483
es_1,es_3	0.27965	0.04100	0.04101	0.03880	0.02048	0.02248	0.03272	0.05599
es_1,es_4	0.94259	0.00001	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
es_2,es_3	0.73513	0.00003	0.00072	0.00111	0.00049	0.00048	0.00088	0.00242
es_2,es_4	0.61017	0.00010	0.00195	0.00282	0.00041	0.00052	0.00087	0.00348
es_3,es_4	0.41381	0.01250	0.01534	0.01878	0.00768	0.01002	0.01540	0.02900

Tableau 9.6: Résultats du test paramétrique modifié pour les couples formés par les échantillons es_1, es_2, es_3 et es_4 en fonction de la stratégie de calcul des matrices de covariance. r_{obs} : valeur observée du coefficient de corrélation de Pearson. La *p-value* de référence est p_0 et les *p-values* p_1 à p_6 sont celles résultant des six stratégies envisagées (détails dans le texte).

Un dernier problème qui n'a pas été considéré jusqu'à présent est celui posé par la validité du recours à la distribution de Student. Pour appliquer le test paramétrique de r il faut supposer que l'échantillon est issu d'une population bivariée normale. Tout et son contraire a été écrit au sujet de la sensibilité du test de r à la violation de l'hypothèse de la normalité (revue dans Kowalski 1972). Dans une étude de Monte-Carlo, Kowalski (1972) conclut que la non-normalité a un impact sur la distribution de r, même quand $\rho = 0$. Cet auteur précise que la variance constitue le paramètre le plus vulnérable. Il serait intéressant d'apprécier la robustesse du test modifié lorsque l'hypothèse de normalité n'est pas raisonnable.

Face aux interrogations posées par l'utilisation du test paramétrique modifié, une alternative intéressante serait de disposer d'un test non paramétrique valide en présence d'autocorrélation.

9.1.6 Test de randomisation stratifiée

De même que les tests paramétriques, les tests de randomisation (ou les tests de permutation qu'ils approximent) supposent l'indépendance des données (Edgington 1986). Plus précisément, c'est *l'échangeabilité* qui est requise (Good 1994), l'indépendance impliquant l'échangeabilité mais pas l'inverse (Galambos 1982). Soit $A_1, A_2, \ldots, A_k$ une séquence d'événements dont la probabilité est :

$$P_k = \Pr\left(\bigcap_{i=1}^{k} A_i\right) \tag{9.12}$$

Cette séquence est échangeable si pour tout $k \geq 1$ la probabilité P_k ne dépend pas de i mais uniquement de k, *i.e.* est invariante par permutation des indices (Galambos 1982).

Le concept d'échangeabilité a été introduit pour remplacer la notion *d'indépendance avec même probabilité inconnue* (*cf.* de Finetti 1979), et joue un rôle fondamental dans l'inférence Bayesienne et l'inférence non paramétrique (Galambos 1982). Dans le contexte Bayesien ou "subjectiviste", un événement correspond à un seul cas complètement spécifié. De même, une quantité aléatoire correspond à une seule quantité complètement spécifiée z_i. Dans ce contexte, il n'est pas nécessaire de faire référence à la notion de variable aléatoire au contraire de l'inférence basée sur des modèles, qui considère que

chaque valeur mesurée z_i est issue d'une variable aléatoire Z_i (de Finetti 1979). Des échantillons aléatoires simples tirés d'une population finie sont échangeables, de même que des données spatialement indépendantes, *i.e.* dont les valeurs sont distribuées dans l'espace selon un ordre aléatoire. Des variables aléatoires dépendantes normalement distribuées $\{Z_i\}$ sont échangeables si la variance de Z_i est une constante indépendante de i et si la covariance entre Z_i et Z_j est constante et indépendante de i et j (Good 1994). Plus généralement, dans le cadre d'une superpopulation telle qu'un modèle de fonction aléatoire $\{Z_i \mid i = 1, \ldots, N\}$, l'échangeabilité implique que les VA $Z_{r_1}, \ldots, Z_{r_N}$ ont, pour toute permutation $r_1, \ldots, r_N$ de $1, \ldots, N$, la même distribution conjointe ξ (Cassel *et al.* 1977).

Dans un contexte spatial, l'inférence fondée sur un test de randomisation devient envisageable lorsque les permutations aléatoires sont autorisées uniquement au sein de sous-ensembles de données pour lesquels l'échangeabilité devient une hypothèse raisonnable. Intuitivement, un test de *randomisation stratifiée*[1] doit limiter les permutations aléatoires aux valeurs similaires, en respectant globalement la structure spatiale. Selon nous, une méthode de randomisation stratifiée implique donc deux étapes indépendantes de :

- définition des sous-ensembles de données échangeables,
- permutation aléatoire au sein de ces sous-ensembles.

Considérons de façon générale une partition spatiale des données en g sous-ensembles. Soit μ_k la fonction d'appartenance de l'ensemble k, la partition est entièrement définie par la matrice d'appartenance $\mathbf{A}$ d'éléments $a_{ik} = \mu_k(z_i)$ avec $k = 1, \ldots, g$ et $i = 1, \ldots, n$. Nous définissons la probabilité de l'événement π_{ij} "permutation de z_i et z_j" comme :

$$\Pr(\pi_{ij}) = \sum_{k=1}^{g} \mu_k(z_i) \cdot \mu_k(z_j)$$

avec $\sum \mu_k(z_i) = \sum \mu_k(z_j) = 1$. Cette expression générale permet de traiter des partitions dures aussi bien que des partitions floues. Dans le cas d'une partition dure, on peut écrire plus simplement :

$$\Pr(\pi_{ij}) = \begin{cases} 1 & \text{si } z_i \text{ et } z_j \text{ sont dans le même sous-ensemble} \\ 0 & \text{sinon} \end{cases}$$

Le problème essentiel de l'approche que nous proposons est celui de la définition de la partition. Il convient notamment de décider si cette partition doit être dure ou floue. Si l'on connaît le nombre de sous-ensembles g, il est possible d'obtenir une partition dure optimale au sens de la minimisation de la variance intra-groupe en utilisant l'algorithme de Fisher (Fisher 1958, Diday *et al.* 1982, Aubry & Egretaud 1994). Une partition floue peut également être obtenue, par exemple en appliquant l'algorithme NEM (*Neighborhood*

[1] En ce qui concerne les tests de randomisation, nous préférons parler de *randomisation stratifiée* (Noreen 1989), plutôt que de *randomisation restreinte* (Sokal *et al.* 1993) ou *contrainte* (Legendre *et al.* 1990), afin d'éviter d'éventuelles confusions avec la *randomisation restreinte* (ou *contrainte*) telle qu'elle est définie dans le cadre des dispositifs expérimentaux (*cf.* Grundy & Healy 1950, White & Welch 1981, Bailey 1983, 1985, 1986).

Expectation-Maximisation) (Dang 1998). Si l'on accepte ces algorithmes de partitionnement, le seul problème qui reste à résoudre est celui de la détermination du nombre de sous-ensembles g, qui est *a priori* inconnu. Un raisonnement circulaire apparaît immédiatement en ce que la valeur correcte de g est celle qui conduit au test de randomisation stratifiée dont la *p-value* est proche de la *p-value* correcte. Or, dans la pratique, on ignore évidemment quelle est la *p-value* correcte. Pour sortir de ce raisonnement circulaire, il est possible de recourir à des heuristiques pour déterminer g, par exemple en utilisant le critère BIC (*Bayesian Information Criterion*) et un modèle (Dang 1998).

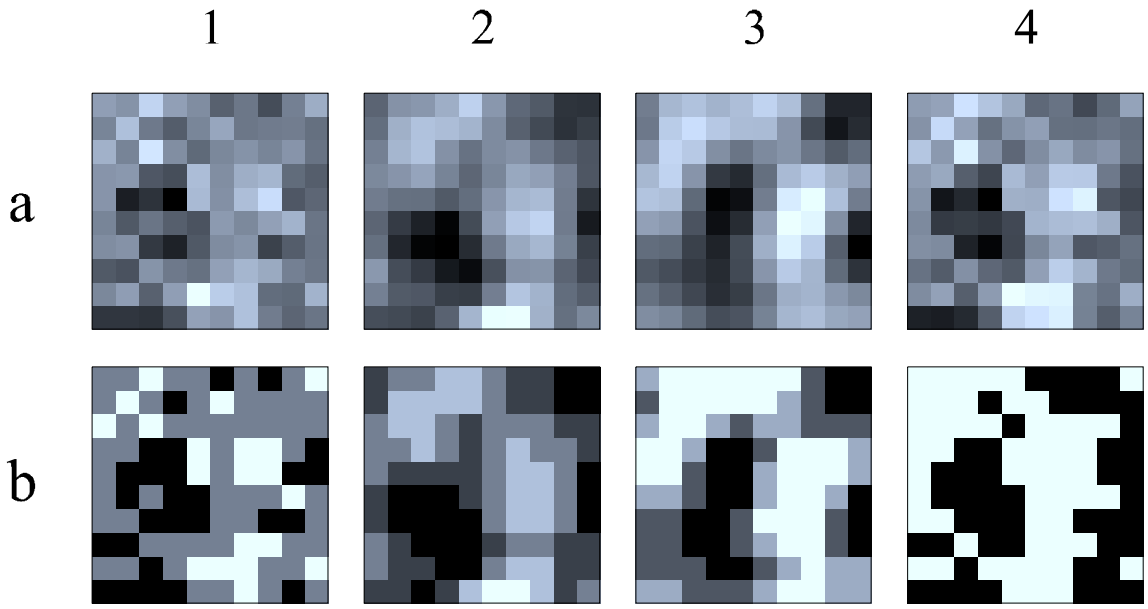
En utilisant des partitions floues obtenues par l'algorithme NEM[2], il nous est apparu que la randomisation stratifiée en contexte flou correspondait en fait à la randomisation classique. Par conséquent, la randomisation stratifiée nécessite des partitions dures. Nous avons appliqué l'algorithme de Fisher pour partitionner les échantillons es_1, es_2, es_3 et es_4, respectivement en 3, 5, 4 et 2 classes (Fig. 9.7). Les tests de randomisation stratifiée résultant de l'utilisation de ces partitions ont conduit aux résultats du Tableau 9.7. Ces résultats apparaissent globalement satisfaisants, avec cependant une estimation médiocre de la *p-value* dans le cas du couple (es_2,es_3), *i.e.* entre les données issues des modèles gaussien et périodique ($r = 0.73513$). La surestimation observée semble liée à la régularité spatiale des VR considérées. Le nombre de classes est du reste plus élevé dans le cas des variables régulières es_2 ($g = 5$) et es_3 ($g = 4$), que dans le cas des variables non régulières es_1 ($g = 3$) et es_4 ($g = 2$). Le comportement à l'origine du variogramme constitue donc un élément important qu'il faudrait considérer explicitement dans la randomisation stratifiée, ce qui impliquerait que l'on ne puisse pas faire l'économie d'une analyse de la structure d'autocorrélation spatiale. Le problème de la prise en compte de l'autocorrélation dans la définition de la partition spatiale reste cependant totalement ouvert.

En essayant un grand nombre de stratégies de partitionnement spatial, il nous est apparu que le test de randomisation stratifiée est excessivement sensible à la partition des données qui est utilisée. Sous sa forme actuelle, la randomisation stratifiée ne peut donc pas être recommandée comme alternative au test paramétrique modifé. En outre, la randomisation stratifiée n'autorise que le test unilatéral parce que la corrélation ne peut jamais changer de signe, à cause du respect de la structure spatiale globale. Toutefois, l'idée d'une randomisation stratifiée par une partition reste séduisante et mériterait une étude approfondie de l'échangeabilité des données spatiales.

9.1.7 Approche du type "corrélation partielle"

En écologie, le problème de l'impact de l'autocorrélation spatiale dans le test de la corrélation entre deux VR $x(\cdot)$ et $y(\cdot)$ est interprété comme un problème de corrélation indirecte entre deux variables X et Y. La question est de savoir si la corrélation observée entre deux variables spatialement structurées est significative ou bien due au fait qu'elles sont toutes les deux liées à l'espace géographique, suivent le même gradient, etc. (*e.g.*, Legendre & Fortin 1989, Rossi 1996, Le Corre *et al.* 1997). En plus des deux variables X et Y, l'espace est donc considéré explicitement comme une troisième "variable" (Houle 1996). L'utilisation systématique du test de Mantel partiel en écologie afin de régler le problème de l'autocorrélation spatiale dans le test de la corrélation entre deux variables régionalisées (Annexe G), nécessite d'examiner en détail la question de sa validité.

[2]Les partitions floues ont été calculées par le Dr. Mô Dang.



Figure˜9.7: Données et partitions associées. (a) Echantillons systématiques issus des populations simulées d'après des modèles de variogrammes : (1) exponentiel, (2) gaussien, (3) périodique, (4) sphérique. (b) Partitions optimales des échantillons systématiques, respectivement en 3, 5, 4 et 2 classes.

$x(\cdot)$	$y(\cdot)$	r_{obs}	p_0	p_1	g
es_2	es_1	0.37663	0.00939	0.01442	3
es_1	es_2	0.37663	0.00456	0.00624	5
es_3	es_1	0.27965	0.04195	0.03518	3
es_1	es_3	0.27965	0.03996	0.03091	4
es_4	es_1	0.94259	0.00001	0.00001	3
es_1	es_4	0.94259	0.00001	0.00001	2
es_3	es_2	0.73513	0.00002	0.00372	5
es_2	es_3	0.73513	0.00001	0.00787	4
es_4	es_2	0.61017	0.00005	0.00062	5
es_2	es_4	0.61017	0.00014	0.00001	2
es_4	es_3	0.41381	0.01252	0.00841	4
es_3	es_4	0.41381	0.01247	0.01243	2

Tableau 9.7: Résultats des tests de randomisation stratifiée pour les couples d'échantillons formés par es_1, es_2, es_3 et es_4. p_0 : *p-value* de référence. p_1 : *p-value* du test de randomisation stratifiée. g : nombre de classes de la partition de la variable randomisée $y(\cdot)$. Le nombre de valeurs constituant la distribution empirique de r sous H_0 s'élève à 10^5.

9.1.7.1 Test de Mantel partiel

De même que l'utilisation de la corrélation partielle permet de traiter le problème de la fausse corrélation (Dagnelie 1986, Sokal & Rohlf 1995), on espère régler le problème de la fausse corrélation due à la variable "espace" (*i.e.*, le problème posé par l'autocorrélation spatiale) en utilisant un test de Mantel partiel. Ce test est une simple généralisation du calcul du coefficient de corrélation partielle à partir du test de Mantel (Smouse *et al.* 1986). Soit la statistique de Mantel (Mantel 1967) :

$$Z = \sum_{i<j} m_{ij}^{(x)} m_{ij}^{(y)} \tag{9.13}$$

où $\mathbf{M}^{(x)}$ et $\mathbf{M}^{(y)}$ sont deux matrices de similarités $n \times n$ symétriques, construites, respectivement, à partir des n valeurs des variables X et Y. Dans un test de permutation, Z n'est pas autre chose qu'une statistique équivalente au coefficient de corrélation de Pearson entre les éléments des deux matrices $\mathbf{M}^{(x)}$ et $\mathbf{M}^{(y)}$. Ce qui caractérise le test de Mantel ce n'est donc pas la statistique mesurant la corrélation entre les distances mais le schéma permutationnel requis pour la tester, *i.e.* par permutation des lignes et des colonnes d'une des deux matrices. Pour simplifier ce qui suit, *test de corrélation de Mantel* désigne le test de $r(\mathbf{M}^{(x)}, \mathbf{M}^{(y)})$ et *test de corrélation de Pearson* désigne le test de $r(X, Y)$, autrement dit, la corrélation linéaire directement mesurée entre les valeurs des variables X et Y. Soit une troisième variable Z donnant lieu à une la matrice $\mathbf{M}^{(z)}$, par stricte analogie avec la corrélation de Pearson partielle, Smouse *et al.* (1986) définissent la corrélation de Mantel partielle comme :

$$r_{xy.z} = \frac{r_{xy} - r_{xz} r_{yz}}{\left[\left(1 - r_{xz}^2\right)\left(1 - r_{yz}^2\right)\right]^{1/2}} \tag{9.14}$$

avec les notations $r_{xy} \equiv r(\mathbf{M}^{(x)}, \mathbf{M}^{(y)})$, $r_{xz} \equiv r(\mathbf{M}^{(x)}, \mathbf{M}^{(z)})$ et $r_{yz} \equiv r(\mathbf{M}^{(y)}, \mathbf{M}^{(z)})$. Le test de $r_{xy.z}$ peut s'effectuer selon deux procédures formellement équivalentes (Smouse *et al.* 1986) :

- par permutation des lignes et colonnes d'une des deux matrices $\mathbf{M}^{(x)}$ ou $\mathbf{M}^{(y)}$ en fixant les deux matrices restantes[3] (considérons arbitrairement que la matrice permutée est $\mathbf{M}^{(y)}$),
- en calculant les matrices des résidus des régressions linéaires de $\mathbf{M}^{(x)}$ et $\mathbf{M}^{(y)}$ par $\mathbf{M}^{(z)}$ que nous notons respectivement $\text{res}(\mathbf{M}^{(x)}, \mathbf{M}^{(z)})$ et $\text{res}(\mathbf{M}^{(y)}, \mathbf{M}^{(z)})$, puis en effectuant un test de Mantel classique pour ces deux matrices de résidus.

Néanmoins, le test de permutation (ou de randomisation si $n > 10$) nécessite deux fois plus de calculs avec la première approche qu'avec la seconde. En effet, dans la première approche il faut recalculer la corrélation de Mantel entre les deux couples de matrices ($\mathbf{M}^{(x)}$, $\mathbf{M}^{(y)}$) et ($\mathbf{M}^{(y)}$, $\mathbf{M}^{(z)}$) au lieu du seul couple ($\text{res}(\mathbf{M}^{(x)}, \mathbf{M}^{(z)})$, $\text{res}(\mathbf{M}^{(y)}, \mathbf{M}^{(z)})$) dans la seconde.

Afin de juger si l'utilisation du test de Mantel partiel permet de traiter le problème de l'autocorrélation spatiale dans le test de corrélation de Pearson, il convient tout d'abord d'étudier le lien entre les tests de corrélation de Mantel et de Pearson.

[3] Legendre & Fortin (1989) se trompent en mentionnant la permutation des trois matrices.

9.1.7.2 Test de corrélation : Mantel vs. Pearson

En absence d'étude formelle du lien entre la corrélation de Pearson mesurée sur les valeurs des variables et celle mesurée sur des similarités entre valeurs des variables, nous proposons une étude empirique. Considérons les valeurs $x_i = i$ avec $i = 1, \ldots, n$ avec $n = 100$. La corrélation positive parfaite $r(X, Y) = 1$ est obtenue pour $y_i = x_i$, pout tout i. Afin de couvrir l'intervalle des valeurs de r entre 0 et 1, il suffit de modifier Y, par exemple en tirant au hasard et sans remise m indices i et en remplaçant chaque valeur y_i par une valeur différente tirée au hasard d'une loi rectangulaire (*e.g.*, entre 1 et 30). Pour inverser le signe de la corrélation, il suffit d'inverser l'ordre des valeurs x_i en faisant $x'_i = x_{n+1-i}$. Pour le test de Mantel, considérons les similarités entre deux valeurs x_i et x_j du type covariance centrée $(x_i - \bar{x})(x_j - \bar{x})$.

Legendre & Fortin (1989) indiquent que les valeurs de la corrélation de Mantel ne se comportent pas comme celles de la corrélation de Pearson, et que des valeurs significatives ne sont pas nécessairement élevées en valeur absolue. En effet, les résultats du Tableau 9.8 montrent bien que les valeurs de r_{xy} sont systématiquement inférieures à celles de $r(X, Y)$ — sauf pour la corrélation parfaite bien entendu. Dans notre exemple $r_{xy} \simeq 0.07$ correspond à $r(X, Y) \simeq 0.28$ pour $n = 100$ et $m = 70$. Nous avons précisé dans Vieira *et al.* (1998) que les valeurs de r_{xy} sont en pratique bien plus faibles que celles de $r(X, Y)$, et que la distribution de r_{xy} sous H_0 est très asymétrique. En outre, avec la définition de la similarité entre valeurs que nous utilisons, nous avons établi empiriquement que l'espérance de r_{xy} sous H_0 vaut $-(n-1)^{-1}$ et non pas 0 comme pour $r(X, Y)$. Ce résultat est en accord avec ceux de Hubert *et al.* (1981) concernant l'équivalence — à un facteur de standardisation près — entre la statistique de Mantel utilisant la similarité $(x_i - \bar{x})(x_j - \bar{x})$ et le I de Moran, dont on sait que l'espérance sous H_0 vaut précisément $-(n-1)^{-1}$ (Upton & Fingleton 1985, p. 170). En conséquence des différences qui existent entre r_{xy} et $r(X, Y)$, seules les *p-values* associées aux valeurs de r_{xy} sont directement interprétables (Vieira *et al.* 1998).

Sokal & Thomson (1987) expliquent qu'en utilisant le test de Mantel il ne s'agit pas de traiter de la corrélation entre variables mais de la corrélation entre distances. De même, Rossi (1996) et Rossi & Quenehervé (1998) notent que la corrélation de Mantel est une corrélation entre matrices et qu'elle n'est pas équivalente à la corrélation entre les variables utilisées pour former ces matrices. Si les valeurs de la corrélation sont effectivement différentes, les résultats du Tableau 9.8 montrent toutefois que les *p-values* du test de corrélation de Mantel sont pratiquement équivalentes aux *p- values* du test de corrélation de Pearson bilatéral, la corrélation étant $r\left(p^{(bi)}, p\right) \simeq 0.99984$ $(p \leq 0.00001)$.

9.1.7.3 Performances du test de Mantel partiel

Dans les conditions en vigueur dans la section précédente, il apparaît que le test de corrélation de Mantel est pratiquement équivalent au test de corrélation de Pearson bilatéral. En supposant que cette quasi-équivalence est conservée en présence d'autocorrélation spatiale, il est licite d'évaluer les performances du test de Mantel partiel en confrontant ses *p-values* avec celles du test bilatéral de r obtenues dans notre étude de Monte-Carlo (Section 9.1.1).

m	$r(X,Y)$	$p^{(bi)}$	r_{xy}	p
0	±1.00000	0.00001	1.00000	0.00001
10	±0.90046	0.00001	0.80814	0.00001
20	±0.80172	0.00001	0.63843	0.00001
30	±0.75927	0.00002	0.57154	0.00001
40	±0.62722	0.00001	0.38694	0.00001
50	±0.44266	0.00006	0.18700	0.00003
60	±0.31924	0.00142	0.09281	0.00111
70	±0.27928	0.00561	0.06953	0.00348
80	±0.30381	0.00117	0.08594	0.00169
90	±0.17255	0.08913	0.02304	0.07197
95	±0.13042	0.19851	0.00876	0.17974
99	±0.02975	0.77316	-0.00894	0.73073
100	±0.02829	0.78132	-0.00918	0.75093

Tableau 9.8: Résultats de la comparaison entre les tests de corrélation de Pearson et de Mantel (détails dans le texte). m : nombre de valeurs modifiées aléatoirement. $r(X,Y)$: corrélation de Pearson. r_{xy} : corrélation de Mantel. $p^{(bi)}$: *p-value* du test de corrélation de Pearson bilatéral. p : *p-value* du test de la corrélation de Mantel. Le nombre de valeurs considérées dans le calcul des *p-values* s'élève à 10^5.

Pour tester la corrélation entre deux variables X et Y localisées dans un espace muni de la métrique euclidienne, les tests de Mantel partiels sont effectués à partir d'une matrice $\mathbf{M}^{(z)}$ de proximités spatiales et de deux matrices d'auto-similarités $\mathbf{M}^{(x)}$ et $\mathbf{M}^{(y)}$ dont les éléments peuvent être calculés comme :

$$\begin{aligned} M_{ij}^{(x)} &= (x_i - \bar{x})\,(x_j - \bar{x}) \\ M_{ij}^{(y)} &= (y_i - \bar{y})\,(y_j - \bar{y}) \end{aligned}$$

L'association entre la matrice d'auto-similarités $\mathbf{M}^{(x)}$ et la matrice de proximités $\mathbf{M}^{(z)}$ est censée modéliser l'autocorrélation spatiale de la variable X. Il en est de même pour Y. La prise en compte de l'autocorrélation spatiale de X et de Y par le test de Mantel partiel nécessite donc que les corrélations r_{xz} et r_{yz} soient les plus élevées possibles. En effet, de faibles corrélations signifieraient que la structure de variation spatiale de chaque variable est mal modélisée. Il apparaît deux problèmes puisqu'il faut :

- modéliser la variation spatiale de deux variables différentes en utilisant une seule et même matrice $\mathbf{M}^{(z)}$,
- définir la proximité spatiale de façon à bien modéliser la variation spatiale.

Afin d'illustrer l'importance du choix de la proximité spatiale, nous proposons *a priori* de calculer $\mathbf{M}^{(z)}$ comme :

$$M_{ij}^{(z)} = h_{ij}^{-1} \tag{9.15}$$

et *a posteriori* de calculer $\mathbf{M}^{(z)}$ comme :

$$M_{ij}^{(z)} = \sum_{k=0}^{q} a_k h_{ij}^{k} \tag{9.16}$$

avec h_{ij} la distance euclidienne entre les supports s_i et s_j. Le modèle (9.16) est un modèle de régression polynomiale de degré q pour la fonction de covariance spatiale. Dans notre exemple, les portées et les seuils des variogrammes ayant servi à simuler les données sont identiques de sorte qu'il est raisonnable de prendre la moyenne des fonctions de covariance (9.8) calculées pour les échantillons es_1, es_2, es_3 et es_4. La covariance spatiale est calculée pour 25 classes $(1.5, 0.75)$. Les classes vides ne sont pas prises en compte dans l'ajustement du modèle polynomial. Un modèle de degré $q = 7$ est ajusté à la fonction de covariance moyenne en utilisant les moindres carrés ordinaires. Dans notre exemple, le modèle de variation spatiale *a priori* conduit systématiquement à des corrélations de Mantel (r_{xz}) inférieures à celles obtenues avec le modèle *a posteriori* (Tab. 9.9).

Echantillon	$r_{xz}^{(1)}$	$r_{xz}^{(2)}$
es_1	0.05646	0.09014
es_2	0.16164	0.21838
es_3	0.15091	0.22204
es_4	0.10288	0.15839

Tableau 9.9: Corrélations de Mantel simples entre la matrice d'auto-similarités et la matrice de proximités spatiales pour les échantillons es_1, es_2, es_3 et es_4. $r_{xz}^{(1)}$ et $r_{xz}^{(2)}$: corrélations pour les proximités spatiales définies *a priori* et *a posteriori*.

Les résultats du Tableau 9.10 montrent la quasi-équivalence entre les tests de corrélation de Pearson et de Mantel en présence d'autocorrélation spatiale. En effet, les corrélations de Mantel (r_{xy}) sont inférieures à celles de Pearson (Tab. 9.6), mais les *p-values* $p_1^{(bi)}$ et p_2 sont du même ordre de grandeur. Dans les deux cas, les *p-values* sous-estiment les *p-values* de référence ($p_0^{(bi)}$) pour les couples dont la corrélation n'est pas très élevée, *i.e.* pour tous les couples sauf (es_1,es_4) et (es_2,es_3). Pour ces couples, les tests de Mantel partiels conduisent à une réduction de la corrélation et par conséquent à une augmentation de la *p-value*. L'augmentation de la *p-value* (*i.e.*, la réduction de sa sous-estimation) est d'autant plus élevée que le modèle de variation spatiale est meilleur ($p_3^{(2)} - p_2 > p_3^{(1)} - p_2$). Mais la réduction de la sous-estimation des *p-values* s'avère très insuffisante. Cet exemple remet sérieusement en question l'assertion selon laquelle le test de Mante partiel résoud le problème posé par l'autocorrélation spatiale (*e.g.*, Legendre & Troussellier 1988).

X, Y	r_{xy}	$r_{xy.z}^{(1)}$	$r_{xy.z}^{(2)}$	$p_0^{(bi)}$	$p_1^{(bi)}$	p_2	$p_3^{(1)}$	$p_3^{(2)}$
es_1,es_2	0.13428	0.12702	0.11791	0.01410	0.00055	0.00017	0.00020	0.00031
es_1,es_3	0.06663	0.05888	0.04801	0.08200	0.00517	0.00599	0.00875	0.01524
es_1,es_4	0.88789	0.88821	0.88840	0.00001	0.00001	0.00001	0.00001	0.00001
es_2,es_3	0.53928	0.52779	0.51580	0.00003	0.00001	0.00002	0.00002	0.00001
es_2,es_4	0.36620	0.35612	0.34416	0.00015	0.00002	0.00001	0.00001	0.00001
es_3,es_4	0.16275	0.14973	0.13252	0.02500	0.00003	0.00007	0.00014	0.00029

Tableau 9.10: Résultats des tests de Pearson et de Mantel pour les couples formés par les échantillons es_1, es_2, es_3 et es_4. r_{xy} : corrélation de Mantel simple. $r_{xy.z}^{(1)}$ et $r_{xy.z}^{(2)}$: corrélations de Mantel partielles pour les proximités spatiales (9.15) et (9.16). $p_0^{(bi)}$ et $p_1^{(bi)}$: *p-values* de référence et du test de randomisation pour le test bilatéral du r de Pearson. p_2 : *p-value* du test de Mantel simple. $p_3^{(1)}$ et $p_3^{(2)}$: *p-values* des tests de Mantel partiels. Chaque *p-value* est calculée à partir d'un ensemble de 10^5 valeurs.

9.1.7.4 Performances des approches du type "corrélation partielle"

En plus du test de Mantel partiel, Sokal & Thomson (1987), Legendre & Fortin (1989) et Oden & Sokal (1992) citent trois autres approches du type "corrélation partielle", proposées par Dow & Cheverud (1985), Hubert (1985) et Manly (1986).

Au moyen d'une étude de Monte-Carlo, Oden & Sokal (1992) évaluent les performances du test de Mantel partiel et les méthodes de Dow & Cheverud (1985) et de Manly (1986), dans le cas de données spatialement autocorrélées. La méthode de Hubert (1985) n'est pas considérée car jugée difficile à interpréter. Oden & Sokal (1992) concluent qu'aucune des méthodes étudiées ne peut être recommandée et expliquent l'échec de toutes ces approches par les schémas permutationnels utilisés, dont aucun ne serait licite en présence d'autocorrélation spatiale.

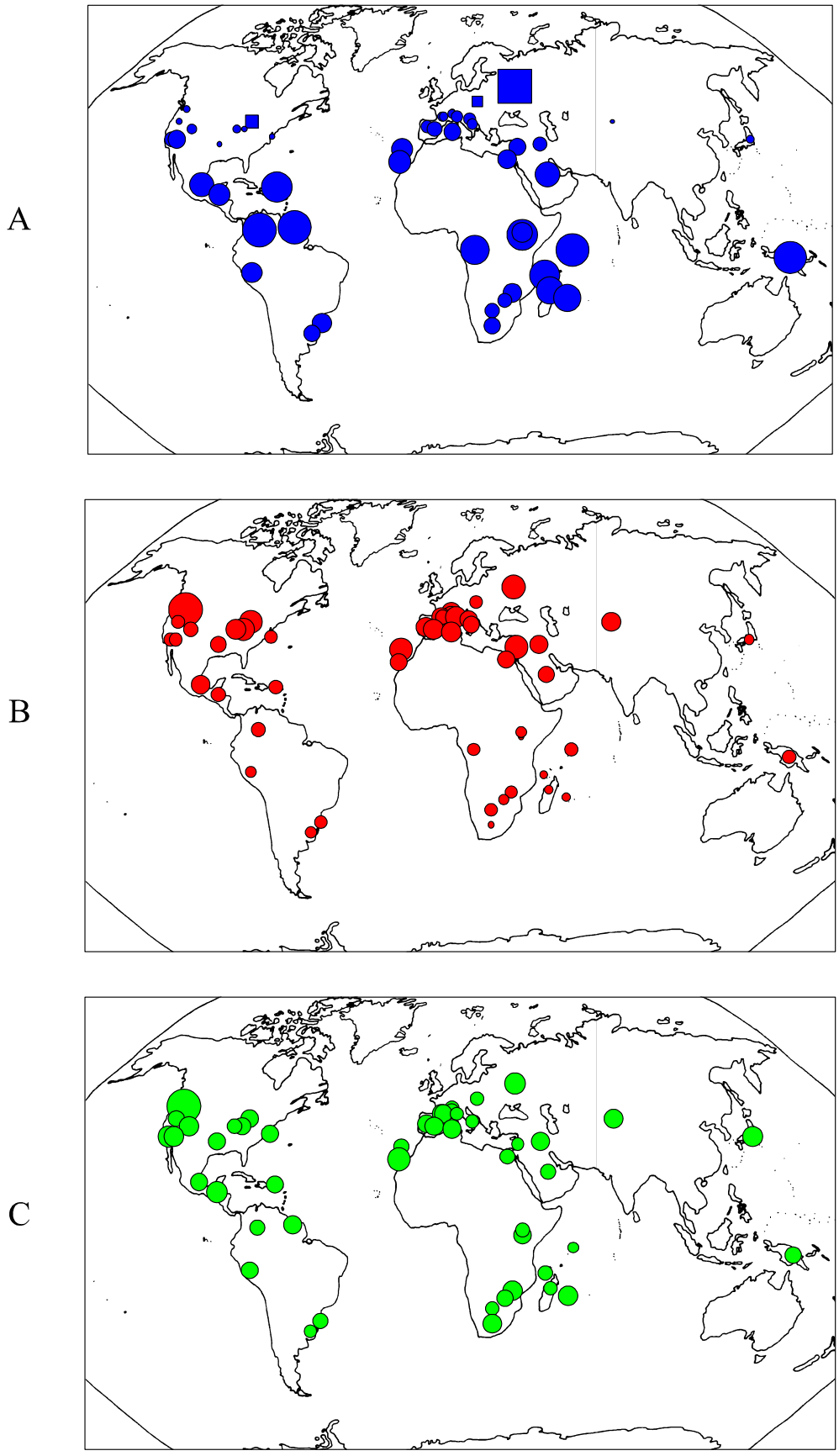
9.1.8 Etude de cas

Dans une récente étude d'écologie génétique, Vieira *et al.* (1998) cherchent à expliquer le gradient du nombre de copies du rétrotransposon 412 chez *Drosophila simulans* observé entre l'hémisphère nord et l'hemisphère sud (Vieira & Biémont 1996). Ce gradient suggère une corrélation entre le nombre de copies et un ou plusieurs facteurs climatiques. Cette corrélation laisse supposer qu'il existe une relation de causalité entre la régulation du nombre de copies de 412 et le climat. L'étude de la corrélation appartient au domaine statistique tandis que l'établissement de la relation de causalité relève de la génétique. Nous nous intéressons ici uniquement à l'aspect statistique et renvoyons à Vieira *et al.* (1998) pour les aspects génétiques de la question.

L'étude concerne $n = 51$ populations réparties à l'échelle planétaire[4]. Le nombre de copies des rétrotransposons 412 et roo/B104 est déterminé dans ces populations et confronté à diverses variables climatiques. Le problème statistique posé par cette étude est celui de la prise en compte de l'autocorrélation spatiale des variables climatiques et du nombre de copies dans le test de la corrélation.

Pour presque toutes les populations, le nombre de copies est déterminé sur deux larves femelles. Dans certains cas on ne dispose que d'une valeur, et les données manquent totalement pour certains des sites géographiques considérés, selon le rétrotransposon considéré. Les valeurs des variables climatiques ont été relevées dans un atlas climatologique de la *World Meteorological Organisation* et donnent une image approximative des conditions climatiques en vigueur dans les sites où ont été prélevées les souches de *D. simulans*. Néanmoins il semble y avoir une corrélation négative entre la température minimale et le nombre de copies de 412 et aucune corrélation avec le nombre de copies de roo/B104. Ce résultat est du reste bien visible sur les représentations cartographiques (Fig. 9.8).

[4]Données communiquées par le Dr. Christian Biémont.



Figure˜9.8: Représentations cartographiques des données à l'échelle mondiale pour 51 populations de *Drosophila simulans*. (A) Température minimale des sites des populations. Nombre moyen de copies des rétrotransposons : (B) 412, (C) roo/B104. Les carrés représentent des valeurs négatives.

En considérant dans un premier temps que les variables ne sont pas spatialement autocorrélées, il suffit de tester le coefficient de corrélation de Pearson, *e.g.* en utilisant le test de Student. Disposant en général de deux valeurs pour chaque population, il est possible de :

1. calculer la moyenne des deux valeurs pour chaque population,
2. considérer l'ensemble des valeurs,
3. considérer les 2^n jeux de données possibles en choisissant tour à tour une seule des deux valeurs pour chaque population.

La première approche considère implicitement que les deux valeurs disponibles sont utilisées afin d'estimer la moyenne du nombre de copies pour chaque population. Estimer une moyenne à partir de seulement deux valeurs ne se justifie que dans la mesure où la variabilité est très faible, ce que semblent démentir les données. Prendre la moyenne des nombres de copies peut se justifier également dans le cas d'une représentation cartographique simplificatrice, mais cela conduit nécessairement à une perte d'information, *a priori* préjudiciable à l'étude statistique. La seconde approche considère que les deux valeurs disponibles pour chaque population ont le même poids, autrement dit, que l'une comme l'autre rend compte aussi bien (ou aussi mal) du nombre de copies moyen dans la population. Enfin, la troisième approche envisage tous les jeux de données qu'il est possible de former en considérant tour à tour chacune des deux valeurs pour chaque population.

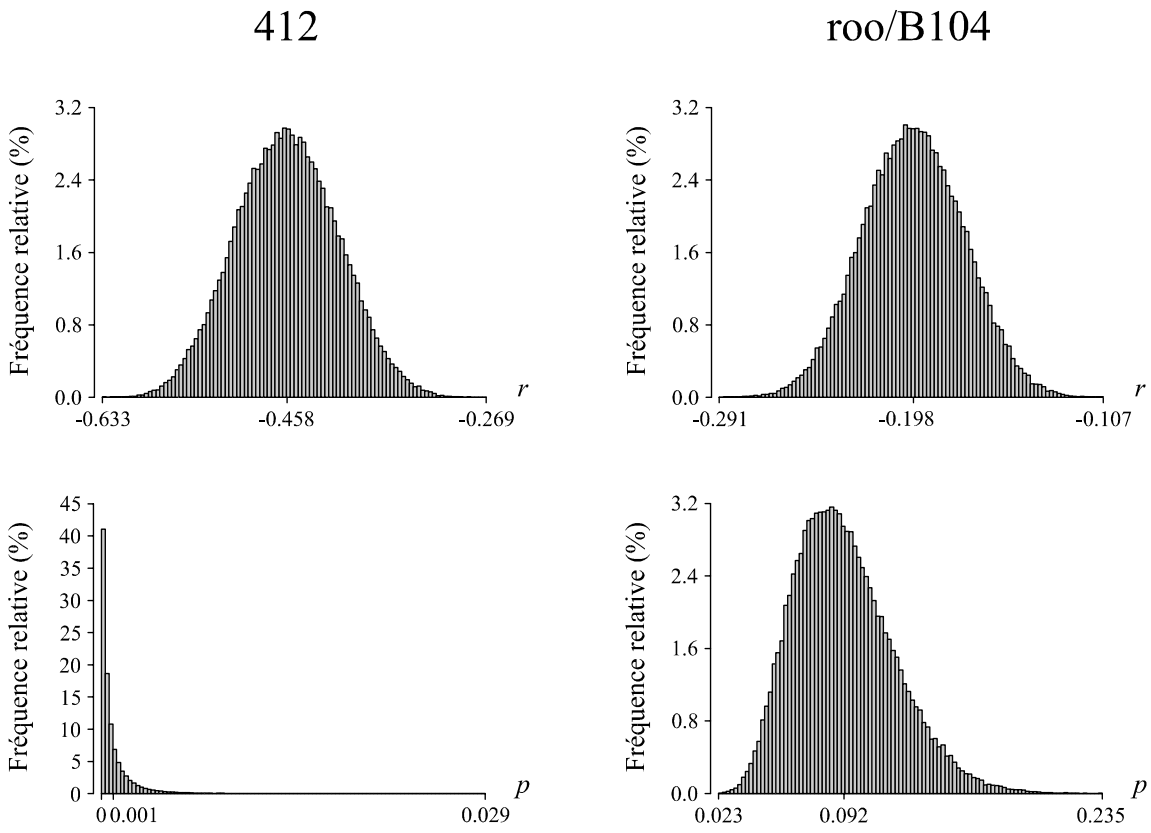
Les résultats des approches 1 & 2 montrent de façon évidente que la température minimale est corrélée négativement au nombre de copies de 412 tandis que roo/B104 ne montre pas de corrélation statistiquement significative (Tab. 9.11). La troisième approche montre également un résultat totalement différent pour 412 et roo/B104 (Fig. 9.9).

Approche	Rétrotransposon	n	r	p
1	412	50	-0.49	0.00018
2	412	99	-0.46	1.09×10^{-6}
1	roo/B104	48	-0.20	0.08340
2	roo/B104	93	-0.15	0.08267

Tableau 9.11: Corrélations entre la température minimale (X) et le nombre de copies des rétrotransposons 412 et roo/B104 (Y) selon l'approche suivie (voir le texte). r : valeur observée du coefficient de corrélation de Pearson. p : *p-value* du test de Student unilatéral.

Dans un second temps, il est nécessaire de considérer la nature régionalisée des variables puisque l'autocorrélation spatiale positive peut conduire à sous-estimer plus ou moins fortement la *p-value* du test de corrélation. Le cas de roo/B104 est sans ambiguïté dans la mesure où l'hypothèse nulle ne serait pas rejetée même si la *p-value* s'avérait sous-estimée. En revanche, dans le cas de 412, la conclusion statistique devrait être nuancée selon l'importance de la sous-estimation de la *p-value* et selon l'approche considérée.

Le test de Student modifié et le test de Monte-Carlo spatial sont difficilement utilisables notamment parce que l'hypothèse de stationnarité est intenable (gradient de température et cline du nombre de copies de 412), et parce que la covariance spatiale est très difficile à estimer à l'échelle planétaire (*i.e.*, sur le géoïde), avec une répartition des sites aussi irrégulière.



Figure˜9.9: Distribution statistique des valeurs observées du coefficient de corrélation de Pearson (r), et des *p-values* associées (p), dans le cas de la troisième approche (voir le texte). En abscisses : valeur minimale, moyenne, valeur maximale.

Dans Vieira *et al.* (1998) nous avons eu recours à des tests de Mantel en considérant uniquement l'autocorrélation spatiale en latitude. La proximité spatiale est calculée comme $Z_{ij} = 1 - h_{ij} / \max(h_{ij})$, avec h_{ij} l'écart en valeur absolue en latitude pour les sites s_i et s_j. Cette proximité varie linéairement et correspond à la variation spatiale selon un gradient ou un cline. Pour la température et le nombre de copies de 412, les autosimilarités sont calculées comme $X_{ij} = (x_i - \bar{x})(x_j - \bar{x})$ et $Y_{ij} = (y_i - \bar{y})(y_j - \bar{y})$. Toutes les valeurs sont considérées, ce qui conduit à des matrices approximativement 100×100. Les corrélations de Mantel sont évaluées par un test de randomisation (Section 3.1.4).

Le modèle de variation spatiale pour la température minimale et le nombre de copies de 412 se révèle très satisfaisant. En effet, les corrélations de Mantel s'élèvent respectivement à 0.40 et 0.33, ce qui correspond à une forte corrélation (Tab. 9.12). La corrélation de Mantel entre la température minimale et le nombre de copies de 412 est fortement réduite par la prise en compte de la latitude puisqu'elle passe de 0.20 pour le test de Mantel simple à 0.08 pour le test de Mantel partiel : la *p-value* augmente en conséquence. Conformément aux résultats de la Section 9.1.7.3, cette forte réduction de la sous-estimation de la *p-value* s'explique par l'excellence du modèle de variation spatiale considéré.

Afin d'apprécier la sensibilité du résultat à la présence de valeurs extrêmes, les calculs ont également été effectués en retirant les données pour trois sites dont les températures

minimales sont extrêmes (sites de Moscou, de Tchéquie et de *Grand Rapids*, aux USA). Dans ces conditions, la *p-value* du test de Mantel partiel augmente encore, mais reste à un niveau qui témoigne d'une corrélation statistiquement significative (Tab. 9.12).

Corrélation	$r^{(1)}$	$p^{(1)}$	$r^{(2)}$	$p^{(2)}$
X, Z	0.40	0.00001	0.39	0.00001
Y, Z	0.33	0.00001	0.34	0.00001
X, Y	0.20	0.00001	0.17	0.00009
$X, Y \mid Z$	0.08	0.00201	0.05	0.01955

Tableau 9.12: Corrélations de Mantel, simples et partielles. X : température minimale. Y : nombre de copies de 412. Z : latitude des sites. $r^{(1)}$ et $p^{(1)}$: corrélation et *p-value* du test de Mantel pour tous les sites. $r^{(2)}$ et $p^{(2)}$: corrélation et *p-value* du test de Mantel sans les trois sites pour lesquels la température minimale est extrême. Les *p-values* ont été obtenues à partir de 10^5 valeurs.

D'après cette étude de cas, l'utilisation du test de Mantel partiel semble résoudre le problème du test de la corrélation en présence de variables spatialement autocorrélées. La corrélation entre la température minimale et le nombre de copies de 412 est statistiquement et biologiquement significative, au moins pour les sites considérés. Cependant, il nous semble impossible d'obtenir une estimation fiable de la *p-value* associée au test de la corrélation en utilisant un test de Mantel partiel. Du reste, les Sections 9.1.7.3 & 9.1.7.4 remettent sérieusement en question la validité de cette approche.

Toutes les fois que c'est envisageable, il s'avère donc préférable de recourir au test de Student modifié ou au test de Monte-Carlo spatial plutôt que d'utiliser un test de Mantel partiel. L'étude de cas considérée ici représente en fait une des situations les moins favorables aux tests que nous recommandons, pour laquelle le test de Mantel partiel ne constitue qu'un pis-aller.

9.2 Association binaire

Si $x(\cdot)$ et $y(\cdot)$ sont deux VR binaires, le test de l'association pose le même type de problème que le test de la corrélation dans le cas des VR quantitatives. En effet, le test de l'indépendance à partir de la table de contingence 2×2 est affecté par la présence de l'autocorrélation spatiale (Fingleton 1983, Upton & Fingleton 1985, pp. 226-232). Ainsi, l'approche classique en écologie consistant à tester l'association de deux espèces en termes de présence/absence au moyen d'un test statistique classique (*e.g.*, Pielou 1969, Dutreix 1986) est invalide en présence d'autocorrélation spatiale. En conséquence, le test paramétrique d'une mesure d'association binaire (*e.g.*, le χ^2 de Pearson ou le G^2 de Wilks) doit être corrigé pour tenir compte de l'autocorrélation spatiale de $x(\cdot)$ et $y(\cdot)$. Un test de Monte-Carlo utilisant un modèle spatial peut également être envisagé.

9.2.1 Etude de Monte-Carlo

Afin d'illustrer l'erreur commise en testant l'association binaire comme si les données étaient spatialement indépendantes, nous avons procédé à une étude de Monte-Carlo en

utilisant des modèles géostatistiques. Les simulations sont exactement les mêmes que pour l'étude de Monte-Carlo menée dans le cas du coefficient de corrélation de Pearson (Section 9.1.1) :

- grille 30×30 de pas $\Delta = 1$,
- modèles exponentiel, gaussien, périodique et sphérique,
- paramétrage fixé à $\theta = (1, 7999, 10)$.

Une première réalisation de chaque modèle est simulée en fixant la même valeur à la graine du générateur de nombres pseudo-aléatoires de sorte que toutes les réalisations sont corrélées positivement deux à deux. Les réalisations $z(\cdot)$ sont recodées en indicatrices $i(\cdot)$ en utilisant le seuil arbitraire $z = 10$. Pour toute réalisation $z(\cdot)$ et tout support s le codage s'écrit :

$$i(s) = \begin{cases} 1 & \text{si } z(s) \geq 10 \\ 0 & \text{sinon} \end{cases} \tag{9.17}$$

Les indicatrices obtenues jouent le rôle des données binaires sous étude. Soit la table de contingence 2×2 établie à partir de deux VR binaires $x(\cdot)$ et $y(\cdot)$:

$\times$	1	0
1	a	b
0	c	d

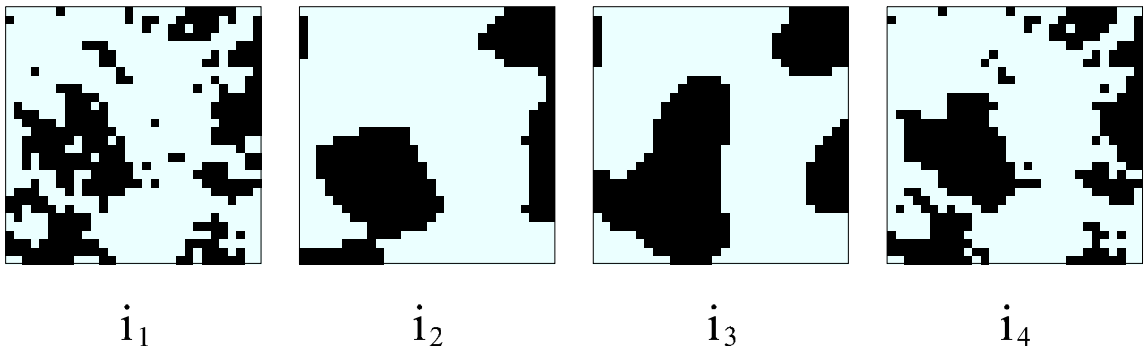
avec $a+b+c+d = n$. L'association binaire peut être mesurée par le χ^2 de Pearson estimé par (Dagnélie 1986, Sokal & Rohlf 1995) :

$$\chi^2_{obs} = \frac{(ad - bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

et distribué sous H_0 selon le χ^2 théorique à 1 degré de liberté dans le cas de données spatialement indépendantes. Nous n'utilisons par la correction de continuité de Yates dont l'intérêt pratique est encore discuté (*cf.* Dagnelie 1986, Sokal & Rohlf 1995). La *p-value* associée au test du χ^2 est estimée en simulant 10^5 réalisations du modèle d'une des deux VR $x(\cdot)$ et $y(\cdot)$. Les simulations pour lesquelles $ad = 0$ ne sont pas considérées. Une légère différence est attendue selon que la distribution empirique du χ^2 sous H_0 est obtenue en simulant le modèle spatial de $x(\cdot)$ ou celui de $y(\cdot)$. Afin d'apprécier cette différence, les simulations sont effectuées pour les deux variables. Le générateur de nombres pseudo-aléatoires étant toujours initialisé avec la même graine (Annexe B), la première réalisation redonne les données sous étude, et la *p-value* est calculée comme :

$$p = \frac{\text{Card}\left(\{\chi^2 \mid \chi^2 \in \Omega,\ \chi^2 \geq \chi^2_{obs}\}\right)}{\text{Card}\left(\Omega\right)}$$

avec Ω l'ensemble des valeurs du χ^2 et $\text{Card}(\Omega) = 10^5$. La *p-value* de référence (p_0) peut être comparée à celle obtenue par le test paramétrique de χ^2_{obs}. La distribution empirique du χ^2 sous H_0 est résumée par la valeur maximale (max), la moyenne ($\bar{\chi}^2$) et l'écart-type (σ_{χ^2}). Dans tous les cas la valeur minimale vaut zéro, *i.e.* pour $ad = bc \neq 0$. Pour les modèles exponentiel, gaussien, périodique et sphérique, les données sont désignées respectivement par i_1, i_2, i_3 et i_4 (Fig. 9.10).



Figure˜9.10: Indicatrices dérivées des populations simulées d'après des modèles de variogrammes : (i_1) exponentiel, (i_2) gaussien, (i_3) périodique, (i_4) sphérique.

Lorsque les deux VR $x(\cdot)$ et $y(\cdot)$ sont spatialement autocorrélées, la distribution empirique du χ^2 sous H_0 ne correspond pas à la distribution théorique pour $\nu = 1$ degré de liberté. En effet, l'espérance et la variance de la distribution théorique du χ^2 sont respectivement $\mathrm{E}_{\chi^2}(\nu) = \nu$ et $\mathrm{Var}_{\chi^2}(\nu) = 2\nu$ (*cf.* Aïvazian *et al.* 1986), ce qui n'est absolument pas compatible avec les résultats du Tableau 9.13. Outre que la distribution théorique du χ^2 se révèle globalement inadéquate, l'écart-type théorique pour $\nu = 1$ sous-estime considérablement les écarts-types réels σ_{χ^2}. En conséquence, le test paramétrique classique sous-estime très fortement la *p-value*. Ces résultats montrent qu'il est nécessaire de tenir compte de la dépendance spatiale lorsque les deux VR binaires sont autocorrélées, par exemple en utilisant un test paramétrique modifié.

$x(\cdot)$	$y(\cdot)$	max	$\bar{\chi}^2$	σ_{χ^2}	χ_{obs}	p_0
i_2	i_1	210.00691	12.06591	16.38604	93.74718	0.00386
i_1	i_2	182.59989	11.32707	15.21993	93.74718	0.00248
i_3	i_1	186.58040	12.18885	16.47164	48.79434	0.04309
i_1	i_3	173.68039	11.16943	14.78697	48.79434	0.03379
i_4	i_1	582.76244	8.02549	11.13914	582.76244	0.00001
i_1	i_4	582.76244	8.22146	11.32133	582.76244	0.00001
i_3	i_2	418.77959	24.34759	32.01740	278.11362	0.00016
i_2	i_3	400.43832	19.45357	25.90889	278.11362	0.00003
i_4	i_2	234.39733	15.67105	20.85010	196.93912	0.00007
i_2	i_4	244.11364	16.71601	22.31308	196.93912	0.00011
i_4	i_3	221.01483	15.74886	20.69597	102.38075	0.00704
i_3	i_4	255.84525	17.36069	23.27299	102.38075	0.01247

Tableau 9.13: Résultats de l'étude de Monte-Carlo concernant la statistique du χ^2 pour les couples formés par i_1, i_2, i_3 et i_4. max, $\bar{\chi}^2$, σ_{χ^2} : valeur maximale, moyenne et écart-type de la distribution empirique du χ^2 sous H_0. p_0 : *p-value* de référence. Les *p-values* calculées en utilisant la loi du χ^2 pour un degré de liberté sont inférieures à 10^{-5} dans tous les cas.

9.2.2 Test paramétrique modifié

Cerioli (1997) expose une version modifiée du test paramétrique de l'indépendance dans une table de contingence 2×2 dans le cas de données spatiales. Soit π_{ij} la probabilité d'observer sur un support s quelconque $x(s) = i$ et $y(s) = j$ $(i, j \in \{0, 1\})$, le vecteur de probabilité $\boldsymbol{\pi} = [\pi_{11}\ \pi_{10}\ \pi_{01}\ \pi_{00}]^{\mathrm{T}}$ est estimé par le vecteur des proportions d'échantillon $\widehat{\boldsymbol{\pi}} = [\widehat{\pi}_{11}\ \widehat{\pi}_{10}\ \widehat{\pi}_{01}\ \widehat{\pi}_{00}]^{\mathrm{T}}$ avec :

$$\begin{aligned}
\widehat{\pi}_{11} &= (a+b)(a+c)/n^2 \\
\widehat{\pi}_{10} &= (a+b)(b+d)/n^2 \\
\widehat{\pi}_{01} &= (c+d)(a+c)/n^2 \\
\widehat{\pi}_{00} &= (c+d)(b+d)/n^2
\end{aligned}$$

Cerioli (1997) propose de tester l'hypothèse nulle d'absence d'association entre $x(\cdot)$ et $y(\cdot)$ en considérant la statistique standardisée :

$$W = \frac{n\left[\ln(\widehat{\alpha})\right]^2}{\widehat{\sigma}^2}$$

avec

$$\widehat{\sigma}^2 = \frac{1}{\widehat{\pi}_{11}} + \frac{1}{\widehat{\pi}_{10}} + \frac{1}{\widehat{\pi}_{01}} + \frac{1}{\widehat{\pi}_{00}}$$

et α le *rapport du produit croisé* ou *odds ratio* (Agresti 1990) estimé par :

$$\widehat{\alpha} = \frac{\widehat{\pi}_{11}\widehat{\pi}_{00}}{\widehat{\pi}_{10}\widehat{\pi}_{01}}$$

Lorsque $x(\cdot)$ et $y(\cdot)$ sont sans autocorrélation spatiale, et pour n élevé, W suit sous H_0 la distribution du χ^2 théorique à 1 degré de liberté. En présence d'autocorrélation spatiale, Cerioli (1997) propose d'ajuster la statistique W en calculant :

$$\frac{W}{1 + \widehat{\lambda}_n}$$

avec

$$\widehat{\lambda}_n = \frac{2}{n\widehat{\pi}_{11}\widehat{\pi}_{00}} \sum N(h)\widehat{C}_X(h)\widehat{C}_Y(h) \tag{9.18}$$

ou bien

$$\widehat{\lambda}_n = \frac{2}{n} \sum N(h)\widehat{\rho}_X(h)\widehat{\rho}_Y(h) \tag{9.19}$$

où $\widehat{C}_X$ est une fonction de covariance et $\widehat{\rho}_X$ une fonction de corrélation. Cerioli (1997) utilise respectivement l'estimateur classique de la covariance spatiale (9.8) et le I de Moran (Section 3.2.2).

En effectuant exactement les mêmes simulations que dans l'étude de Monte-Carlo du χ^2, on obtient un ensemble de *p-values* de référence (p_0) que l'on peut comparer aux *p-values* obtenues en considérant le W ajusté et la distribution du χ^2 théorique à 1 degré de liberté (p_1). La covariance (9.8) est calculée pour un ensemble de 20 classes de pas constant $\Delta = 1$ et de tolérance $\varepsilon = 0.5$ et W est ajusté en utilisant l'expression (9.18).

Les résultats du Tableau 9.14 montrent un bon accord global entre les *p-values* de référence et celles estimées par la méthode proposée par Cerioli (1997). Des résultats identiques peuvent être obtenus en utilisant le I de Moran et l'expression (9.19).

$x(\cdot)$	$y(\cdot)$	max	$\bar{W}$	σ_W	W_{obs}	p_0	p_1
i_2	i_1	305.61787	12.68906	18.05445	90.00465	0.00812	0.00993
i_1	i_2	218.12740	11.79687	16.46908	90.00465	0.00533	0.00993
i_3	i_1	228.55424	12.71107	17.84829	47.83147	0.05172	0.05974
i_1	i_3	286.18469	11.58378	15.84773	47.83147	0.04121	0.05974
i_4	i_1	961.02474	8.25169	11.98091	961.02474	0.00001	0.00000
i_1	i_4	961.02474	8.45296	12.16536	961.02474	0.00001	0.00000
i_3	i_2	985.91328	26.83248	38.75964	319.61888	0.00086	0.00079
i_2	i_3	1050.83190	21.26853	31.05385	319.61888	0.00030	0.00079
i_4	i_2	295.46558	16.62297	23.31655	200.69210	0.00043	0.00130
i_2	i_4	477.10546	17.98390	25.72665	200.69210	0.00096	0.00130
i_4	i_3	464.33626	16.59501	22.86505	102.79471	0.01172	0.02159
i_3	i_4	383.03767	18.44408	26.14784	102.79471	0.01841	0.02159

Tableau 9.14: Résultats de l'étude de Monte-Carlo concernant la statistique W pour les couples formés par i_1, i_2, i_3 et i_4. max, $\bar{W}$, σ_W : valeur maximale, moyenne et écart-type de la distribution empirique de W sous H_0. p_0 : *p-value* de référence. Les *p-values* calculées en utilisant la loi du χ^2 pour un degré de liberté sont inférieures à 10^{-5} dans tous les cas.

Chapitre 10

Association spatiale

> *"The comparison of an observed pattern with a theoretically derived pattern is of considerable importance since it is usually undertaken in order to verify how closely the model from which the theoretical pattern was obtained conforms to reality."* (Cliff 1970)
>
> *"Il conviendra ensuite de comparer ces images à des cartes réelles* [...] *une comparaison pixel à pixel n'a guère de sens. Car il s'agit de modéliser une organisation spatiale et non de retrouver les détails de la réalité aux mêmes localisations, à moins qu'elles ne soient fondamentales."* (Chiarello 1994)
>
> *"Finalement, la forme se réduit à un agrégat de carrés. C'est bien entendu une façon très barbare de représenter les formes."* (Thom 1991)

La comparaison entre deux variables régionalisées $f(\cdot)$ et $g(\cdot)$ définies sur un même domaine D apparaît dans de nombreuses problématiques écologiques. Il convient tout d'abord de distinguer deux situations selon que $f(\cdot)$ et $g(\cdot)$ correspondent à un même phénomène ou à deux phénomènes différents.

Dans la première situation, $f(\cdot)$ et $g(\cdot)$ peuvent correspondre à des mesures effectuées à des dates différentes. L'objectif consiste alors à apprécier la stabilité de la structure spatiale du phénomène, qu'il s'agisse de la répartition spatiale de populations (*e.g.*, Legendre *et al.* 1997, Gerhards *et al.* 1997) ou de la structure spatiale de la différentiation génétique au sein d'une population (*e.g.*, Hossaert-McKey *et al.* 1996). Dans le cadre de la modélisation des processus qui génèrent les structures spatiales observées, une approche classique consiste à comparer la sortie d'un modèle $\mathcal{M}$ et une image de la réalité, afin de tester la plausibilité de $\mathcal{M}$ (*e.g.*, Li *et al.* 1993, Chiarello 1994, Heil & van Deursen 1996, Hill *et al.* 1998).

Dans la seconde situation, il s'agit généralement de comparer les répartitions spatiales de plusieurs espèces (*e.g.*, Kershaw 1961, Fisher 1968, Pielou 1969, pp. 159-171, Lieth & Moore 1971, Crovello 1981, Brunel 1986, Dutreix 1986, Birks 1987, MacDonald & Waters 1988, Rossi *et al.* 1992, Manly 1995, Delaville *et al.* 1996, Rossi *et al.* 1996, Legendre *et al.* 1997, Jonsson & Moen 1998), celles des écophases d'une même espèce (*e.g.*, Cardina *et al.* 1996, Legendre *et al.* 1997), les structures spatiales de deux variables environnementales (*e.g.*, Söderström & Eriksson 1996), la variation géographique d'une espèce et la structure spatiale d'une variable environnementale (*e.g.*, Cesaroni *et al.* 1997)

ou l'association entre la répartition spatiale d'un groupe d'espèces et un ensemble de variables environnementales (*e.g.*, Strahler 1978, ter Braak 1987, Lebreton *et al.* 1988a, 1988b, Hill 1991, Borcard *et al.* 1992, ter Braak & Juggins 1993, Chessel & Mercier 1993, Franquet & Chessel 1994, Borcard & Legendre 1994, Zhang 1994, Prodon & Lebreton 1994, Diniz-Filho & Bini 1996, Rodríguez & Lewis 1997, Adjeroud 1997, Fariña *et al.* 1997, Méot *et al.* 1998, etc.).

Les objectifs qui motivent l'étude conjointe de deux ou plusieurs VR sont donc très variés, mais les méthodes statistiques utilisées sont généralement classiques et rarement adaptées à la nature spatiale des données (Switzer 1983). Par exemple, pour tester l'association et la stabilité de la répartition spatiale de juvéniles et d'adultes de deux bivalves (*Macomona liliana* et *Austrovenus stutchburyi*), Legendre *et al.* (1997) utilisent le t de Student apparié et le coefficient de corrélation de Pearson r. Cesaroni *et al.* (1997) calculent r entre des grilles de valeurs interpolées par krigeage en FAI-k concernant, d'une part des caractéristiques morphologiques et génétiques des criquets cavernicoles *Dolichopoda laetitiae* et *D. geniculata*, et d'autre part des variables environmentales. Cesaroni *et al.* (1997) utilisent également des tests de Mantel entre matrices de distances afin de confirmer les résultats obtenus avec r. Le coefficient de corrélation r est également utilisé par Birrell *et al.* (1996) afin de comparer des cartes de rendement de récoltes, par Schlesinger *et al.* (1996) dans le cas des nutriments présents dans le sol d'écosystèmes désertiques, et par Cannavacciuolo *et al.* (1998) dans la comparaison de la distribution spatiale de la biomasse des adultes de deux espèces de lombrics (*Lumbricus terrestris* et *Aporrectodea caliginosa*).

Afin de tester si deux ensembles de valeurs diffèrent uniquement à cause d'erreurs de mesure ou d'échantillonnage, Legendre & McArdle (1997) proposent plusieurs modèles d'ANOVA et d'ANCOVA, en fonction des dispositifs d'échantillonnage utilisés, et de la présence ou de l'absence de réplication. Cependant, dans leur comparaison de deux campagnes d'échantillonnage du macrobenthos du lac Erie, Minns *et al.* (1996) concluent que les différences significatives mises en évidence par l'ANOVA et l'ANCOVA sont largement dues au fait que les hypothèses statistiques sous-jacentes ne sont pas satisfaites. Söderström & Eriksson (1996) utilisent un test de Kruskal-Wallis afin d'étudier les différences de concentration de cadmium dans le sol selon le type de roche sous-jacente.

Dans le cas des variables qualitatives, le recours aux tables de contingence et aux statistiques associées est assez général, qu'il s'agisse de comparer des cartes de présence/absence d'espèces, réelles (Dutreix 1986) ou prédites (Fielding & Bell 1997), des cartes de différentes variables environnementales (Davis & Dozier 1990), ou les deux types de cartes entre eux (Debinski & Humphrey 1997), d'étudier le déterminisme d'une épidémie (Lannou & Savary 1991), ou de comparer le résultat d'une simulation et une image de la réalité (Li *et al.* 1993).

Enfin, la comparaison des répartitions d'espèces s'effectue classiquement au moyen des méthodes multivariées (Fisher 1968, Dutreix 1986, Debinski & Humphrey 1997).

Le problème est que les méthodes statistiques classiques qui viennent d'être évoquées n'exploitent pas le caractère régionalisé des VR. Bien au contraire, l'autocorrélation spatiale des VR constitue souvent un paramètre de nuisance pour ces statistiques. Cet aspect a été illustré dans le Chapitre 9 consacré à l'étude de l'impact de l'autocorrélation dans le test de la corrélation (variables quantitatives) ou de l'association (variables binaires). Dans ce contexte, l'approche généralement proposée consiste à aborder le problème dès

l'échantillonnage en garantissant que les supports sont espacés d'une distance au moins égale à la portée de l'autocorrélation. Ce type de solution est notamment suggégé dans l'étude des associations d'espèces par Jonsson & Moen (1998) afin de s'assurer que les co-occurrences des espèces reflètent uniquement les interactions inter-spécifiques, et pas les interactions au sein de chaque espèce, autrement dit, en se débarrassant d'emblée de l'autocorrélation spatiale. Selon un point de vue similaire, Belgrano *et al.* (1995a) considèrent que l'existence d'une structure spatiale partagée par les espèces et les variables environnementales peut entraîner une surestimation de l'interaction espèces/variables[1].

Dans ce chapitre, nous adoptons un point de vue radicalement différent du précédent en ce sens que la structure d'autocorrélation spatiale n'est pas considérée comme un paramètre de nuisance mais comme une information essentielle, qu'il faut exploiter lors de l'analyse. Il convient d'abord de distinguer la corrélation — au sens du coefficient de Pearson par exemple — de l'association spatiale, bien que ces concepts soient souvent confondus (*e.g.*, Besag & Clifford 1989, Dutilleul 1993, Koenig 1999). En effet, toutes les mesures de corrélation ou d'association classiquement utilisées en écologie statistique ne tiennent pas compte du caractère régionalisé des variables, et sont donc fondamentalement *a-spatiales*. A l'opposé, nous désignons sous le terme générique de mesure d'*association spatiale* toute définition opératoire de l'association entre VR qui fait explicitement intervenir la position relative des supports des valeurs[2].

Dans ce qui suit, nous examinons successivement la notion d'association spatiale globale de deux VR, puis la notion de corégionalisation définie par la géostatistique et les fonctions structurales croisées qui en découlent, et enfin l'association spatiale entre cartes binaires et quantitatives.

10.1 Association spatiale globale

Considérons deux variables régionalisées $f(\cdot)$ et $g(\cdot)$ définies dans D sur un même ensemble de supports $s = \{s_i \mid i = 1, \ldots, n\}$. L'association a-spatiale entre $f(\cdot)$ et $g(\cdot)$ est calculée à partir des couples de valeurs $\{f(s_i), g(s_i)\}$ pour $i = 1, \ldots, n$, au moyen d'une statistique classique, *e.g.* le r de Pearson, le ρ de Spearman ou le τ de Kendall (Sokal & Rohlf 1995). Dans cette approche, la structure spatiale de $f(\cdot)$ et de $f(\cdot)$ n'intervient pas[3]. L'association a-spatiale présente deux inconvénients majeurs :

- elle ne peut être utilisée que lorsque la correspondance des valeurs support-à-support a un sens, ce qui n'est pas le cas lorsqu'on s'intéresse à la ressemblance entre des types d'organisations spatiales, que ce soit dans le cadre de la comparaison de la sortie d'un modèle avec des données de terrain (*e.g.*, Chiarello 1994) ou dans l'étude des associations d'espèces (Section 10.3),
- il est impossible de savoir si l'association correspond à un processus spatial ou pas (Hubert *et al.* 1985).

Une approche différente consiste à étudier l'association entre $f(\cdot)$ et $g(\cdot)$ en considérant les couples de supports (s_i, s_j) formés pour $i \neq j = 1, \ldots, n$, afin de déterminer si

[1] Les mêmes auteurs (Belgrano *et al.* 1995b) montrent néanmoins que la répartition spatiale des espèces peut être complètement expliquée par la structure spatiale des variables environnementales.

[2] La différence entre les associations spatiale et a-spatiale est illustrée dans Hubert *et al.* (1985).

[3] La structure spatiale intervient uniquement lors du test de la statistique observée.

la covariation des valeurs des deux VR est spatialement structurée. Lorsque l'association spatiale est appréciée à l'aide d'une statistique unique, *i.e.* calculée à partir de toutes les données, nous parlerons d'*association spatiale globale*. A notre connaissance, une des toutes premières tentatives pour mesurer l'association spatiale globale est celle de Tjøstheim (1978). Par la suite, deux articles importants dus à Hubert & Golledge (1982) et Hubert *et al.* (1985), ont permis :

- d'éclaircir la nature de la statistique proposée par Tjøstheim (1978),
- de dériver une mesure d'association spatiale globale qui n'est pas autre chose que la statistique de Mantel, généralisée au cas de deux variables.

10.1.1 Statistique de Tjøstheim

Considérons deux variables régionalisées $f(\cdot)$ et $g(\cdot)$ prenant leurs valeurs dans D sur des supports ponctuels $s_i = (x_i, y_i)$ pour $i = 1, \ldots, n$. Soit $c(u)$ une fonction définie par :

$$c(u) = \begin{cases} 1 & \text{si } u > 0 \\ \frac{1}{2} & \text{si } u = 0 \\ 0 & \text{si } u < 0 \end{cases} \tag{10.1}$$

Sans considérer le cas des *ex aequo*, le rang $R_f(x_i, y_i)$ de $f(\cdot)$ au point (x_i, y_i) peut être défini formellement comme (Tjøstheim 1978) :

$$R_f(x_i, y_i) = \sum_{j=1}^{n} c\{f(x_i, y_i) - f(x_j, y_j)\} \tag{10.2}$$

et identiquement pour $g(\cdot)$. Notons $x_f(i)$ l'abscisse du support de la valeur de $f(\cdot)$ de rang i et $\delta(\alpha, \beta)$ le delta de Kronecker tel que :

$$\delta(\alpha, \beta) = \begin{cases} 1 & \text{si } \alpha = \beta \\ 0 & \text{sinon} \end{cases} \tag{10.3}$$

La fonction $x_f(i)$ peut être définie formellement comme (Tjøstheim 1978) :

$$x_f(i) = \sum_{j=1}^{n} x_j \cdot \delta\{i, R_f(x_j, y_j)\} \tag{10.4}$$

Les fonctions $y_f(i)$, $x_g(i)$ et $y_g(i)$ sont définies comme en (10.4). Soit $d\{(x_i, y_i); (x_j, y_j)\}$ une fonction mesurant la proximité entre les points s_i et s_j, la statistique de Tjøstheim s'écrit (Tjøstheim 1978) :

$$\Lambda = \sum_{i=1}^{n} d[\{x_f(i), y_f(i)\}; \{x_g(i), y_g(i)\}] \tag{10.5}$$

autrement dit, la statistique (10.5) cumule les proximités spatiales entre valeurs de $f(\cdot)$ et de $g(\cdot)$ de rangs identiques. Alors que les mesures d'association a-spatiale telles que le ρ de Spearman ou le τ de Kendall mesurent la similarité entre les rangs des deux variables aux mêmes positions, la statistique de Tjøstheim mesure la similarité entre les positions des supports, pour des rangs identiques (Hubert & Golledge 1982).

En utilisant une distance $d(\cdot,\cdot)$ dans la définition (10.5), l'association spatiale entre $f(\cdot)$ et $g(\cdot)$ au sens de Λ est positive lorsque les supports des données de rangs identiques sont proches, *i.e.* lorsque Λ est petit, et négative dans le cas contraire, *i.e.* lorsque Λ est grand. Si $d(\cdot,\cdot)$ est le carré de la distance euclidienne, la statistique (10.5) peut s'écrire :

$$\Lambda = \sum_{i=1}^{n} \left[\{x_f(i) - x_g(i)\}^2 + \{y_f(i) - y_g(i)\}^2 \right] \tag{10.6}$$

Pour un ensemble fixé de points on a $\sum x_f^2(i) = \sum x_g^2(i) = \sum x_i^2$, et de même pour les ordonnées. En conséquence, la partie "active" de (10.6) est (Tjøstheim 1978, Hubert & Golledge 1982) :

$$\sum_{i=1}^{n} \{x_f(i) \cdot x_g(i) + y_f(i) \cdot y_g(i)\} \tag{10.7}$$

La proposition de Tjøstheim est importante comme point de départ d'une réflexion sur l'association spatiale, mais elle présente au moins trois faiblesses.

Premièrement, la statistique (10.5) n'exploite pas réellement la structure spatiale des VR. Par exemple, considérons que $f(\cdot)$ et $g(\cdot)$ sont spatialement structurées, associées support-à-support, et présentent une valeur Λ_{obs}. En randomisant en parallèle les couples de valeurs $\{f(s_i), g(s_i)\}$ afin de préserver l'association support-à-support, il est possible d'obtenir une valeur $\Lambda < \Lambda_{obs}$ traduisant une plus grande association spatiale, alors même que la structure spatiale de $f(\cdot)$ et $g(\cdot)$ a été détruite par la randomisation. Le moins que l'on puisse dire c'est que ce type de résultat ne s'accorde pas avec l'idée que l'on peut se faire d'une mesure d'association spatiale.

Deuxièmement, le traitement des variables quantitatives ne peut pas s'effectuer autrement qu'en les dégradant sous la forme de variables ordinales, ce qui s'accompagne d'une perte d'information *a priori* préjudiciable à l'analyse.

Enfin, le test de la valeur observée présente une anomalie dans la mesure où le modèle d'inférence est spatialement "contaminé" par une statistique d'association a-spatiale (Hubert *et al.* 1985). Il s'avère préférable d'évaluer l'association spatiale *per se*, conditionnellement à un certain degré d'association support-à-support (Hubert *et al.* 1985).

10.1.2 Statistique de Mantel généralisée

En notant d_{ij} la distance $d[\{x_f(i), y_f(i)\}; \{x_g(j), y_g(j)\}]$, la statistique de Tjøstheim (10.5) peut s'écrire (Hubert & Golledge 1982) :

$$\Lambda = \sum_{i=1}^{n} d_{ii} \tag{10.8}$$

En introduisant une matrice $\mathbf{C}$ d'éléments c_{ij} tels que :

$$c_{ij} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{sinon} \end{cases} \tag{10.9}$$

et en considérant la matrice des distances $\mathbf{D} = ((d_{ij}))$, la statistique (10.8) peut également s'écrire (Hubert & Golledge 1982) :

$$\Lambda = \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} d_{ij} \tag{10.10}$$

Sans se restreindre aux rangs des valeurs des variables, il est possible de considérer la distance $d_{ij} = d(s_i, s_j)$ et de généraliser la forme de c_{ij} à une fonction quelconque des valeurs $f = \{f_i \mid i = 1, \ldots, n\}$ et $g = \{g_i \mid i = 1, \ldots, n\}$, soit $c_{ij} = \varphi(f_i, g_j)$. La statistique de Tjøstheim (10.5) apparaît donc comme un cas particulier de la forme générale (10.10) où les valeurs de f et de g sont des rangs et où $\varphi(f_i, g_j) = 1$ si $f_i = g_j$ et $\varphi(f_i, g_j) = 0$ si $f_i \neq g_j$ (Hubert & Golledge 1982).

En fait, la statistique (10.10) n'est pas autre chose qu'une généralisation de la statistique de Mantel — définie pour des matrices pas nécessairement symétriques — dans laquelle la matrice $\mathbf{C}$ de similarité (ou de dissimilarité) entre valeurs concerne deux variables au lieu d'une seule. Dans ce qui suit, nous considérons que $\mathbf{C}$ et $\mathbf{D}$ sont symétriques, ce qui permet d'écrire la statistique (10.10) sous la même forme que la statistique de Mantel de la Section 3.1.4 (p. 39) :

$$\Lambda = \sum_{i<j} c_{ij} d_{ij} \tag{10.11}$$

Soit la définition opératoire de la statistique de Mantel donnée dans la Section 3.5.1 (p. 58). La matrice symétrique $\mathbf{D}$ peut être définie à partir de la proximité spatiale $d_{ij} = 1/h_{ij}$ avec h_{ij} la distance euclidienne entre les supports s_i et s_j. Pour une variable quantitative, nous proposons de définir la matrice symétrique $\mathbf{C}$ à partir de la covariance moyenne entre valeurs $c_{ij} = \left[\left(f_i - \bar{f}\right)(g_j - \bar{g}) + \left(f_j - \bar{f}\right)(g_i - \bar{g})\right]/2$, avec $\bar{f}$ et $\bar{g}$ les moyennes arithmétiques respectives des valeurs de f et de g. Dans ce contexte, la *statistique de Mantel généralisée* (10.11) permet de tester si la covariation entre les valeurs de f et de g exprimée sous la forme de la matrice $\mathbf{C}$ est spatialement structurée, au sens de la matrice de proximité spatiale $\mathbf{D}$.

La mesure d'association spatiale globale (10.11) apparaît donc finalement comme une forme dérivée de la mesure d'autocorrélation spatiale globale (Hubert *et al.* 1985), et comme telle, elle pose les mêmes types de problèmes d'interprétation que ceux évoqués dans la Section 3.5.1.

10.2 Corégionalisation

La géostatistique étend la notion de régionalisation définie pour une seule variable régionalisée $z(\cdot)$ à un ensemble de K variables $\{z_k(\cdot) \mid k = 1, \ldots, K\}$, autrement dit, à une multivariable régionalisée (MVR). La modélisation probabiliste des MVR est similaire à celle d'une seule VR au sens où la MVR est vue comme une réalisation particulière d'un ensemble de K fonctions aléatoires $\{Z_k(\cdot) \mid k = 1, \ldots, K\}$ (Journel & Huijbregts 1978). Dans ce qui suit, nous nous contentons de traiter du cas bivarié $K = 2$ et de la description de la corégionalisation au moyen de fonctions structurales croisées.

10.2.1 Isotopie vs. hétérotopie

Considérons deux variables régionalisées $z_1(\cdot)$ et $z_2(\cdot)$ définies dans D sur des ensembles de supports respectifs $s_1 = \{s_{1i} \mid i = 1, \ldots, n_1\}$ et $s_2 = \{s_{2i} \mid i = 1, \ldots, n_2\}$. L'association spatiale entre les valeurs $z_1 = \{z_{1i} \mid i = 1, \ldots, n_1\}$ et $z_2 = \{z_{2i} \mid i = 1, \ldots, n_2\}$ nécessite, par définition, de tenir compte de l'information concernant les ensembles de supports s_1 et s_2. Il est possible de définir les situations (Wackernagel 1993) :

- d'*isotopie*, lorsque les deux VR sont mesurées ou observées sur n supports identiques, soit $s_{1i} = s_{2i}$ pour $i = 1, \ldots, n$,
- d'*hétérotopie totale*, lorsque s_1 et s_2 sont disjoints, soit $s_1 \cap s_2 = \emptyset$,
- d'*hétérotopie partielle*, lorsque $s_1 \neq s_2$ et $s_1 \cap s_2 \neq \emptyset$, avec les cas particuliers $s_1 \subset s_2$ ou $s_2 \subset s_1$.

La distinction de ces trois situations est importante dans la mesure où elle conditionne le type de fonction structurale croisée qui peut être employé pour décrire l'association spatiale entre $z_1(\cdot)$ et $z_2(\cdot)$.

10.2.2 Covariance croisée

Considérons deux FAST-2 $Z_1(\cdot)$ et $Z_2(\cdot)$ d'espérances $\mathrm{E}[Z_1(x)] = m_1$ et $\mathrm{E}[Z_2(x)] = m_2$ pour tout $x \in D$. La *covariance croisée théorique* se définit comme (Journel & Huijbregts 1978, Wackernagel 1993, Goovaerts 1997) :

$$C_{12}(h) = \mathrm{E}[\{Z_1(x) - m_1\}\{Z_2(x+\mathbf{h}) - m_2\}] = \mathrm{E}[Z_1(x)\, Z_2(x+\mathbf{h})] - m_1 m_2 \quad (10.12)$$

La fonction (10.12) n'est pas nécessairement symétrique puisqu'il est possible d'avoir $C_{12}(\mathbf{h}) \neq C_{21}(\mathbf{h})$, et elle n'est pas non plus nécessairement paire puisqu'en général $C_{12}(\mathbf{h}) \neq C_{12}(-\mathbf{h})$ (Wackernagel 1993). En revanche, on a toujours $C_{12}(\mathbf{h}) = C_{21}(-\mathbf{h})$ (Journel & Huijbregts 1978, Wackernagel 1993, Goovaerts 1997).

Une propriété intéressante de la covariance croisée est qu'elle ne nécessite pas de connaître les VR $z_1(\cdot)$ et $z_2(\cdot)$ pour les mêmes supports $x \in D$, autrement dit, cette fonction peut être utilisée en situation d'hétérotopie totale.

10.2.2.1 Estimateurs

La fonction de *covariance croisée expérimentale* peut se calculer classiquement selon (Wackernagel 1993, p. 54) :

$$\widehat{C}_{12}(\mathbf{h}) = \frac{1}{N(\mathbf{h})} \sum_{(i,j)|\mathsf{h}_{ij}=\mathsf{h}} (z_{1i} - \bar{z}_1)(z_{2j} - \bar{z}_2) \quad (10.13)$$

avec $\bar{z}_1$ et $\bar{z}_2$ les moyennes arithmétiques respectives des valeurs z_1 et z_2, $\mathbf{h}_{ij}$ le vecteur dont l'origine est le support de z_{1i} et dont l'extrémité est le support de z_{2j} et $N(\mathbf{h})$ le

nombre de couples tels que $\mathbf{h}_{ij} = \mathbf{h}$. L'estimateur de type *non ergodique* s'écrit (Isaaks & Srivastava 1989, p. 62, Goovaerts 1997, p. 46) :

$$\widehat{C}_{12}(\mathbf{h}) = \frac{1}{N(\mathbf{h})} \sum_{(i,j)|\mathsf{h}_{ij}=\mathsf{h}} \left(z_{1i} - m_{1_{-\mathsf{h}}}\right)\left(z_{2j} - m_{2_{+\mathsf{h}}}\right) \tag{10.14}$$

avec $m_{1_{-h}}$ la moyenne de toutes les valeurs de z_1 dont les supports sont situés à $-\mathbf{h}$ des supports de z_2 :

$$m_{1_{-\mathsf{h}}} = \frac{1}{N(\mathbf{h})} \sum_{i|\mathsf{h}_{ij}=\mathsf{h}} z_{1i} \tag{10.15}$$

et $m_{2_{+\mathsf{h}}}$ la moyenne de toutes les valeurs de z_2 dont les supports sont situés à $+\mathbf{h}$ des supports de z_1 :

$$m_{2_{+\mathsf{h}}} = \frac{1}{N(\mathbf{h})} \sum_{j|\mathsf{h}_{ij}=\mathsf{h}} z_{2j} \tag{10.16}$$

10.2.2.2 Effet de retard

Une différence substantielle entre $C_{12}(\mathbf{h})$ et $C_{12}(-\mathbf{h})$ signifie que l'une des variables est décalée par rapport à l'autre, cet effet étant connu sous le nom d'*effet de retard* (*lag effect*) (Journel & Huijbregts 1978, Wackernagel 1993, Goovaerts 1997). Un effet de retard se traduit par un décalage du maximum de la covariance croisée de l'origine vers $\mathbf{h} \neq \mathbf{0}$ (Wackernagel 1993, Goovaerts 1997). Le plus souvent, les différences entre directions opposées $\mathbf{h}$ et $-\mathbf{h}$ résultent du petit nombre de couples $N(\mathbf{h})$ disponibles et reflètent donc essentiellement des fluctuations d'échantillonnage (Goovaerts 1997).

10.2.3 Variogramme croisé

Considérons deux FAI-0 $Z_1(\cdot)$ et $Z_2(\cdot)$ d'incréments $Z_1(x+\mathbf{h}) - Z_1(x)$ et $Z_1(x+\mathbf{h}) - Z_1(x)$, dont les espérances sont nulles pour tout $x \in D$; le *variogramme croisé théorique* est défini comme (Journel & Huijbregts 1978, Wackernagel 1993, Goovaerts 1997) :

$$\gamma_{12}(\mathbf{h}) = \frac{1}{2} \mathrm{E}\left[\{Z_1(x+\mathbf{h}) - Z_1(x)\}\{Z_2(x+\mathbf{h}) - Z_2(x)\}\right] \tag{10.17}$$

La fonction (10.17) est symétrique puisque $\gamma_{12}(\mathbf{h}) = \gamma_{21}(\mathbf{h})$ (Journel & Huijbregts 1978). Le variogramme croisé peut prendre des valeurs négatives lorsqu'un accroissement pour l'une des VR correspond à une décroissance pour l'autre (Journel & Huijbregts 1978). En écologie, cette association spatiale négative peut se rencontrer notamment dans le cas d'une compétition entre deux espèces qui se partagent l'espace en s'excluant mutuellement.

Sous l'hypothèse de stationnarité d'ordre 2, il est possible de relier la fonction (10.12) et la fonction (10.17) selon (Journel & Huijbregts 1978, Wackernagel 1993, Goovaerts 1997) :

$$\gamma_{12}(\mathbf{h}) = C_{12}(\mathbf{0}) - \frac{1}{2}\left[C_{12}(\mathbf{h}) + C_{12}(-\mathbf{h})\right] \tag{10.18}$$

En remarquant que la fonction de covariance croisée (10.12) peut s'écrire sous la forme (Wackernagel 1993, Goovaerts 1997) :

$$C_{12}(\mathbf{h}) = \frac{1}{2}\underbrace{[C_{12}(\mathbf{h}) + C_{12}(-\mathbf{h})]}_{\text{terme pair}} + \frac{1}{2}\underbrace{[C_{12}(\mathbf{h}) - C_{12}(-\mathbf{h})]}_{\text{terme impair}} \tag{10.19}$$

on constate que $\gamma_{12}(\mathbf{h})$ incorpore uniquement le terme pair de la covariance croisée $C_{12}(\mathbf{h})$, et par conséquent, $\gamma_{12}(\mathbf{h})$ est une fonction paire ($\gamma_{12}(\mathbf{h}) = \gamma_{12}(-\mathbf{h})$).

Du fait de sa définition (10.17), et de la parité qui en découle, le variogramme croisé présente deux inconvénients majeurs par rapport à la covariance croisée :

- il nécessite l'isotopie,
- il ne permet pas de mettre en évidence un éventuel effet retard.

10.2.3.1 Estimateur

Le *variogramme croisé expérimental* peut se calculer selon (Isaaks & Srivastava 1989, p. 64, Wackernagel 1993, p. 54) :

$$\widehat{\gamma}_{12}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{(i,j)|\mathsf{h}_{ij}=\mathsf{h}} (z_{1j} - z_{1i})(z_{2j} - z_{2i}) \tag{10.20}$$

avec les mêmes notations que dans la Section 10.2.2.1. Le calcul de $\widehat{\gamma}_{12}(\mathbf{h})$ nécessite l'isotopie : dans le cas d'une situation d'hétérotopie partielle, il faut donc se restreindre à l'ensemble des supports communs $s = s_1 \cap s_2$, ce qui peut représenter une perte d'information importante.

10.2.4 Pseudo-variogramme croisé

Afin d'obtenir une fonction structurale croisée qui ressemble au variogramme croisé tout en étant applicable en situation d'hétérotopie, un *pseudo-variogramme croisé théorique* peut être défini de façon générale comme (Myers 1991b, Papritz *et al.* 1993) :

$$\gamma_{12}^{p}(\mathbf{h}) = \frac{1}{2}\operatorname{Var}[Z_1(x+\mathbf{h}) - Z_2(x)] \tag{10.21}$$

et si les FA $Z_1(\cdot)$ et $Z_2(\cdot)$ ont même espérance ($m_1 = m_2$), le pseudo-variogramme croisé (10.21) peut s'écrire :

$$\gamma_{12}^{c}(\mathbf{h}) = \frac{1}{2}\operatorname{E}\left[\{Z_1(x+\mathbf{h}) - Z_2(x)\}^2\right] \tag{10.22}$$

Dans le cas général où $m_1 \neq m_2$, les fonctions (10.21) et (10.22) sont liées par la relation (Myers 1991b) :

$$\gamma_{12}^{c}(\mathbf{h}) = \gamma_{12}^{p}(\mathbf{h}) + \frac{1}{2}(m_1 - m_2)^2 \tag{10.23}$$

Contrairement au variogramme croisé (10.17), mais de la même façon que la covariance croisée (10.12), la fonction (10.21) n'est pas nécessairement une fonction paire puisqu'en général $\gamma_{12}^{p}(\mathbf{h}) \neq \gamma_{12}^{p}(-\mathbf{h})$, et elle n'est pas non plus nécessairement symétrique puisqu'il est possible d'avoir $\gamma_{12}^{p}(\mathbf{h}) \neq \gamma_{21}^{p}(\mathbf{h})$ (Papritz *et al.* 1993). En revanche, de même que pour la covariance croisée et le variogramme croisé, on a toujours $\gamma_{12}^{p}(\mathbf{h}) = \gamma_{21}^{p}(-\mathbf{h})$ (Myers 1991b, Papritz *et al.* 1993).

Si $Z_1(\cdot)$ et $Z_2(\cdot)$ sont deux FAST-2 de covariances respectives $C_{11}(\mathbf{h})$ et $C_{22}(\mathbf{h})$ et de covariance croisée $C_{12}(\mathbf{h})$, il existe la relation (Myers 1991b, Wackernagel 1993, Papritz *et al.* 1993) :

$$\gamma_{12}^{p}(\mathbf{h}) = \frac{1}{2}\left[C_{11}(\mathbf{0}) + C_{22}(\mathbf{0})\right] - C_{12}(\mathbf{h}) \tag{10.24}$$

D'après Papritz *et al.* (1993), la relation (10.24) montre que :

- le pseudo-variogramme croisé peut servir à décrire l'association spatiale négative[4] et détecter des effets de retard au même titre que la covariance croisée,
- le "seuil" du pseudo-variogramme croisé est égal à la moyenne des seuils des deux autovariogrammes $\gamma_{11}(\mathbf{h})$ et $\gamma_{22}(\mathbf{h})$ — respectivement $C_{11}(\mathbf{0})$ et $C_{22}(\mathbf{0})$ — puisque pour des variogrammes bornés on a :

$$\lim_{\mathsf{h}\to\infty} C_{12}(\mathbf{h}) = \mathbf{0} \tag{10.25}$$

- si $C_{12}(\mathbf{h}) < 0$ et croît de façon monotone vers 0 lorsque la distance augmente, alors $\gamma_{12}^{p}(\mathbf{h})$ est maximal en $\mathbf{h} = \mathbf{0}$ puis décroît vers son "seuil",
- le pseudo-variogramme croisé étant composé de variances, on a toujours $\gamma_{12}^{p}(\mathbf{h}) \geq 0$. Alors que dans le cas du variogramme croisé $\gamma_{12}(\mathbf{0}) = 0$, dans le cas du pseudo-variogramme croisé, $\gamma_{12}^{p}(\mathbf{0}) = 0$ n'est obtenu que si $Z_1(\cdot)$ et $Z_2(\cdot)$ présentent une corrélation positive parfaite en $\mathbf{h} = \mathbf{0}$, *i.e.* si $C_{12}(\mathbf{0}) = \{C_{11}(\mathbf{0})\, C_{22}(\mathbf{0})\}^{1/2}$ et si les variances sont égales, soit $C_{11}(\mathbf{0}) = C_{22}(\mathbf{0})$. En général, on a donc $\gamma_{12}^{p}(\mathbf{h}) > 0$.

Il est possible d'écrire le variogramme croisé en fonction du pseudo-variogramme croisé, la réciproque étant généralement fausse, à moins d'imposer des conditions de stationnarité plus restrictives (Papritz *et al.* 1993) :

$$\gamma_{12}(\mathbf{h}) = \frac{1}{2}\left[\gamma_{12}^{p}(\mathbf{h}) + \gamma_{21}^{p}(\mathbf{h})\right] - \gamma_{12}^{p}(\mathbf{0}) \tag{10.26}$$

Lorsque $Z_1(\cdot)$ et $Z_2(\cdot)$ sont deux FAI-0 strictes, $\gamma_{12}^{p}(\mathbf{h})$ n'existe pas, à moins que les autovariogrammes $\gamma_{11}(\mathbf{h})$ et $\gamma_{22}(\mathbf{h})$ ne croissent à la même vitesse pour les classes de grandes distances (Papritz *et al.* 1993).

10.2.4.1 Estimateur

Le *pseudo-variogramme croisé expérimental* peut se calculer classiquement selon (Goovaerts 1997, p. 48) :

$$\widehat{\gamma}_{12}^{p}(\mathbf{h}) = \frac{1}{2N(h)} \sum_{(i,j)|\mathsf{h}_{ij}=\mathsf{h}} (z_{1j} - z_{2i})^2 \tag{10.27}$$

[4] Wackernagel (1993, p. 73) considère au contraire que le pseudo-variogramme croisé ne peut pas traduire une éventuelle association spatiale négative entre deux variables.

avec les mêmes notations que dans la Section 10.2.2.1. L'inconvénient de l'estimateur (10.27) est que le pseudo-variogramme croisé apparaît très influencé par la variable qui présente les plus grandes valeurs. Il est par conséquent conseillé de transformer les données, par exemple en les standardisant (Goovaerts 1997). Toutefois, l'estimateur (10.27) peut être utilisé dans le cas d'une même variable mesurée à deux dates différentes, du moins si les valeurs varient à l'intérieur d'un intervalle similaire (Goovaerts 1997). Une autre approche consiste à inclure directement le centrage des valeurs dans la définition de l'estimateur lui-même, ce qui donne (Papritz *et al.* 1993) :

$$\widehat{\gamma}_{12}^{p}(\mathbf{h}) = \frac{1}{2N(h)} \sum_{(i,j)|\mathsf{h}_{ij}=\mathsf{h}} \left[(z_{1j} - \bar{z}_1) - (z_{2i} - \bar{z}_2)\right]^2 \tag{10.28}$$

Néanmoins, l'estimateur (10.28) est biaisé, l'amplitude du biais pour les faibles distances étant de l'ordre de $\mathrm{Var}\left[\bar{z}_1 - \bar{z}_2\right]$ (Papritz *et al.* 1993).

10.2.5 Autres fonctions croisées

En dehors des fonctions croisées issues de la géostatistique, il est important de souligner que toutes les fonctions d'autocorrélation décrites dans le Chapitre 3 peuvent être étendues à l'étude de l'association spatiale, simplement en considérant deux VR $z_1(\cdot)$ et $z_2(\cdot)$ au lieu d'une seule VR $z(\cdot)$. Par exemple, l'autocorrélogramme du I de Moran peut être généralisé en un corrélogramme croisé :

$$I_{12}(\mathbf{h}) = \frac{\widehat{C}_{12}(\mathbf{h})}{\sigma_1 \sigma_2} \tag{10.29}$$

avec

$$\sigma_1 = \left[\frac{1}{n_1} \sum_{i=1}^{n_1} (z_{1i} - \bar{z}_1)^2\right]^{1/2} \tag{10.30}$$

$$\sigma_2 = \left[\frac{1}{n_2} \sum_{i=1}^{n_2} (z_{2i} - \bar{z}_2)^2\right]^{1/2} \tag{10.31}$$

et $\widehat{C}_{12}(\mathbf{h})$ la fonction de covariance croisée expérimentale (10.13).

En outre, l'hypothèse nulle d'une absence de structure spatiale de la covariation des valeurs de $z_1(\cdot)$ et $z_2(\cdot)$ peut être testée pour tout $\mathbf{h}$ au moyen d'un test de randomisation (Chapitre 3), les permutations aléatoires concernant les valeurs d'une des deux VR. A notre connaissance, cette possibilité est encore très peu exploitée[5].

10.2.6 Application

Le variogramme croisé permet de mettre en évidence la covariation positive ($\widehat{\gamma}_{12}(\mathbf{h}) > 0$) et négative ($\widehat{\gamma}_{12}(\mathbf{h}) < 0$), au sein d'un même domaine d'étude D. Il peut être utilisé pour

[5] Le modèle de permutation est cependant considéré par Reich *et al.* (1994) afin de tester un indice de corrélation croisée analogue au I de Moran croisé, mais construit à partir de la covariance non centrée.

analyser la covariation spatiale d'une même espèce à deux dates différentes ou bien entre deux espèces (*cf.* Rossi *et al.* 1992), entre deux écophases d'une même espèce, entre deux nutriments du sol, etc. (Annexe G).

A notre connaissance, le pseudo-variogramme croisé n'a jamais été utilisé en écologie, du moins à des fins strictement descriptives. Son intérêt pratique dans cette discipline reste donc à évaluer, mais son interprétation est sans aucun doute plus délicate que celle du variogramme croisé ou de la covariance croisée.

Les fonctions croisées sont utiles pour décrire de façon synthétique la covariation spatiale entre deux VR, mais les représentations cartographiques des données peuvent s'avérer nécessaires à leur interprétation (*e.g.*, Rossi *et al.* 1992). Par ailleurs, lorsque l'échantillonnage de plusieurs VR est effectué selon une même grille, la comparaison directe des représentations cartographiques peut suffire à mettre en évidence les covariations spatiales. Enfin, l'analyse de la covariation spatiale à l'aide des fonctions structurales croisées se prolonge parfois par la cartographie des VR à des fins de comparaison visuelle entre cartes (*e.g.*, Cardina *et al.* 1996). Une autre approche consiste donc à comparer directement des cartes entre elles, qu'elles soient binaires ou quantitatives.

10.3 Association spatiale entre cartes binaires

En écologie, il est fréquent que les données spatiales soient du type présence/absence, parce que la quantification du phénomène étudié nécessiterait un travail de terrain démesuré par rapport aux moyens disponibles. C'est notamment le cas en biogéographie où il s'avère souvent impossible d'évaluer de façon suffisamment précise l'abondance des individus d'un groupe d'espèces, sauf lorsque la richesse spécifique du groupe n'est pas trop élevée, *i.e.* inférieure à 100 (Debinski & Humphrey 1997), et lorsque le groupe est largement étudié et ne pose pas de problèmes techniques insurmontables. L'étude de la répartition spatiale des espèces d'insectes fournit un exemple typique des situations où l'information manipulée est presque toujours de type présence/absence. Dans ce qui suit, nous traitons donc plus particulièrement des cartes de présence/absence d'espèces d'insectes.

La structure de répartition spatiale d'une espèce est souvent intéressante en elle-même, mais les facteurs expliquant cette structure s'exercent généralement sur d'autres espèces. L'étude simultanée de la répartition de plusieurs espèces peut permettre d'identifier les facteurs écologiques essentiels déterminant la façon dont ces espèces se répartissent dans l'espace géographique. En effet, si deux espèces sont affectées par les mêmes facteurs environnementaux et/ou si elles exercent une influence l'une sur l'autre (prédation, compétition, commensalisme, etc.), leurs structures de répartition spatiale seront associées. L'association spatiale ou l'absence d'association spatiale parmi un groupe d'espèces est par conséquent d'un intérêt écologique évident (Pielou 1969).

Considérons par exemple les cartes de présence/absence des lépidoptères Zygaenidae en région Bourgogne[6]. La région Bourgogne définit un domaine D discrétisé en 349 carrés UTM[7] de 10 km de côté, inclus dans une grille comportant 25 lignes et 22 colonnes. Les relevés de présence/absence dans chaque carré UTM pour les 22 espèces de Zygaenidae

[6] Données communiquées par le Dr. Claude Dutreix.

[7] Les grilles UTM (*Universal Transverse Mercator*) de 10×10 km et 50×50 km sont très utilisées en Europe pour les cartographies d'espèces (Cartan 1978).

présentes en région Bourgogne fournissent 22 cartes binaires en mode image (ou mode raster).

La région Bourgogne ne constitue qu'un sous-ensemble de l'aire de répartition des Zygaenidae, tandis que la famille des Zygaenidae ne constitue qu'un sous-ensemble des espèces animales présentes en région Bourgogne. Selon le type d'analyse effectué, l'étude est conditionnelle aux espèces ou au domaine. Ainsi, un des objectifs de la biogéographie écologique est de définir des groupes d'espèces dont la répartition spatiale est similaire, l'interprétation écologique de ces groupes faisant intervenir par la suite des cartes de facteurs biotiques et abiotiques (couverture végétale, facteurs climatologiques, etc.) (Birks 1987). Dans ce type d'analyse (R-analyse), les conclusions sont obtenues à partir d'un sous-ensemble de l'aire de répartition des espèces considérées. L'analyse duale (Q-analyse) vise à définir des régions écologiques au sein du domaine étudié, à partir d'un ensemble d'espèces (taxicoenose, guilde, etc.). Dans le cadre de l'association spatiale, nous considérons uniquement la R-analyse, la Q-analyse faisant référence au partitionnement spatial, sujet qui n'est pas traité dans ce mémoire.

Dans un contexte de classification automatique, la R-analyse nécessite de définir une similarité mesurant la ressemblance entre deux images binaires F et G. Soit la table de contingence 2×2 établie en croisant les valeurs des pixels des images F et G :

$\times$	1	0
1	a	b
0	c	d

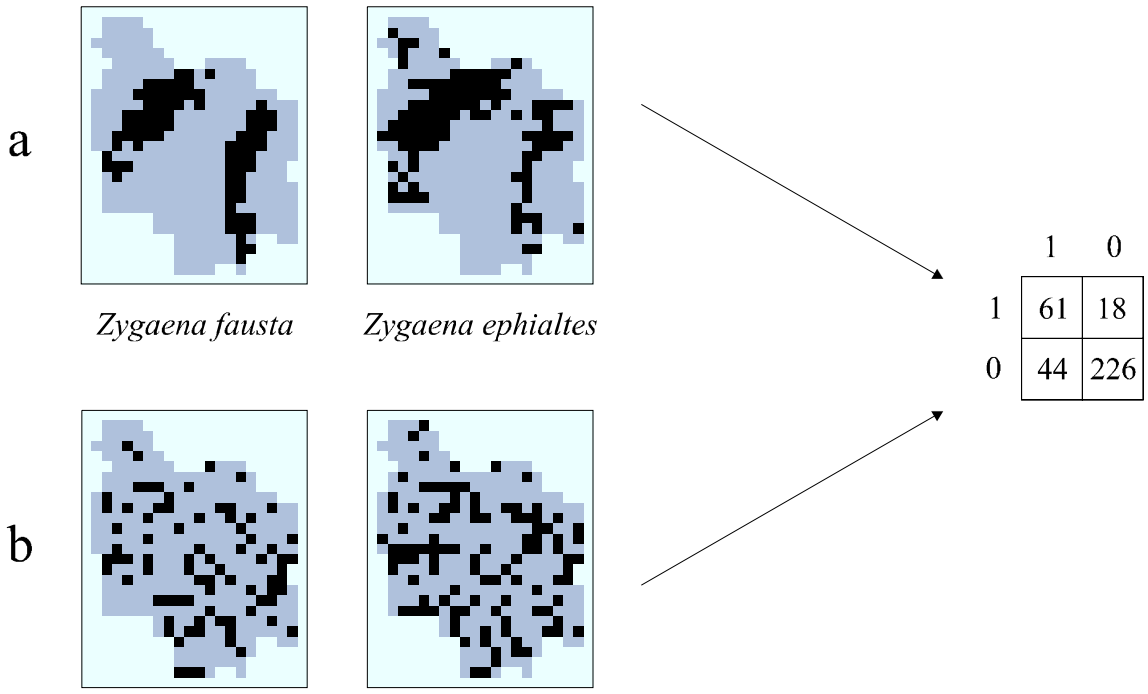
avec $a+b+c+d=n$ et n le nombre de pixels considérés. L'association binaire peut être quantifiée par une statistique exploitant la table de contingence 2×2, par exemple le χ^2_{obs} de Pearson, le coefficient de corrélation de point ϕ (Pielou 1969, pp. 160-166, Upton & Fingleton 1985, pp. 224-232, Li *et al.* 1993) ou bien d'autres mesures telles que le t de Tschuprow ou le v de Cramer (Legendre & Legendre 1984a, p. 174, Hermann 1986, p. 122, 129).

Cependant, le plus souvent les écologistes utilisent une similarité de forme générale $f(a,b,c,d)$. Il en existe de nombreuses définitions, revues notamment dans Cheetham & Hazel (1969), Blanc *et al.* (1976), Janson & Vegelius (1981), Hubálek (1982), Benzécri (1984, pp. 72-74), Legendre & Legendre (1984b, pp. 5-10), Roux (1985, pp. 127-132), Gower & Legendre (1986). Dans le cadre de la R-analyse des répartitions d'espèces, les écologistes font souvent référence à l'indice de Jaccard $S_J \in [0,1]$ (Cheetham & Hazel 1969) :

$$S_J = \frac{a}{a+b+c} \tag{10.32}$$

Cette similarité est maximale lorsque $b=c=0$, autrement dit, lorsque la concordance des présences des deux espèces est parfaite, et minimale lorsqu'il n'existe aucune concordance ($a=0$). Il est important de noter que le terme d n'est pas pris en compte dans la définition (10.32). L'indice de Jaccard S_J peut être transformé en une distance D_J en prenant son complément à 1, soit $D_J = 1 - S_J$.

Quelle que soit la mesure d'association binaire, seule l'information contenue dans la table de contingence 2×2 est utilisée, et la structure spatiale des cartes F et G n'est pas prise en compte. Considérons par exemple la répartition de *Zygaena fausta* et de *Zygaena ephialtes* en région Bourgogne. La distance de Jaccard entre les deux cartes de présence/absence est $D_J = 0.50407$ ($a = 61$, $b = 18$, $c = 44$, $d = 226$, $n = 349$). Cette distance reste inchangée si la structure spatiale des deux cartes est détruite par 10^6 permutations aléatoires réalisées en parallèle afin de conserver l'association pixel-à-pixel (Fig. 10.1).

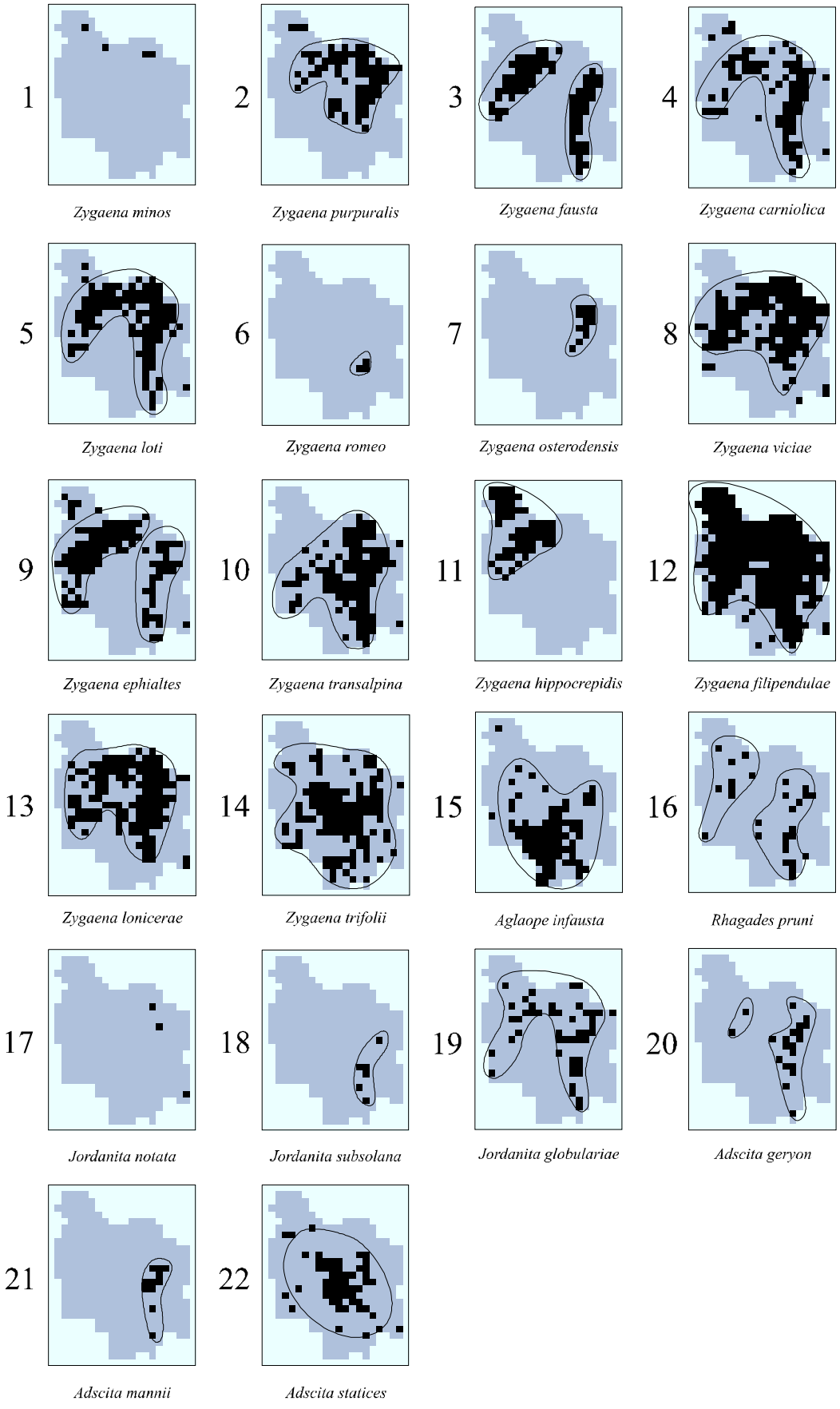


Figure˜10.1: Résumé statistique de l'association entre deux cartes de présence/absence au moyen d'une table de contingence 2×2. (a) Cartes de répartition de *Zygaena fausta* et *Zygaena ephialtes*. (b) Cartes randomisées en parallèle par 10^6 permutations aléatoires afin de détruire leur structure spatiale tout en conservant l'association pixel-à-pixel.

Cet exemple illustre bien le caractère a-spatial de l'approche classique, et révèle une prise en compte incomplète de la nature de l'information traitée, *a priori* préjudiciable à l'analyse. Il faut toutefois distinguer deux types d'images binaires :

1. les images de haute résolution, issues de la télédétection, obtenues de façon instantanée,
2. les images de faible résolution, issues d'observations de terrain, menées sur le long terme, et concernant des espèces dont la rareté est très variable.

Pour le premier type d'image, la distance de Jaccard est en général satisfaisante, dans la mesure où l'information véhiculée par la table de contingence suffit à traduire de façon précise la notion d'association spatiale. En revanche, pour évaluer la ressemblance entre des images du second type, l'écologiste a tout d'abord tendance à synthétiser l'information disponible en délimitant des aires de répartition (Fig. 10.2).



Figure˜10.2: Cartes de répartition des 22 espèces de lépidoptères Zygaenidae présentes en région Bourgogne. Un pixel représente un carré UTM de 10 km de côté. Les aires de répartition ont été délimitées manuellement.

La délimitation permet en effet une superposition assez aisée de la représentation des aires de répartition de plusieurs espèces, bien que les limites tracées soient subjectives (Cartan 1978). Cette pratique témoigne de ce que la notion d'association spatiale de l'écologiste fait intervenir la forme des aires de répartition. Or, la ressemblance globale entre certaines formes ne se traduit pas nécessairement par une bonne concordance au niveau élémentaire de chaque pixel. L'écologiste a plutôt tendance à faire abstraction des concordances locales pour ne retenir finalement que la coïncidence globale des formes, en tenant compte du fait que :

- l'intensité de la prospection géographique est souvent inhomogène (sur-prospection ou sous-prospection de certains carrés UTM),
- la facilité d'observation des espèces est très variable (espèces communes ou rares),
- les images ne correspondent pas à des cartographies instantanées et précises de la réalité.

Il convient de respecter la notion de ressemblance de l'écologiste plutôt que de lui imposer une définition opératoire restrictive et mal adaptée. Deux approches sont possibles :

- reproduire le mode d'analyse de l'écologiste, au moyen d'un système à base de connaissances,
- construire un algorithme raisonnable dont la conclusion est en accord avec l'expertise de l'écologiste.

Nous nous contentons ici de proposer un algorithme, en considérant que le problème de l'association spatiale entre cartes binaires doit être formulé comme un problème de ressemblance entre formes. Dans ce contexte, deux approches sont envisageables (Miclet 1984) :

- l'approche statistique,
- l'approche structurelle.

L'approche statistique est illustrée ici par les mesures de ressemblance exploitant la table de contingence 2×2, notamment la disance de Jaccard. A tout prendre, la distance de Jaccard est satisfaisante lorsque les formes sont finement discrétisées et représentent fidèlement la réalité. Dans le contexte des images incertaines de faible résolution, nous considérons que l'approche statistique est insuffisante, et qu'il faut également utiliser l'approche structurelle.

Le recours à l'approche structurelle nécessite de coder les images binaires dans un autre espace de représentation que celui utilisé par l'approche statistique. Une distance peut être calculée dans ce nouvel espace de représentation, non pas par une formule mais plutôt par un algorithme, tenant compte par exemple d'une hiérarchie ou d'un ordre dans les coordonnées des pixels (Miclet 1984). Les approches statistique et structurelle sont naturellement opposées quant à leur formalisme, mais elles s'avèrent complémentaires, et il semble judicieux d'utiliser les deux approches conjointement afin de bénéficier de leurs avantages respectifs (Miclet 1984). Pour pouvoir appliquer l'approche structurelle, il nous faut choisir :

- un mode de représentation qui tienne compte de la structure d'une image binaire, définissant ainsi un espace de représentation structurel $\mathcal{E}$,
- une distance définie dans $\mathcal{E}$.

10.3.1 Structure des images binaires

Nous considérons des images binaires en noir et blanc, les pixels noirs correspondant à la présence et les pixels blancs à l'absence. Toute image de ce type peut être étendue à une image carrée de taille $2^r \times 2^r$ en ajoutant le nombre nécessaire de pixels blancs en lignes et en colonnes. La structure d'une image binaire $2^r \times 2^r$ peut être représentée sous forme hiérarchique grâce à un *quadtree* (Section 2.3.4.1). Nous considérons donc par la suite un espace de représentation $\mathcal{E}$ constitué par tous les *quadtrees* codant des images binaires $2^r \times 2^r$, pour r fixé.

10.3.2 Distance structurelle entre images binaires

La technique la plus simple pour mesurer l'association entre images binaires dans l'espace de représentation $\mathcal{E}$ consiste à définir une distance entre arbres. Dans la mesure où cette distance correspond à une définition cohérente, il est alors possible de comparer les formes représentées par ces arbres (Miclet 1984). Nous proposons de mesurer la distance entre deux *quadtrees* A et B au moyen d'une distance d'édition, *i.e.* en comptant le plus petit nombre de transformations élémentaires nécessaires pour passer de A à B (Miclet 1984). Les *quadtrees* étant des arbres étiquetés récursifs, l'algorithme de Selkow (1977) s'avère parfaitement adapté pour calculer la distance d'édition entre A et B.

Un arbre étiqueté récursif est un ensemble fini non vide de sommets T muni d'une fonction d'étiquetage $\lambda(\cdot)$ telle que (Selkow 1977, Miclet 1984) :

- il existe un sommet distinct : la racine de T,
- les sommets restants sont partitionnés en $m \geq 0$ ensembles disjoints $T_1, \ldots, T_m$ et chacun de ces ensembles est un arbre (sous-arbre de T),
- à chaque sommet $\nu \in T$ est associée une étiquette $\lambda(\nu)$, et $\lambda(T)$ est l'étiquette de la racine de T,
- pour $0 \leq i \leq m$, $T\langle i \rangle$ est l'arbre obtenu à partir de T en supprimant les sous-arbres $T_{i+1}, \ldots, T_m$ (en particulier $T = T\langle m \rangle$),
- si A est un arbre de sous-arbres $A_1, \ldots, A_m$ et B un arbre de sous-arbres $B_1, \ldots, B_n$ alors $A = B$ si $\lambda(A) = \lambda(B)$, $m = n$ et $A_i = B_i$ pour $1 \leq i \leq m$.

Etant donné un arbre T dont les étiquettes appartiennent à l'ensemble $\{e_1, \ldots, e_q\}$, avec $\lambda(T) = e_j$ et les sous-arbres $T_1, \ldots, T_m$, on peut définir trois opérations élémentaires (Selkow 1977, Miclet 1984) :

1. opération de changement d'étiquette $L(e_j, e_k)$ qui fait passer de T à T^*, avec $\lambda(T^*) = e_k$ et $T_1, \ldots, T_m$ les sous-arbres de T^*,
2. pour $0 \leq i \leq m$ et un arbre A, l'opération d'insertion $I(A)$ appliquée à T en i donne T^*, avec $\lambda(T^*) = e_j$ et $T_1, \ldots T_i, A, T_{i+1}, \ldots, T_m$ les sous-arbres de T^*,
3. pour $1 \leq i \leq m$, l'opération de délétion $D(T_i)$ appliquée à T en i donne T^*, avec $\lambda(T^*) = e_j$ et $T_1, \ldots, T_{i-1}, T_{i+1}, \ldots, T_m$ les sous-arbres de T^*.

Chacune des trois opérations précédentes constitue une opération d'édition.

Un coût non négatif est associé à chaque opération d'édition :

- pour chaque paire d'étiquettes (e_i, e_j) l'opération $L(e_i, e_j)$ s'effectue selon un coût $C_L(e_i, e_j)$,
- pour chaque étiquette e_i, les coûts $C_I(e_i)$ et $C_D(e_i)$ sont associés respectivement à l'insertion $I(T)$ et à la délétion $D(T)$ où T est un arbre à un seul sommet et $\lambda(T) = e_i$.

Les coûts des opérations d'insertion et de délétion pour un arbre arbitraire T peuvent être définis comme :

$$C_I(T) = \sum_{\nu \in T} C_I(\lambda(\nu)) \tag{10.33}$$

$$C_D(T) = \sum_{\nu \in T} C_D(\lambda(\nu)) \tag{10.34}$$

Pour tout triplet d'étiquettes (e_i, e_j, e_k) on suppose que C_L respecte $C_L(e_i, e_i) = 0$ et $C_L(e_i, e_j) \leq C_L(e_i, e_k) + C_L(e_k, e_j)$. Etant donnés deux arbres A et B et l'ensemble des séquences d'opérations d'édition qui permettent de passer de A à B, on note $d(A, B)$ le minimum des sommes des coûts de chaque séquence. Soit A un arbre de sous-arbres $A_1, \ldots, A_m$ et B un arbre de sous-arbres $B_1, \ldots, B_n$, alors :

$$d(A, B) \leq C_L(\lambda(A), \lambda(B)) + \sum_{i=1}^{m} C_D(A_i) + \sum_{i=1}^{n} C_I(B_i) \tag{10.35}$$

Selkow (1977) démontre la proposition suivante :

Proposition 2 *Soient A un arbre de sous-arbres $A_1, \ldots, A_m$ ($m \geq 0$) et B un arbre de sous-arbres $B_1, \ldots, B_n$ ($n \geq 0$), on a*

$$d(A\langle 0\rangle, B\langle j\rangle) = C_L(\lambda(A), \lambda(B)) + \sum_{k=1}^{j} C_I(B_k) \tag{10.36}$$

$$d(A\langle i\rangle, B\langle 0\rangle) = C_L(\lambda(A), \lambda(B)) + \sum_{k=1}^{i} C_D(A_k) \tag{10.37}$$

$$d(A\langle i\rangle, B\langle j\rangle) = \min \begin{cases} d(A\langle i-1\rangle, B\langle j-1\rangle) + d(A_i, B_j), \\ d(A\langle i\rangle, B\langle j-1\rangle) + C_I(B_j), \\ d(A\langle i-1\rangle, B\langle j\rangle) + C_D(A_i) \end{cases} \tag{10.38}$$

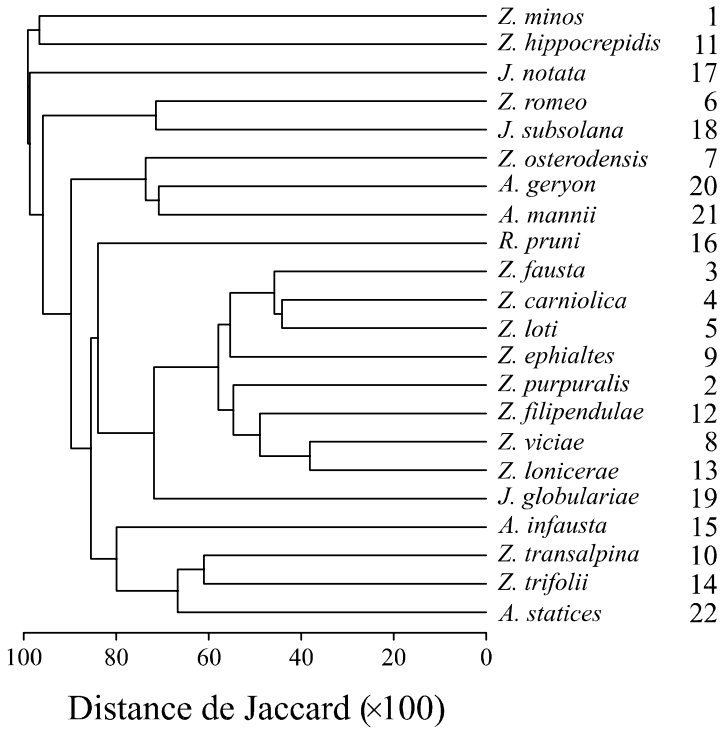
Un algorithme récursif dérive de cette proposition de calcul de $d(A, B)$ par programmation dynamique (Selkow 1977).

10.3.3 Etude de cas

Chaque carte de présence/absence des 22 espèces de Zygaenidae constitue une image binaire de 25×22 carrés UTM de 10 km de côté. La présence est codée 1 (pixels noirs) et l'absence 0 (pixels blancs).

10.3.3.1 Distance de Jaccard

Parmi les 550 pixels d'une image 25×22, seuls $n = 349$ pixels correspondent effectivement à des relevés de présence/absence dans les carrés UTM couvrant la région Bourgogne. Les carrés UTM situés en dehors de la région Bourgogne sont codés 0, ce qui est neutre vis-à-vis du calcul de la distance de Jaccard puisque D_J ne tient pas compte des absences concordantes. Le calcul des distances de Jaccard pour toutes les paires d'images conduit à une matrice de distances $\mathbf{A}$. Un dendrogramme est construit à partir de $\mathbf{A}$ par CAH selon le critère d'agrégation du lien moyen (Fig. 10.3).



Figure˜10.3: Dendrogramme obtenu par CAH (critère d'agrégation du lien moyen) à partir de la matrice des distances de Jaccard.

Le dendrogramme met en évidence le groupe $g_1 = (3, 4, 5, 9, 2, 12, 8, 13)$ auquel viennent s'agréger tardivement les cartes 19 puis 16. Le groupe $g_2 = (10, 14, 22)$ rassemble des aires de répartition différentes mais qui se recouvrent partiellement. Le groupe $g_3 = (6, 18)$ correspond à des aires très limitées. Trois aires partiellement chevauchantes sont rassemblées dans le groupe $g_4 = (7, 20, 21)$. Les cartes restantes $\{1, 11, 15, 17\}$ sont agrégées tardivement dans le dendrogramme et ne participent à aucun groupe. Un examen visuel des cartes suggère que :

- la carte 7, et surtout la carte 21, soient incorporées dans le groupe g_3,
- l'aire 12, qui couvre pratiquement toute la région Bourgogne, soit exclue du groupe g_1 qui rassemble des aires en forme de *boomerang* continu ou fragmenté en deux parties,
- la carte 10 soit séparée de la carte 14 dont l'aire de répartition n'a pas la même forme, et incorporée dans le groupe g_1,
- les cartes 16 et 19, et éventuellement 20, soient incorporées dans le groupe g_1.

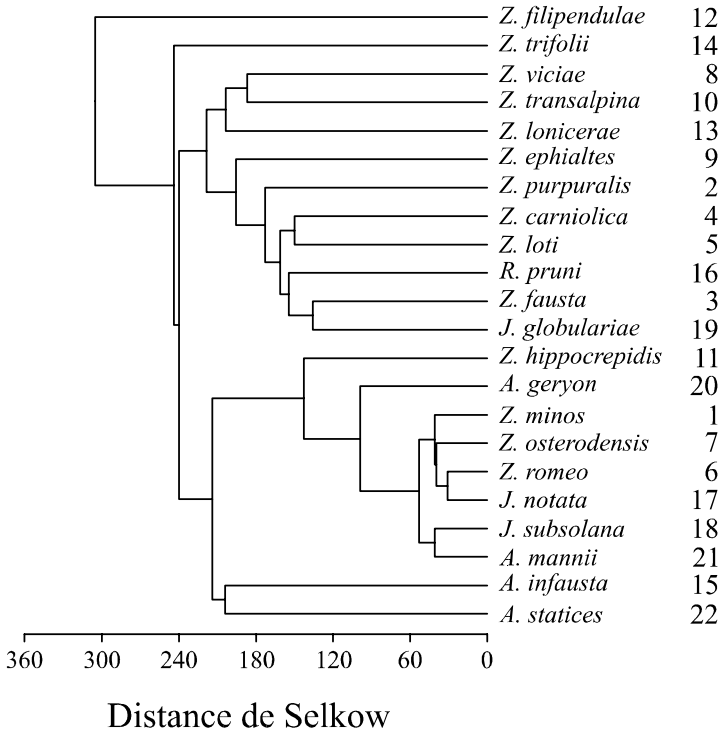
10.3.3.2 Distance de Selkow

Dans le cas d'un *quadtree* d'image binaire, les étiquettes de l'arbre codent le blanc, le noir et le gris (pour les quadrants comportant des pixels blancs et noirs). Pour l'opération de changement d'étiquette, nous utilisons une fonction de coût qui respecte les trois axiomes d'une métrique :

1. axiome de séparation : $C_L(e_i, e_j) = 0$ si et seulement si $e_i = e_j$,
2. axiome de symétrie : $C_L(e_i, e_j) = C_L(e_j, e_i)$,
3. inégalité triangulaire : $C_L(e_i, e_j) \leq C_L(e_i, e_k) + C_L(e_k, e_j)$.

Il n'y a pas de raison *a priori* de pénaliser une transformation plutôt qu'une autre de sorte que nous choisissons $C_L(e_i, e_j) = 1$ pour tout $e_i \neq e_j$, $C_I(e_i) = 1$ et $C_D(e_i) = 1$ pour tout e_i.

Les 22 images binaires correspondant aux cartes de présence/absence des Zygaenidae sont complétées par des pixels blancs afin de former des images carrées $2^5 \times 2^5$ (32×32). Un *quadtree* est construit pour chaque image en noir et blanc $2^5 \times 2^5$. Le calcul de la distance d'édition de Selkow (1977) pour toutes les paires de *quadtrees* produit une matrice de distances $\mathbf{B}$. Un dendrogramme est construit à partir de $\mathbf{B}$ par CAH selon le critère d'agrégation du lien moyen (Fig. 10.4).



Figure˜10.4: Dendrogramme obtenu par CAH (critère d'agrégation du lien moyen) à partir de la matrice des distances de Selkow.

Le dendrogramme met en évidence un grand groupe rassemblant les aires en forme de *boomerang* continu ou fragmenté en deux parties : $(8, 10, 13, 9, 2, 4, 5, 16, 3, 19)$. Un second groupe rassemble les cartes des espèces très localisées $(1, 7, 6, 17, 18, 21)$. Les cartes 20 puis

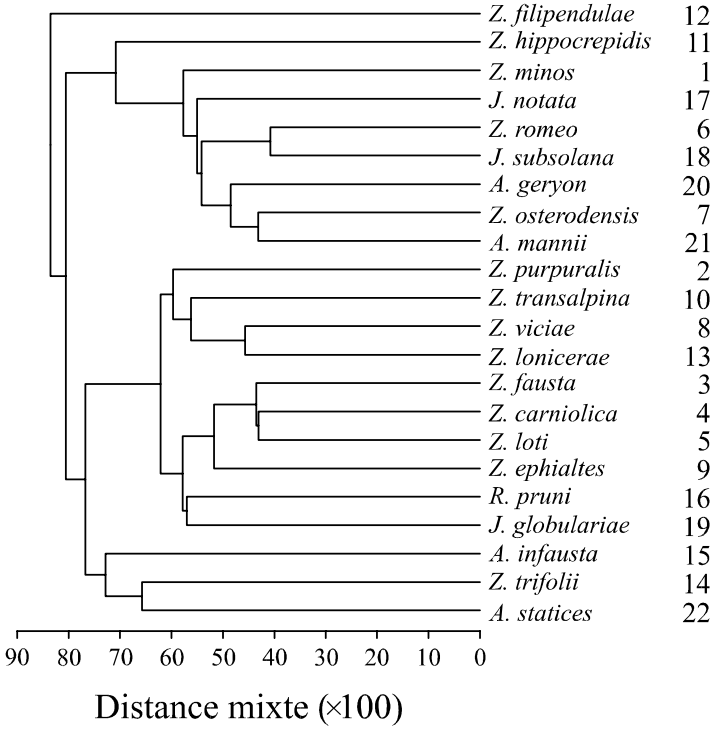
11 s'agrègent tardivement au groupe précédent. Les cartes 15 et 22 constituent un groupe agrégé tardivement dans le dendrogramme. Enfin les cartes 14 et 12 s'avèrent inclassables.

10.3.3.3 Distance mixte

Définissons une distance mixte combinant les approches statistique (distance de Jaccard) et structurelle (distance de Selkow). A partir des deux matrices $\mathbf{A} = ((a_{ij}))$ et $\mathbf{B} = ((b_{ij}))$, nous construisons une matrice de distances $\mathbf{C}$ dont les éléments c_{ij} sont calculés comme des combinaisons linéaires de distances standardisées :

$$c_{ij} = \alpha \frac{a_{ij}}{\max(\mathbf{A})} + \beta \frac{b_{ij}}{\max(\mathbf{B})} \tag{10.39}$$

avec $\max(\mathbf{A})$ la plus grande distance dans $\mathbf{A}$, $\max(\mathbf{B})$ la plus grande distance dans $\mathbf{B}$ et $\alpha + \beta = 1$. Nous n'avons pas de raison *a priori* de favoriser une distance plutôt que l'autre, aussi nous choisissons $\alpha = \beta = \frac{1}{2}$. Un dendrogramme est construit à partir de $\mathbf{C}$ par CAH selon le critère d'agrégation du lien moyen (Fig. 10.5).



Figure˜10.5: Dendrogramme obtenu par CAH (critère d'agrégation du lien moyen) à partir de la matrice des distances mixtes (détails dans le texte).

L'aire de répartition 12 couvrant presque toute la région Bourgogne se singularise en étant agrégée dans le dendrogramme en dernier. Le groupe $(6, 18, 20, 7, 21)$ rassemble les aires très localisées, auxquelles s'agrègent plus tardivement les cartes 17 et 1 qui correspondent également à des espèces très localisées mais situées ailleurs dans le territoire. L'aire 11, compacte et isolée au nord-ouest de la région Bourgogne, ne participe à aucun groupe. Un grand groupe $(2, 10, 8, 13, 3, 4, 5, 9, 16, 19)$ rassemble toutes les aires en forme de *boomerang* continu ou fragmenté en deux parties.

Ce groupe est lui-même composé :

- du groupe des aires compactes en forme de *boomerang* peu prononcé $(2, 10, 8, 13)$,
- du groupe des aires en forme de *boomerang* prononcé ou fragmenté en deux parties $(3, 4, 5, 9)$,
- du groupe des aires en forme de *boomerang* lacunaire $(16, 19)$.

Enfin, les aires centrales 14 et 22 sont associées tardivement dans le dendrogramme et l'aire 15 s'avère inclassable.

10.3.3.4 Interprétation

Il convient tout d'abord de rappeler que les résultats de l'étude de l'association entre répartitions d'espèces sont conditionnels aux définitions du domaine et des supports des observations, autrement dit, à l'échelle spatiale considérée (Pielou 1969, Jonsson & Moen 1998).

La distance de Jaccard (approche statistique) et la distance de Selkow (approche structurelle) constituent deux définitions opératoires différentes de la dissemblance des cartes de présence/absence. La distance de Jaccard, calculée d'après une table de contingence, tient compte de la concordance des présences au niveau des pixels. Cette distance a pour avantage de traiter la superposition des cartes de façon exacte. Son inconvénient majeur est d'ignorer les structures spatiales puisque chaque pixel est considéré individuellement. En revanche, la distance de Selkow entre *quadtrees* tient compte des formes présentes dans les images. En effet, un *quadtree* décompose hiérarchiquement une image en quadrants homogènes, de sorte que le calcul d'une distance d'édition entre *quadtrees* prend en compte plusieurs échelles de structuration spatiale. L'inconvénient de cette approche est de faire passer la concordance spatiale au second plan.

En tenant compte à la fois de la concordance spatiale exacte, grâce à la distance de Jaccard, et des formes, grâce à la distance entre *quadtrees*, les cartes de présence/absence étudiées sont regroupées de façon très satisfaisante (Fig. 10.5). Bien que l'étude du déterminisme des aires de répartition des Zygaenidae en région Bourgogne soit en dehors de notre propos, il est remarquable de noter que la forme du *boomerang* correspond à la répartition du calcaire jurassique : les exigences écologiques des plantes hôtes des Zygaenidae constituent évidemment un facteur essentiel du déterminisme des répartitions spatiales observées.

10.4 Association spatiale entre cartes quantitatives

Les cartes quantitatives se présentent essentiellement sous la forme d'images en niveaux de gris ou de cartes isoplèthes. Par définition, une image est constituée d'une grille de pixels auxquels sont associées les valeurs d'une VR, réelles ou estimées. A partir de cette grille de valeurs, il est possible de calculer une carte isoplèthe (*e.g.*, Davis 1986, pp. 365-377). Inversement, la rasterisation d'une carte isoplèthe produit une image en niveaux de gris (Section 2.3.4). En conséquence, une image en niveaux de gris et une carte isoplèthe peuvent être vues comme deux modèles géomatiques alternatifs d'une même VR. Dans

ce qui suit, nous considérons essentiellement les images en niveaux de gris qui se prêtent plus facilement aux traitements analytiques que les cartes isoplèthes.

Nous examinons succinctement les comparaisons licites dans le cas de deux VR quantitatives, puis les méthodes utilisées pour comparer les cartes quantitatives.

10.4.1 Types de comparaisons

Considérons deux variables régionalisées quantitatives $f(\cdot)$ et $g(\cdot)$ connues par les valeurs $f = \{f_i \mid i = 1, \ldots, n_f\}$ et $g = \{g_i \mid i = 1, \ldots, n_g\}$. Ces deux VR sont modélisées par deux cartes $F(\cdot)$ et $G(\cdot)$ définies sur des supports organisés selon une même grille Ω afin d'assurer une situation d'isotopie. Dans l'absolu, il est possible d'envisager des comparaisons (*e.g.*, Söderström & Eriksson 1996) :

a. entre les données f et g, sauf dans le cas d'une situation d'hétérotopie totale,
b. entre les données f et la carte $G(\Omega)$ obtenue à partir des données g,
c. entre les deux cartes $F(\Omega)$ et $G(\Omega)$ obtenues, respectivement, à partir des données f et g.

Cependant, il convient de remarquer que les deux cartes $F(\Omega)$ et $G(\Omega)$ peuvent résulter :

1. d'un échantillonnage systématique des VR sur Ω, conduisant aux données f et g,
2. d'une estimation spatiale des VR sur Ω, à partir des données f et g réparties de façon irrégulière dans D,
3. d'un échantillonnage systématique de $f(\cdot)$ sur Ω conduisant aux données f, et d'une estimation spatiale de $g(\cdot)$ sur Ω, à partir des données g réparties de façon irrégulière dans D.

Les trois cas doivent être distingués car les valeurs estimées ne présentent pas la même variabilité que les valeurs réelles à cause du phénomène de lissage propre à la plupart des méthodes d'interpolation (Switzer 1983). Dans le cas 1, les comparaisons (a), (b) et (c) sont strictement équivalentes et licites. Dans le cas 2, la comparaison (a) est possible s'il existe des supports communs aux données f et g. En toute rigueur, la comparaison (b) n'est pas valide parce qu'elle confronte des valeurs réelles à des valeurs estimées. La comparaison (c) est licite, du moins en tant que comparaison entre modèles. Dans le cas 3, la comparaison (a) est possible s'il existe des supports communs aux données f et g, mais les comparaisons (b) et (c) ne sont pas valides parce qu'elles confrontent des valeurs réelles à des valeurs estimées. En conséquence :

1. la situation idéale consiste à comparer deux échantillons systématiques organisés selon la même grille Ω,
2. la comparaison entre deux cartes estimées $F(\Omega)$ et $G(\Omega)$ est licite, mais sous réserve que $F(\Omega)$ et $G(\Omega)$ aient été obtenues dans des conditions similaires, tant du point de vue de la quantité de données et de leur répartition spatiale, que de la procédure d'estimation elle-même,
3. dans la mesure du possible, les autres comparaisons doivent être évitées.

Dans le contexte de la comparaison 2, Davis (1986), puis Söderström & Eriksson (1996), considèrent qu'il n'est pas possible de juger de la signification statistique des corrélations parce qu'elles reposent entièrement sur des valeurs interpolées. Il convient tout d'abord de rappeler que l'obstacle majeur au test usuel de la corrélation est l'autocorrélation des valeurs (Section 9.1.1). En admettant que le test tienne compte de l'autocorrélation spatiale, la question qui se pose est de savoir si l'objectif consiste à tester la corrélation des VR sous-jacentes $f(\cdot)$ et $g(\cdot)$, auquel cas il est évident que l'utilisation des cartes estimées $F(\Omega)$ et $G(\Omega)$ n'est pas valide, ou simplement des cartes elles-mêmes, en tant que modèles, et dans ce cas rien ne s'oppose à tester la corrélation entre $F(\Omega)$ et $G(\Omega)$.

10.4.2 Méthodes de comparaison

Bien que d'un intérêt primordial dans l'étude de l'association, les méthodes statistiques de comparaison spatiale de deux cartes quantitatives $F(\Omega)$ et $G(\Omega)$ sont extrêmement peu développées (Minns *et al.* 1996). L'approche la plus élémentaire — mais peut-être encore la plus satisfaisante à l'heure actuelle — est la comparaison visuelle des cartes (Annexe G). Afin de rendre la comparaison entre cartes quantitatives plus objective que l'examen visuel, les approches classiquement envisagées sont de trois types :

- calcul d'une dissimilarité globale,
- calcul d'une distance entre modèles,
- calcul d'une image de la différence.

Il a été récemment proposé de comparer les cartes à l'aide de statistiques résumant le chevauchement de frontières, celles-ci étant définies comme des zones de changement rapide au sein de chaque carte (Jacquez 1995, Fortin *et al.* 1996). Cette approche est intéressante, mais elle pose le problème de la définition objective des frontières, sujet que nous ne traitons pas ici[8] (*cf.* Campbell 1978, Barbujani *et al.* 1989, Cornelius & Reynolds 1991, Oden *et al.* 1993, Bocquet-Appel & Bacro 1994, Fortin 1994, 1997, 1999, Fortin & Drapeau 1995).

10.4.2.1 Calcul d'une dissimilarité globale

La façon la plus simple de comparer deux images $F(\Omega)$ et $G(\Omega)$ est de calculer une dissimilarité globale (Davis 1986). En dehors du coefficient de corrélation de Pearson, de nombreuses statistiques peuvent être utilisées (revue dans Legendre & Legendre 1984b, pp. 10-35), notamment la métrique L_p ou métrique de Minkowski :

$$d\left[F(\Omega), G(\Omega)\right] = \left[\sum_{x\in\Omega} \left|F(x) - G(x)\right|^p\right]^{1/p} \tag{10.40}$$

avec x un pixel de la grille Ω. Seules les métriques L_1 (distance de Manhattan) ou L_2 (distance euclidienne) sont utilisées en pratique, les valeurs $p > 2$ donnant un poids trop important aux forts écarts absolus (Legendre & Legendre 1984b).

[8] Le problème peut être abordé en termes de segmentation d'image : voir Gonzales & Wintz (1987).

L'inconvénient majeur de ce type d'approche est que la statistique observée peut refléter correctement la correspondance entre les deux images ou, au contraire, être le résultat d'écarts importants dans une petite région de l'image seulement (Davis 1986). Ceci est dû à la nature a-spatiale des statistiques utilisées qui traitent chaque pixel de façon indépendante. Une discussion similaire à celle menée dans la Section 10.3 peut conduire à étendre notre approche de coopération statistique/structurelle utilisant à la fois une dissimilarité (*e.g.*, la métrique L_1) et la distance de Selkow entre les *quadtrees* associés aux images.

10.4.2.2 Calcul d'une distance entre modèles

Le principal inconvénient du calcul d'une dissimilarité globale entre les deux images est de traiter les couples de valeurs $\{F(x), G(x)\}\ \forall x \in \Omega$ comme s'ils étaient indépendants. Pour Merriam & Sneath (1966), il est évident que les couples $\{F(x), G(x)\}$ ne sont pas indépendants, et les différences entre $F(\Omega)$ et $G(\Omega)$ peuvent être vues comme des conséquences logiques de la variation spatiale. Dans cette optique, la comparaison de $F(\Omega)$ et $G(\Omega)$ peut s'effectuer en deux étapes :

- ajustement d'un modèle de variation spatiale à chaque image,
- calcul d'une distance entre les deux modèles.

Les modèles de variation spatiale classiquement considérés sont des surfaces de tendance polynomiales de degré k, ajustées aux moindres carrés ordinaires (Merriam & Sneath 1966, Davis 1986, pp. 459-461). Notons les $n = (k+1)(k+2)/2$ coefficients des polynômes ajustés à $F(\Omega)$ et $G(\Omega)$, respectivement $\beta_i^{(F)}$ et $\beta_i^{(G)}$, pour $i = 1, \ldots, n$. Merriam & Sneath (1966) proposent de calculer le coefficient de corrélation de Pearson entre les coefficients ou bien la distance :

$$d = \left[\frac{1}{n} \sum_{i=1}^{n} \left(\beta_i^{(F)} - \beta_i^{(G)} \right)^2 \right]^{1/2} \tag{10.41}$$

Merriam & Sneath (1966) recommandent d'éviter les surfaces de degré élevé et utilisent des polynomes de degré $k = 3$.

Le recours à une distance entre modèles constitue une approche intéressante. Néanmoins, le calcul d'une distance entre les coefficients de surfaces de tendance soulève plusieurs objections. D'abord, il est nécessaire d'ajuster des polynômes de mêmes degrés, ce qui représente une contrainte imposée par le calcul de (10.41) et pas par la variabilité spatiale présente dans chacune des images. Il n'y a aucune raison, *a priori*, de considérer par exemple qu'une surface polynomiale cubique ($k = 3$) constitue un modèle satisfaisant pour n'importe quelle image de phénomène écologique[9] : une modélisation fidèle de la variabilité spatiale nécessite vraisemblablement une surface polynomiale de degré plus élevé. En outre, les surfaces de tendance présentent l'inconvénient de produire des ondulations artefactuelles vers les bords du domaine afin d'ajuster les valeurs centrales (Ripley 1981, p. 30). Enfin, les coefficients de régression sont parfois sujets à d'extrêmes fluctuations d'origine strictement numérique, particulièrement lorsque k est élevé (Davis 1986). En

[9] Legendre *et al.* (1997) justifient l'utilisation de surfaces de tendance cubiques comme un moyen de rechercher des phénomènes dont l'échelle est identique (ou supérieure) à celle du domaine d'étude.

effet, comme tout problème de régression polynomiale aux moindres carrés, l'ajustement d'une surface de tendance polynomiale est un problème mal conditionné (Ripley 1981, p. 30). Le problème se révèle d'autant plus mal conditionné que le degré du polynôme augmente (Unwin 1975).

10.4.2.3 Calcul d'une image de la différence

Afin de comparer deux images $F(\Omega)$ et $G(\Omega)$ sans recourir à un résumé global, l'approche la plus simple consiste à calculer une image de la différence (Upton & Fingleton 1985, Davis 1986). Soit $\Delta(\Omega)$ l'image résultant du calcul de la différence entre $F(\Omega)$ et $G(\Omega)$ au sens d'un opérateur diff $[F(\Omega), G(\Omega)]$. Il est souhaitable que l'opérateur diff $(\cdot, \cdot)$ soit symétrique en ses arguments, par exemple :

$$\Delta(\Omega) = \{\Delta(x) \mid \Delta(x) = |F(x) - G(x)|,\ x \in \Omega\} \tag{10.42}$$

autrement dit, l'image $\Delta(\Omega)$ est obtenue en calculant la métrique L_1 pour les couples de valeurs $\{F(x), G(x)\}\ \forall x \in \Omega$, indépendamment les uns des autres. Cependant, à la différence des mesures de la Section 10.4.2.1, le résultat de ce calcul n'est pas global et la nature régionalisée de l'information est conservée. En procédant ainsi, le problème de l'association spatiale entre deux images se trouve ramené à celui de l'analyse de $\Delta(\Omega)$.

A nouveau, les méthodes d'analyse de $\Delta(\Omega)$ apparaissent extrêmement peu développées. Le plus simple est de localiser visuellement les régions où $F(\Omega)$ et $G(\Omega)$ varient de la même façon (Söderström & Eriksson 1996). Un point de vue davantage statistique — mais pas nécessairement plus utile — consiste à tester la présence d'autocorrélation spatiale significative dans $\Delta(\Omega)$ en considérant que la présence d'une structure spatiale dans la différence indique que les images ne correspondent pas (Cliff 1970, Legendre & McArdle 1997). Par exemple, dans leur comparaison de deux campagnes d'échantillonnage du macrobenthos du lac Erie, Minns *et al.* (1996) testent l'autocorrélation de la différence en utilisant un test de Mantel (Section 3.1.4).

En fait, en considérant que $F(\Omega)$ et $G(\Omega)$ sont spatialement structurées, l'absence d'autocorrélation spatiale significative au sein de $\Delta(\Omega)$ n'est possible que si $F(\Omega)$ et $G(\Omega)$ sont identiques, à quelques fluctuations près, ces fluctuations étant spatialement indépendantes. Dans l'immense majorité des cas, il faut s'attendre à ce que $\Delta(\Omega)$ soit autocorrélée, et la question est de savoir comment interpréter le niveau de structuration spatiale de $\Delta(\Omega)$ en termes d'association spatiale de $F(\Omega)$ et $G(\Omega)$, positive, non significative ou négative.

Il nous semble impossible de donner des règles précises quant à l'interprétation des images des différences, parce qu'il existe tout un continuum de situations entre l'identité stricte et la complémentarité stricte de $F(\Omega)$ et $G(\Omega)$, autrement dit, entre l'association positive parfaite et l'association négative parfaite. L'absence de résultat général provient également de l'existence d'un continuum de degrés de régularité spatiale pour $F(\Omega)$ et $G(\Omega)$. En termes de variogramme, ce continuum s'étend de l'effet de pépite très prononcé au comportement parabolique à l'origine. Toutefois, quelques généralités peuvent être énoncées, de façon assez intuitive.

L'analyse de $\Delta(\Omega)$ en termes de signe et de degré d'association spatiale de $F(\Omega)$ et $G(\Omega)$ doit porter à la fois sur l'amplitude des différences et sur leur degré de structuration spatiale. Considérons par exemple une famille d'images $F(\Omega)$ et $G(\Omega)$ comparables entre elles, dont les valeurs varient par exemple dans l'intervalle $[0, 1]$. En utilisant la même dynamique d'image, les niveaux de gris de $\Delta(\Omega)$ traduisent précisément l'amplitude des écarts absolus entre $F(\Omega)$ et $G(\Omega)$. Que la régularité spatiale soit faible (Fig. 10.6.1a & 10.6.1b) ou élevée (Fig. 10.6.3a & 10.6.3b), l'association spatiale positive se traduit par une image $\Delta(\Omega)$ globalement claire (Fig. 10.6.1c & 10.6.3c). Inversement, que la régularité spatiale soit faible (Fig. 10.6.2a & 10.6.2b) ou élevée (Fig. 10.6.4a & 10.6.4b), l'association spatiale négative se traduit par une image $\Delta(\Omega)$ globalement foncée (Fig. 10.6.2c & 10.6.4c). En considérant des images dont le variogramme présente un seuil c, le type d'association spatiale entre $F(\Omega)$ et $G(\Omega)$ peut être caractérisé — au moins grossièrement — par le niveau de gris global de $\Delta(\Omega)$ et son degré de structuration spatiale, révélé notamment par l'aspect du variogramme et la valeur de c. Cependant, nous ne savons pas précisément à partir de quel seuil l'association spatiale doit être considérée comme non significative (Tab. 10.1).

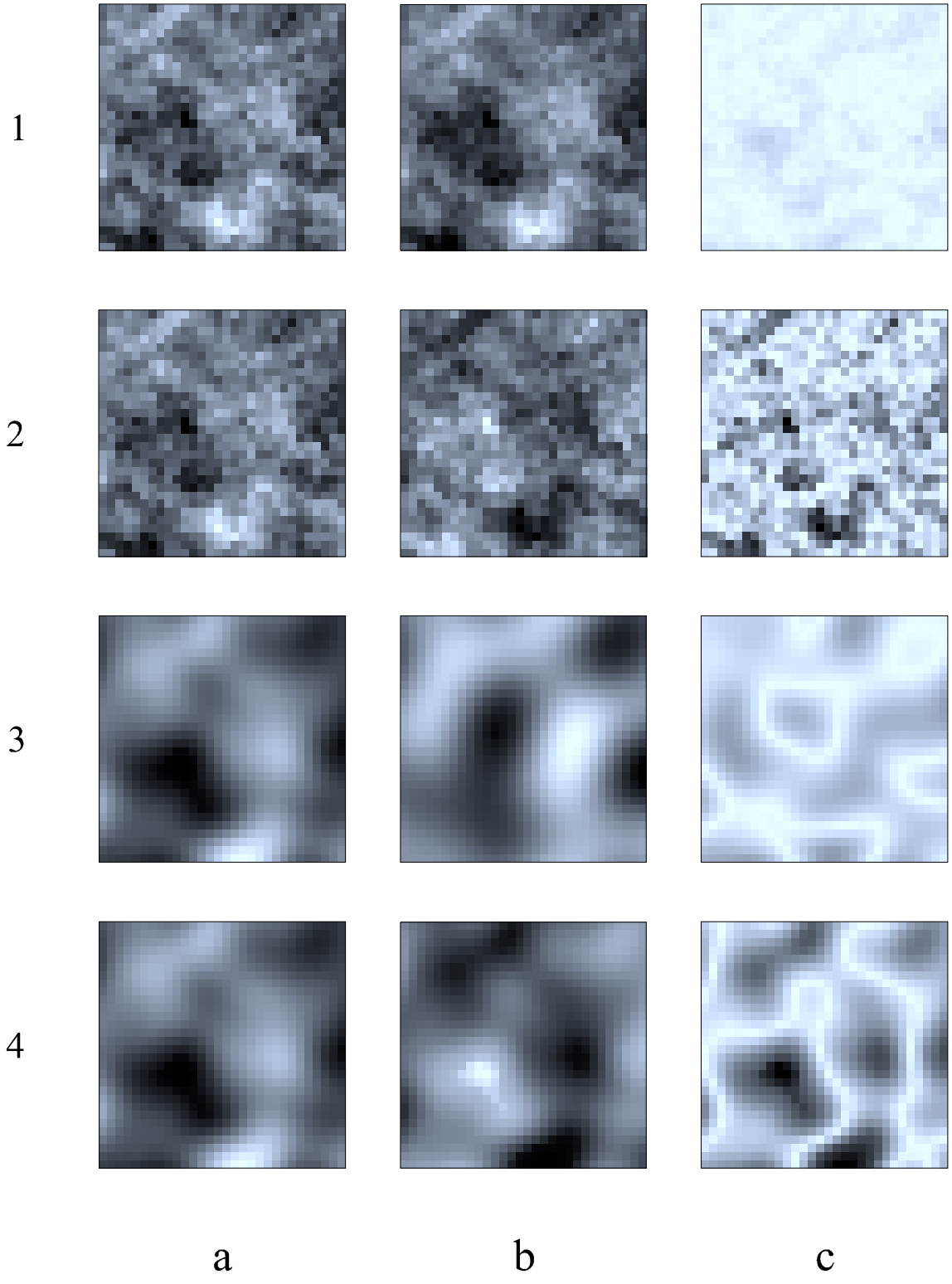
Relation entre $F(\Omega)$ et $G(\Omega)$	Niveau de gris de $\Delta(\Omega)$	Variogramme de $\Delta(\Omega)$
Identité	Blanc, uniforme	Plat, $c = 0$
Association positive	Gris clair, hétérogène	structuré, c faible
Association non significative	?	?
Association négative	Gris foncé, hétérogène	structuré, c élevé
Complémentarité (0/1)	Noir, uniforme	Plat, $c = 0$

Tableau 10.1: Essai de caractérisation du type de relation entre deux images en niveaux de gris $F(\Omega)$ et $G(\Omega)$ à partir du niveau de gris global et de la structure du variogramme de l'image de la différence $\Delta(\Omega)$ (détails dans le texte).

La recherche d'une catégorisation globale des images $\Delta(\Omega)$ selon le signe et le degré d'association spatiale est difficile et présente en fait peu d'intérêt en pratique. En effet, s'il existe une association spatiale significative entre deux images $F(\Omega)$ et $G(\Omega)$, la relation n'est généralement valide que sur une partie de Ω, inconnue *a priori* (Donnay 1994). Ainsi, ce qui apparaît vraiment intéressant, c'est le partitionnement de Ω en régions d'association positive, non significative, et négative. Afin d'obtenir ce type de partition, une solution envisageable consiste à utiliser un *indice de conformité locale* défini en géomatique par Donnay (1994). Cet indice présente l'intérêt (Donnay 1994) :

- d'être calculé à partir de mesures vectorielles dérivées des images $F(\Omega)$ et $G(\Omega)$ (gradient et orientation de pente),
- de rendre compte de l'intensité et du sens de la relation à un niveau local — en pratique au sein d'une fenêtre mobile de 3×3 pixels — et non de manière globale sur Ω,
- de produire une image susceptible d'être partitionnée par seuillage de l'indice de conformité locale.

L'intérêt pratique de cette approche en écologie reste cependant à évaluer à partir d'images simulées et de données réelles.



Figure˜10.6: Exemples d'associations spatiales entre deux images en niveaux de gris. (1), (3) Associations spatiales positives. (2), (4) Associations spatiales négatives. (a) Image $F(\Omega)$. (b) Image $G(\Omega)$. (c) Image des écarts absolus entre $F(\Omega)$ et $G(\Omega)$ (détails dans le texte).

Chapitre 11

Complexité spatiale

> *"La question centrale* [...] *est celle de la mesure de l'organisation spatiale* [de différentes unités] *afin de les comparer et de les ordonner en fonction d'une complexité croissante"* (Chiarello 1994)
>
> *"The common situation is this: An experimentalist performs a resolution analysis and finds a limited-range power law with a value of D smaller than the embedding dimension.* [...] *the experimentalist then often chooses to label the object for which she or he finds this power law a 'fractal'."* (Avnir *et al.* 1998)

Un des objectifs de la théorie écologique est de relier la diversité biologique et la complexité spatiale (Baudry & Baudry-Burel 1982, Scheiner 1992, Chiarello 1994, He *et al.* 1994) : ceci nécessite de définir de façon opératoire, d'une part la diversité biologique, et d'autre part la complexité spatiale.

La diversité biologique est souvent définie à l'aide d'une mesure de diversité appliquée à un ensemble d'espèces (peuplement, taxicoenose, guilde) (voir notamment Hurlbert 1971, Kempton & Wedderburn 1978, Daget 1979, pp. 9-28, Daget 1980, Frontier 1983b, Legendre & Legendre 1984a, pp. 187-204, Frontier & Pichod-Viale 1991, pp. 284-292, Cousins 1991, Solow & Polasky 1994). La mesure de diversité spécifique la plus utilisée est certainement l'*entropie informationnelle* H de Shannon[1] ou *néguentropie* (Brillouin 1959, pp. 147-156) :

$$H = -\sum_{i=1}^{N} P_i \log_2 P_i \tag{11.1}$$

avec N le nombre de modalités d'un événement et P_i la probabilité attachée à la modalité i. Le H de Shannon constitue une forme de mesure de la variabilité d'une distribution statistique (en l'occurrence discrète), l'entropie (11.1) étant fonctionnellement liée à la variance de cette distribution (Mukherjee & Ratnaparkhi 1986).

La diversité spécifique définie par H intègre à la fois le nombre d'espèces N ou *richesse spécifique* et leurs fréquences relatives P_i. La valeur maximale que peut atteindre

[1] Shannon introduit la fonction H en 1948 dans le domaine des télécommunications pour mesurer l'entropie d'un message. Par la suite, le H de Shannon a été étudié en tant que mesure de l'organisation du vivant, notamment par Atlan (1972).

H correspond à l'équifréquence des N espèces, soit $P_i = N^{-1}$ pour $i = 1, \ldots, N$, d'où :

$$H_{\max} = \log_2 N \tag{11.2}$$

Une mesure de redondance caractérisant l'existence de contraintes peut être définie comme :

$$R = 1 - \frac{H}{H_{\max}} \tag{11.3}$$

La redondance (11.3) prend sa valeur minimale $R = 0$ pour $H = H_{\max}$, *i.e.* lorsque toutes les espèces sont équifréquentes, et $R \to 1$ lorsque la plupart des espèces sont rares et que peu d'espèces sont très fréquentes. A la limite, avec la convention $0 \cdot \infty = 0$, la redondance est maximale ($R = 1$) lorqu'une seule espèce est présente parmi les N espèces considérées ($H = 0$).

De façon similaire à la diversité spécifique définie au sens de (11.1), la notion de complexité spatiale intègre à la fois la richesse en unités spatiales différentes et leurs fréquences relatives. Cependant, la complexité spatiale doit également tenir compte de la forme de ces unités, de leur organisation spatiale et du degré de fragmentation de l'ensemble (Baudry & Baudry-Burel 1982, Chiarello 1994).

En fonction du problème posé, l'accent est souvent mis sur un certain aspect de la complexité spatiale. Par exemple, en biologie des populations il est possible de considérer essentiellement **le degré de fragmentation** de l'habitat parce qu'il est fondamentalement lié à la dynamique des métapopulations (Fahrig & Merriam 1994, Hanski *et al.* 1995, Hill & Caswell 1999) ou bien, avec une perspective de conservation, l'accent peut être davantage mis sur **la forme** des réserves naturelles parce qu'elle est en rapport avec les flux d'immigrants (Kunin 1997). En revanche, l'écologie du paysage considère la complexité spatiale de façon globale et propose en conséquence de nombreuses mesures pour quantifier ses différents aspects (Milne 1988, O'Neill *et al.* 1988, Turner 1989, Lagro 1991, Scheiner 1992, Li & Reynolds 1993, 1994, Metzger & Muller 1996, Blackburn & Milton 1996, Geoghegan *et al.* 1997, Gustafson 1998), et notamment la diversité de Shannon (11.1), la redondance[2] (11.3), ainsi que d'autres indices classiquement utilisés en écologie (*e.g.*, Davis & Dozier 1990, Hess & Bay 1997).

Au-delà d'une conception sur le plan phénoménologique, la complexité spatiale doit être définie de façon opératoire, à l'aide d'une formule, ou plus généralement à l'aide d'un algorithme appliqué à la variable régionalisée considérée $z(\cdot)$. Soit un jeu de données $\{z(s_i) \mid i = 1, \ldots, N\}$ décomposé en un ensemble de valeurs $z = \{z_i \mid i = 1, \ldots, N\}$ et un ensemble de supports $s = \{s_i \mid i = 1, \ldots, N\}$. Indépendamment de la structure algébrique de la VR, il convient de distinguer les mesures de complexité qui traitent :

- uniquement des valeurs de z,
- uniquement des supports de s,
- à la fois des valeurs et des supports.

Les mesures qui opèrent uniquement sur z ou bien sur s ne peuvent pas être considérées à proprement parler comme des mesures de la complexité de la VR, mais elles peuvent y participer et ne doivent pas être négligées.

[2]En écologie du paysage, la *redondance* est aussi nommée *dominance* (*e.g.*, Ruuska & Helenius 1996), bien que la *dominance* soit parfois définie comme $D = H_{\max} - H$, *i.e.* sans standardisation par $H_{\max}$ (*e.g.*, O'Neill *et al.* 1988).

En revanche, la mise en relation de z et de s correspond au concept d'autocorrélation spatiale (Chapitre 3). L'autocorrélation spatiale peut être vue comme une forme de contrainte dans une structure spatiale, et les liens avec la mesure de redondance (11.3) issue de la théorie de l'information ont été explorés (Gatrell 1977). Il apparaît donc que les mesures de complexité spatiale portant sur la VR sont étroitement liées aux mesures d'autocorrélation spatiale (Monmonier 1974, Cliff & Ord 1981, p. 13). En effet, une VR positivement autocorrélée peut être vue comme moins complexe qu'une VR non autocorrélée, et d'autant moins complexe qu'elle s'avère spatialement régulière. Cette conception correspond bien à la complexité vue comme une mesure de désordre, et à la perception des écologistes pour lesquels le milieu est d'autant plus complexe que les unités y sont réparties au hasard (Baudry & Baudry-Burel 1982). A partir de ce qui précède, nous pouvons proposer les axiomes suivants :

Axiome 1 *Toute mesure d'autocorrélation spatiale constitue une mesure de la complexité spatiale d'une variable régionalisée.*

Axiome 2 *La complexité spatiale d'une variable régionalisée peut être définie comme l'opposé de sa régularité spatiale.*

Sans considérer les mesures d'autocorrélation spatiale (Chapitre 3), nous traitons dans ce qui suit de la définition opératoire de la complexité spatiale selon le type de carte (Section 2.3) :

- cartes choroplèthes,
- cartes isoplèthes,
- images.

11.1 Complexité des cartes choroplèthes

Soit $\mathcal{C}$ une carte choroplèthe comportant N polygones et G son graphe d'adjacence. En vertu de l'axiome 1, les mesures d'autocorrélation spatiale du type c de Geary et I de Moran calculées sur la base de G et des valeurs $\{z_i \mid i = 1, \ldots, N\}$ constituent des mesures de la complexité de $\mathcal{C}$. Ce type de mesure de complexité spatiale tient compte à la fois des valeurs de la VR et de la topologie de $\mathcal{C}$ qui est décrite par son graphe d'adjacence. Cependant, l'autocorrélation spatiale ne permet pas de discriminer des situations très différentes en termes de géométrie des polygones de $\mathcal{C}$.

Bien qu'il soit possible de concevoir une mesure de complexité géométrique qui porterait sur $\mathcal{C}$ en tant qu'objet composé, dans un premier temps il peut s'avérer suffisant de résumer la complexité géométrique de l'ensemble des polygones qui constituent $\mathcal{C}$. Ce résumé peut se présenter sous la forme de la distribution ou de la diversité des valeurs des attributs des polygones. En particulier le périmètre et l'aire sont des attributs qui apparaissent souvent dans la mesure de la complexité spatiale (Monmonier 1974, Baudry & Baudry-Burel 1982, Blackburn & Milton 1996, Jorge & Garcia 1997, Gustafson 1998). Par exemple, Jorge & Garcia (1997) modélisent la distribution des aires des *patches* de végétation afin de caractériser le degré de fragmentation d'un paysage.

Nous nous intéressons en particulier à la définition opératoire de la complexité géométrique d'un polygone P. Comme les polygones forment une sous-classe des polylignes (Chapitre 2), il peut s'avérer suffisant de traiter le problème général de la complexité géométrique des polylignes. En négligeant le recours aux définitions *ad hoc* (revue dans Monmonier 1974), au moins deux cadres mathématiques permettent de définir la complexité géométrique des polylignes :

- la géométrie fractale[3],
- la géométrie probabiliste.

Dans ce qui suit, nous examinons les deux cadres mathématiques, puis nous illustrons les méthodes retenues à l'aide d'un exemple classique : la courbe de von Koch.

11.1.1 Géométrie fractale

Le recours à la géométrie fractale implique que les objets modélisés par des polylignes soient auto-similaires. Un objet est dit *auto-similaire* s'il est construit à partir d'un même motif de base, qui se répète à l'infini, à toutes les échelles (Voss 1988).

Soit une polyligne $s_1 \rightsquigarrow s_\infty$ comportant un nombre infini de sommets, définie comme une fractale géométrique exacte. La complexité géométrique de cette polyligne peut être mesurée par sa *dimension fractale* (Voss 1988). Il existe principalement deux méthodes pour déterminer la dimension fractale d'une polyligne :

- la méthode du compas,
- la méthode des boîtes.

D'autres méthodes sont revues notamment dans Rigaut (1987), Voss (1988), Sugihara & May (1990), Ramstein & Raffy (1990), Jelinek & Fernandez (1998). Citons en particulier la méthode aire-périmètre (Lovejoy 1982), largement utilisée pour quantifier la complexité du paysage (*e.g.*, O'Neill *et al.* 1988, Haslett 1994, Luque *et al.* 1994, Ruuska & Helenius 1996, Jorge & Garcia 1997, Oleschko *et al.* 1998, Kampichler 1999), mais qui ne peut pas s'appliquer à une seule polyligne (Sugihara & May 1990).

11.1.1.1 Méthode du compas

La longueur L d'une polyligne fractale $s_1 \rightsquigarrow s_\infty$ peut s'exprimer comme une loi puissance fonction de l'échelle de mesure δ (Mandelbrot 1967, Maling 1989 p. 289, Sugihara & May 1990) :

$$L(\delta) = K\delta^{1-D} \tag{11.4}$$

avec K une constante et D est la dimension fractale de $s_1 \rightsquigarrow s_\infty$. Pour une polyligne fractale au sens strict, on a :

$$\lim_{\delta \to 0} L(\delta) = \infty \tag{11.5}$$

la convergence s'effectuant à la vitesse donnée par la loi puissance (11.4).

[3]Les concepts liés aux fractales sont revus notamment dans Voss (1988), Sugihara & May (1990) et Hastings & Sugihara (1993).

La méthode la plus ancienne pour mesurer la dimension fractale D est la *méthode du compas* ou *méthode de Richardson-Mandelbrot* (Maling 1989). Soit δ l'écartement d'un compas ou la longueur d'une règle ; la longueur de $s_1 \rightsquigarrow s_\infty$ peut être approximée comme $L(\delta) = N(\delta)\,\delta$ avec $N(\delta)$ le nombre de fois qu'il faut reporter le compas ou appliquer la règle pour parcourir la polyligne. En répétant cette procédure pour différentes valeurs de δ, il est possible de représenter $\log L(\delta)$ comme une fonction linéaire de $\log \delta$, de pente m. La dimension fractale est alors immédiatement déterminée comme $D = 1 - m$ (Whalley & Orford 1989, Roach & Fowler 1993, Jaggi *et al.* 1993). Des variantes de cette méthode sont revues dans Longley & Batty (1989a, 1989b).

Bien que largement utilisée, la méthode du compas pose un certain nombre de difficultés pratiques. Idéalement, le dernier coup de compas coïncide avec l'extrémité de la polyligne d'où $L(\delta) = N(\delta)\,\delta$, avec $N(\delta)$ entier. En pratique on a plutôt $L(\delta) = N(\delta)\,\delta + \varepsilon$, avec ε une longueur résiduelle (Maling 1989). Il est parfois conseillé d'appliquer la procédure plusieurs fois pour une même valeur δ, à partir de différents points situés sur $s_1 \rightsquigarrow s_\infty$ (Sugihara & May 1990), ce qui se justifie essentiellement dans le cas des polylignes fermées sur elles-mêmes (frontières de polygones). En outre, le graphe log-log n'est en fait jamais linéaire mais concave[4], et s'il apparaît linéaire, c'est parce qu'il comporte peu de points obtenus pour des valeurs δ correspondant aux résolutions extrêmes (Rigaut 1987, Rigaut *et al.* 1998). En général, ce comportement est interprété comme un changement de la dimension fractale avec l'échelle (Sugihara & May 1990), et le graphe log-log est ajusté par plusieurs segments de droites, ce qui donne lieu à une succession de dimensions fractales (*e.g.*, Goodchild 1980, Fig. 1, Whalley & Orford 1989, Fig. 6), ou bien certains points ne sont pas pris en compte dans la régression (*e.g.,* Pennycuick & Kline 1986), ou encore le graphe log-log n'est tout simplement pas figuré (*e.g.*, van Hees 1994).

L'interprétation du changement de dimension fractale selon l'échelle est *ad hoc* et s'avère incohérente. Il convient d'abord de reconnaître qu'une polyligne véritablement fractale ne peut être définie qu'en intention — notamment grâce à une procédure récursive traduisant son auto-similarité — et pas en extension, parce qu'elle comporte un nombre infini de sommets. Ensuite, il est indiscutable que l'application de la méthode du compas avec de nombreuses valeurs de δ sur la représentation finie — donc approximative — d'une polyligne fractale conduit à un graphe log-log non linéaire. Or, il n'y a aucun sens à invoquer le changement de dimension fractale avec l'échelle dans le cas d'un objet qui, même s'il ne constitue qu'une approximation finie, est caractérisé par une dimension fractale unique.

11.1.1.2 Méthode des boîtes

La méthode des boîtes consiste à appliquer une grille régulière de maille δ sur la représentation de la polyligne $s_1 \rightsquigarrow s_\infty$ (rasterisation). Le nombre de mailles $P(\delta)$ contenant un fragment de la polyligne peut s'exprimer comme une loi puissance fonction de la maille δ (Sugihara & May 1990) :

$$P(\delta) = K\delta^{-D} \tag{11.6}$$

avec K une constante et D la dimension fractale de $s_1 \rightsquigarrow s_\infty$. En répétant cette procédure pour différentes valeurs de δ il est possible de représenter $\log P(\delta)$ comme une fonction

[4] De nombreux exemples de graphes log-log concaves sont figurés dans Longley & Batty (1989a, 1989b).

linéaire de $\log \delta$, de pente m. La dimension fractale est alors déterminée comme $D = -m$ (Sugihara & May 1990).

Bien que certainement d'utilisation plus générale que la méthode du compas, la méthode des boîtes pose le même type de difficulté, le graphe log-log s'écartant de la linéarité pour les petites et les grandes valeurs de δ (Boddy *et al.* 1999). La solution *ad hoc* consiste simplement à limiter la régression à la partie linéaire du graphe log-log (Berntson & Stoll 1997, Boddy *et al.* 1999).

11.1.1.3 Critique générale

D'un point de vue strictement technique, Loehle & Li (1996) affirment que l'estimation de la dimension fractale par régression linéaire d'un graphe log-log n'est pas statistiquement valide parce que les hypothèses sous-jacentes ne sont pas respectées. En effet, lorsque la régression linéaire est considérée dans le cadre OLS — modélisant les résidus par des variables aléatoires indépendantes et de même variance — la procédure n'est pas valide dans la mesure où les résidus sont autocorrélés. En conséquence, l'intervalle de confiance généralement associé à la dimension fractale — intervalle de confiance de la pente de la droite de régression — est invalide (Reeve 1992). Une solution à ce problème peut consister à utiliser la régression WLS ou GLS, ou plus simplement, à considérer la régression linéaire d'un point de vue strictement géométrique, sans calcul d'intervalle de confiance.

Avec un relatif succès, With (1994) et Wiens *et al.* (1995) utilisent la méthode du compas pour mesurer la complexité de trajectoires d'insectes (Acrididae, Tenebrioinidae et Formicidae). Mais Wiens *et al.* (1995) considèrent finalement que la dimension fractale n'est pas suffisante et devrait être utilisée conjointement avec d'autres mesures. Par exemple, Anderson *et al.* (1997) caractérisent la trajectoire du nématode *Caenorhabditis elegans* à la fois par la dimension fractale calculée par la méthode du compas et par la distribution des changements d'angles (*turning angle distribution*).

De même, Snover & Commito (1998) et Eshel (1998) semblent utiliser la méthode des boîtes avec succès. En fait, les régressions obtenues par Snover & Commito (1998) et Eshel (1998) portent respectivement sur 7 et 4 points, ce qui permet d'éviter ou de négliger la concavité du graphe log-log, toujours fort embarrassante dans un contexte de régression linéaire.

Le problème posé par la concavité du graphe log-log a conduit notamment à une modification progressive du modèle fractal, qualifié de *semi-fractal* (Rigaut 1987), de *pseudo-fractal* (Whalley & Orford 1988) ou encore d'*asymptotiquement fractal* (Rigaut *et al.* 1998).

De ce qui précède, une question évidente émerge : en dépit d'une large publicité et de très nombreux articles d'application, la géométrie fractale est-elle vraiment adaptée aux objets étudiés ? Sauf dans le cas d'une structure de branchement telle que celle d'une structure de fourragement de fourmis (Theraulaz *et al.* 1994, Fig. 2), du réseau racinaire d'une plante (Berntson & Stoll 1997, Eshel 1998), du mycelium d'un champignon (Boddy *et al.* 1999, Fig. 1 à 6), nous considérons que le concept d'auto-similarité de la géométrie fractale est généralement mal adapté aux objets modélisés par des polylignes (*e.g.*, cours d'une rivière, frontière d'une île, trajectoire d'un animal).

Il convient tout d'abord de reconnaître qu'une polyligne modélisant un objet ou un phénomène naturel n'est jamais auto-similaire au sens strict (Roach & Fowler 1993). En conséquence, le recours à la géométrie fractale ne va pas de soi mais doit être vu comme un véritable acte de modélisation. Dans le contexte de la géométrie fractale, parler de "Nature fractale" constitue un abus de langage qui consiste à confondre les propriétés des phénomènes ou des objets étudiés avec celles d'un modèle mathématique[5], aussi séduisant soit-il *a priori*. Berntson & Stoll (1997) considèrent à juste titre que les méthodes d'estimation de la dimension fractale ont souvent été appliquées de façon non critique, violant les hypothèses à propos de la nature des structures fractales, et conduisant à des résultats apparemment excitants mais qui peuvent être en fait de simples artefacts[6].

Bien qu'il soit classiquement admis que la géométrie fractale constitue un outil unique en son genre et de grand potentiel pour résoudre des questions écologiques importantes (*e.g.*, Kampichler 1999), nous recommandons plutôt de faire appel à la géométrie probabiliste.

11.1.2 Géométrie probabiliste

La mathématique est suffisamment souple et riche pour qu'il soit possible de modeler la définition mathématique de la complexité à l'image de l'objet étudié et non l'inverse. Dans cette optique, Mendès-France (1984, 1987) propose de mesurer la complexité d'une courbe Γ par son entropie S définie comme :

$$S(\Gamma) = -\sum_{n=1}^{\infty} P_n \log P_n \tag{11.7}$$

avec P_n la probabilité qu'une droite aléatoire coupe Γ en exactement n points, et en convenant que $0 \cdot \infty = 0$.

Soit L la longueur d'une polyligne Γ et ℓ la longueur de son enveloppe convexe (Section 2.2.2.3), avec $\ell \leq L$. La géométrie probabiliste permet de calculer le nombre moyen de points d'intersection de Γ avec une droite aléatoire comme (Santaló 1953, pp. 13-14, Steinhaus 1954, Mendès-France 1984, 1987) :

$$\sum_{n=1}^{\infty} n P_n = \frac{2L}{\ell} \tag{11.8}$$

A l'aide de ce résultat, l'entropie $S(\Gamma)$ peut être définie comme (Mendès-France 1984, 1987, Stewart 1990) :

$$S(\Gamma) = \log \frac{2L}{\ell} + \frac{\beta}{\mathrm{e}^{\beta} - 1} \tag{11.9}$$

avec

$$\beta = \log \frac{2L}{2L - \ell} \tag{11.10}$$

Par analogie avec les variables d'état d'un système thermodynamique, il est possible de définir la température $T = \beta^{-1}$, le volume $V = L$ et la pression $P = \ell^{-1}$ d'une

[5] La position épistémologique du modèle et de l'objet est du reste inversée par Mandelbrot *et al.* (1984) qui écrivent : *"it makes good sense to use a metal for modelling a fractal"*.

[6] Concernant le statut empirique des fractales, lire la récente polémique publiée dans *Science* (Avnir *et al.* 1998, Mandelbrot 1998, Pfeifer 1998, Biham *et al.* 1998).

courbe (Mendès-France 1984, Stewart 1990). Dans le cas de la polyligne la plus simple (un segment $s_1 \leftrightarrow s_2$) on a $T = 0$ et $S = 0$. Les variables d'état T et S sont évidemment corrélées, et en pratique il est presque équivalent de considérer l'une ou l'autre. En effet, plus Γ est longue et contorsionnée au sein de son enveloppe convexe, plus elle est "chaude", et plus son entropie est élevée.

En adoptant la même définition que (11.7), il est possible de concevoir l'entropie d'une carte choroplèthe $\mathcal{C}$ dans son ensemble plutôt qu'un résumé de l'ensemble des entropies des polygones qui la composent. A notre connaissance, il n'existe cependant pas de formule analogue à (11.9) définie pour une tessellation. Une solution consisterait à utiliser une méthode de Monte-Carlo afin d'approximer les probabilités P_n en :

- générant un grand nombre de droites aléatoires,
- appliquant un opérateur géomatique calculant l'intersection de chaque droite aléatoire avec les frontières des polygones,
- comptant le nombre d'intersections,
- calculant les fréquences relatives pour chaque nombre d'intersections.

L'opération la plus compliquée sur le plan algorithmique et la plus coûteuse en temps de calcul est évidemment celle qui fait intervenir l'opérateur géomatique d'intersection d'un segment avec les frontières des polygones qui composent $\mathcal{C}$.

11.1.3 Exemple

Considérons les cinq premières itérations mises en oeuvre dans la construction d'objets décrits par la géométrie fractale, par exemple la courbe de von Koch (Fig. 11.1.a) ou le flocon de von Koch (Fig. 11.1.b). En toute rigueur, la courbe ou le flocon de von Koch correspondent à des polylignes infinies, définies par une infinité d'itérations de la procédure de construction. La complexité géométrique de la courbe ou du flocon de von Koch peut être quantifiée par leur dimension fractale $D = \log 4 / \log 3 \simeq 1.262$ (Voss 1988, Sugihara & May 1990, Hastings & Sugihara 1993, p. 21).

La complexité géométrique des polylignes correspondant aux cinq premières itérations de la construction de la courbe de von Koch peut être définie dans le cadre de la géométrie fractale ou dans celui de la géométrie probabiliste. La dimension fractale est calculée en utilisant :

- la méthode du compas en faisant varier δ à partir de $\delta = 1$, par pas de 1, jusqu'à ce que $N(\delta) < 8$ (Fig. 11.2.a),
- la méthode des boîtes à partir d'une capture d'écran d'une résolution de 620×180 pixels, en doublant la taille des pixels jusqu'à la résolution limite de 10×3 pixels (Fig. 11.2.b).

Les valeurs de D sont calculées à partir des pentes m des droites de régression des graphes log-log selon $D = 1 - m$ pour la méthode du compas, et $D = -m$ pour la méthode des boîtes (Tab. 11.1, Fig. 11.3).

Le calcul de l'entropie s'effectue très simplement en construisant l'enveloppe convexe de la polyligne (Section 2.2.2.3), en calculant les longueurs de l'enveloppe convexe et de la polyligne, puis en utilisant la formule (11.9).

Pour les cinq polylignes considérées, la méthode des boîtes donne des valeurs de D systématiquement inférieures à celles obtenues par la méthode du compas (Tab. 11.1). Les valeurs produites par la méthode des boîtes sont trop faibles puisque $D < 1$ pour l'itération 1, et $D \simeq 1$ pour l'itération 2. La méthode des boîtes produit une séquence de valeurs de D qui croît de façon monotone avec le rang de l'itération, tandis que la méthode du compas semble atteindre une dimension moyenne $D \simeq 1.226$ dès la quatrième et la cinquième itération.

Dans le cas des polylignes correspondant aux itérations successives de la construction de la courbe de von Koch, l'enveloppe convexe reste inchangée, et seule la longueur de la polyligne augmente. Le calcul de l'entropie est à la fois simple, exact, et parfaitement cohérent avec l'augmentation de la complexité géométrique de la polyligne au fur et à mesure que le rang de l'itération augmente. En conséquence, afin de mesurer la complexité d'une polyligne, nous recommandons fortement de calculer son entropie au sens de la formule (11.9).

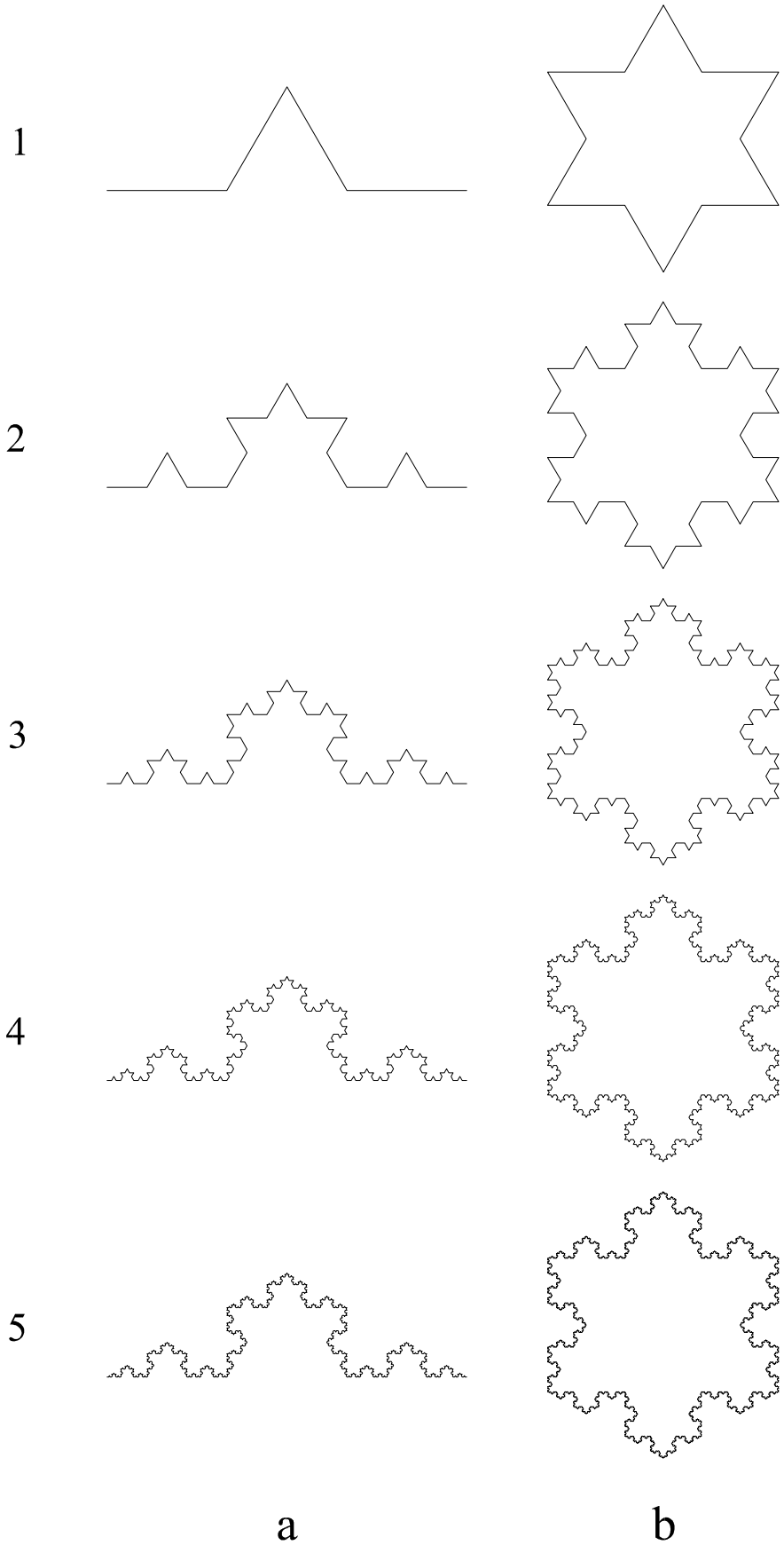
Itération	1	2	3	4	5
m_1	-0.031	-0.094	-0.169	-0.227	-0.225
D_1	1.031	1.094	1.169	1.227	1.225
m_2	-0.966	-1.006	-1.051	-1.132	-1.215
D_2	0.966	1.006	1.051	1.132	1.215
S	0.605	1.106	1.516	1.882	2.224

Tableau 11.1: Mesures de la complexité géométrique des polylignes correspondant au cinq premières itérations de la construction de la courbe de von Koch. m_x, D_x : pente de la droite de régression du graphe log-log et dimension fractale associée dans le cas de la méthode du compas ($x = 1$) et dans le cas de la méthode des boîtes ($x = 2$). S : entropie de la polyligne.

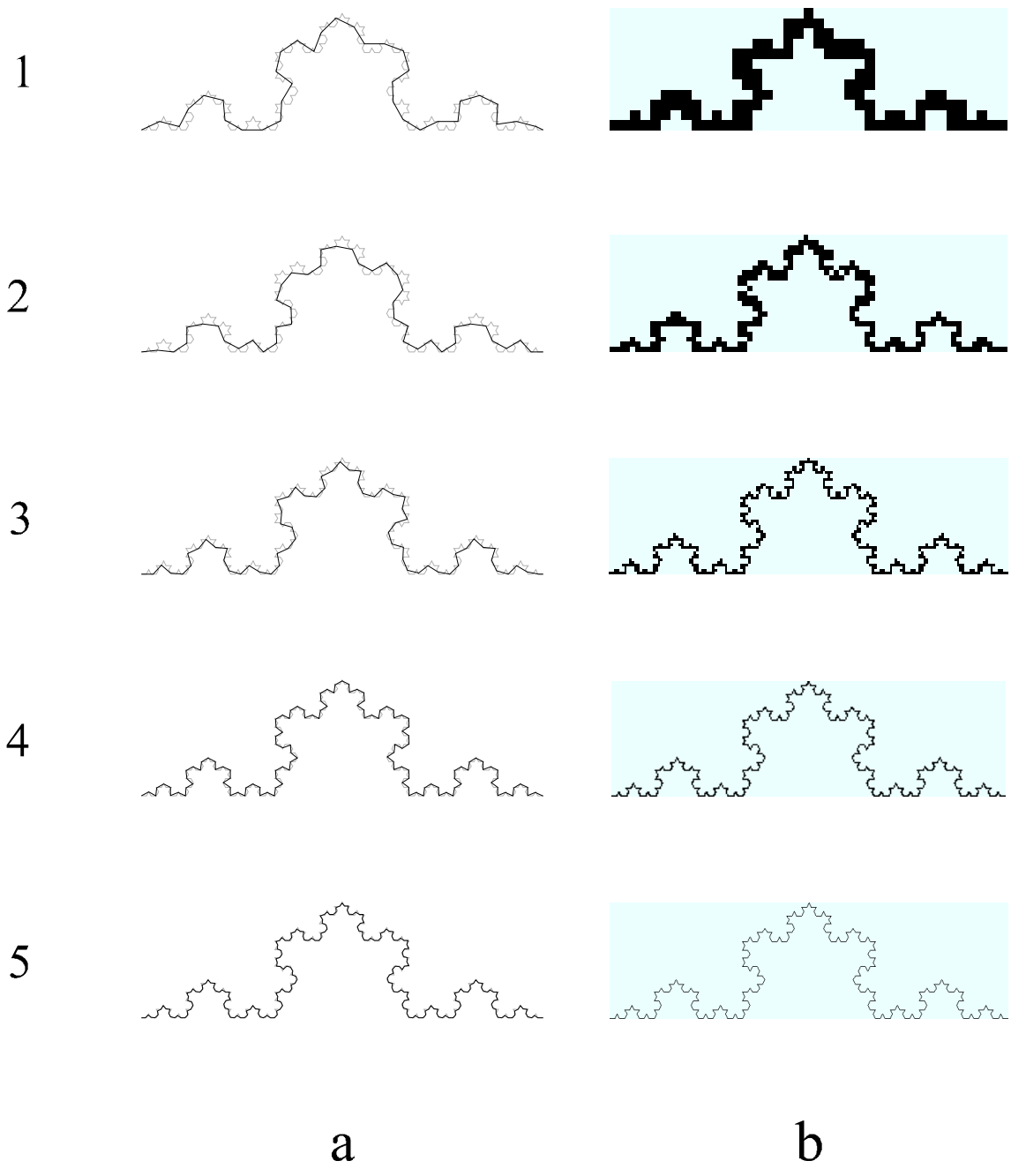
11.2 Complexité des cartes isoplèthes

Intuitivement, on conçoit qu'une carte isoplèthe correspondant à une VR spatialement irrégulière doit présenter de nombreuses isolignes irrégulières, et en conséquence, une droite traversant la carte doit traverser de nombreuses fois les isolignes. On s'attend évidemment à observer l'inverse dans le cas d'une carte isoplèthe correspondant à une VR spatialement régulière. Dans ce contexte, McCarty & Salisbury (1961, *op. cit.* Monmonier 1974) proposent une mesure de complexité composite dont un des éléments est le nombre d'isolignes traversées par les deux diagonales de la carte (en l'occurrence rectangulaire).

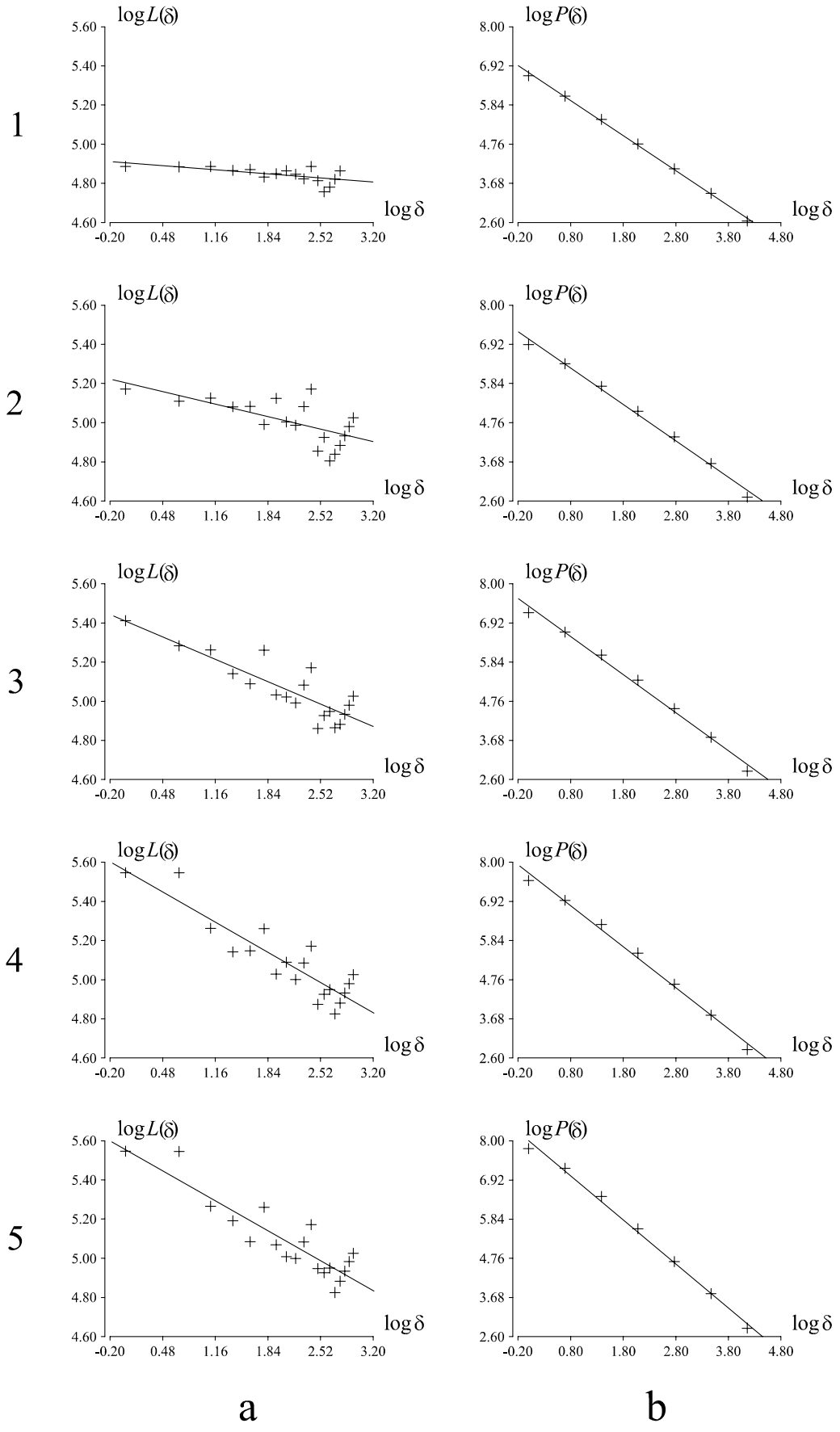
Une carte isoplèthe peut éventuellement être vue comme un cas particulier de carte choroplèthe dans laquelle les polygones définissent des domaines correspondant à des intervalles de cotes. En adoptant ce point de vue, toutes les mesures de complexité définies dans la Section 11.1 peuvent en principe être appliquées aux cartes isoplèthes.



Figure˜11.1: Polylignes correspondant aux cinq premières itérations de la construction de la courbe de von Koch (a) et du flocon de von Koch (b).



Figure˜11.2: Mesure de la dimension fractale de la polyligne correspondant à la quatrième itération de la construction de la courbe de von Koch. (a) Méthode du compas. (b) Méthode des boîtes. (1) à (5) : pour la méthode du compas, diminution de δ de 5 à 1 ; pour la méthode des boîtes, résolutions 39×12 pixels à 620×180 pixels.



Figure˜11.3: Graphes log-log destinés à calculer la dimension fractale pour les cinq premières itérations de la construction de la courbe de von Koch. (a) Méthode du compas. (b) Méthode des boîtes.

Il est notamment possible de calculer l'entropie d'une carte isoplèthe de la même façon que celle d'une carte choroplèthe (Section 11.1.2, p. 336), ce qui correspond *grosso modo* à un développement de l'approche suivie par McCarty & Salisbury (1961).

D'autres mesures de la complexité géométrique peuvent éventuellement être envisagées. Par exemple, une pratique consiste à calculer la dimension fractale des isolignes (*e.g.*, Pennycuick & Kline 1986, Ramstein & Raffy 1990, Jaggi *et al.* 1993). Néanmoins, cette approche ignore le fait que les isolignes résultent généralement d'un traitement des données, de sorte que leur complexité géométrique n'est pas nécessairement représentative de la complexité spatiale de la VR, et *a fortiori* du phénomène régionalisé sous-jacent. Avant de pouvoir définir une mesure de complexité spatiale pour les cartes isoplèthes, il convient donc tout d'abord d'évaluer l'impact de leur mode de calcul.

Les cartes isoplèthes sont généralement calculées pour un domaine D, à partir de valeurs réparties selon une grille (Mallet 1974, Ripley 1981, Davis 1986, Myers 1994b). Soit $\mathcal{A}$ un algorithme de calcul des isolignes à partir d'une grille de valeurs. Il convient de distinguer deux situations selon que :

1. la grille résulte d'un échantillonnage systématique,
2. la grille est obtenue par estimation à partir d'un échantillon irrégulier, par exemple grâce au krigeage.

Dans le premier cas, les données autorisent le calcul direct des isolignes par l'algorithme $\mathcal{A}$, du moins lorsque l'échantillon systématique s'avère suffisamment dense (*e.g.*, grille 30×30). Lorsque le semis des supports n'est pas régulier, le calcul d'isolignes selon $\mathcal{A}$ requiert une étape préliminaire d'estimation de la VR aux noeuds d'une grille aussi dense que l'on voudra. Pour une même VR, mesurée soit sur une grille, soit sur un semis quelconque, il apparaît immédiatement que les cartes isoplèthes calculées par $\mathcal{A}$ seront différentes. En effet, l'étape d'estimation utilisée dans le second cas réduit la variabilité des valeurs (phénomène de lissage), tout en autorisant le choix arbitraire de la résolution de la grille, l'augmentation de la résolution de la grille se traduisant par une complexité croissante de la géométrie des isolignes. En mettant de côté le problème de la représentativité de la complexité géométrique des isolignes vis-à-vis de la complexité spatiale de la VR, il convient d'examiner l'intérêt de la mesure de la complexité géométrique des isolignes.

En géomatique, une carte isoplèthe peut être modélisée de différentes façons afin de faciliter certains traitements, par exemple comme :

- un ensemble de polylignes valuées $\mathcal{L}$ définissant les isolignes,
- un ensemble de polygones $\mathcal{P}$ définissant les domaines compris entre les isolignes.

En conséquence, il est formellement possible de calculer toutes les mesures de complexité géométrique définies pour des polylignes ou des polygones. La question est de savoir si ces mesures sont à la fois utiles et robustes. La représentation sous forme de polylignes ou de polygones conduit à une décomposition de la carte isoplèthe en objets élémentaires, ce qui équivaut à une perte d'information. En outre, les mesures de complexité géométrique opérant sur les polygones ou les polylignes dépendent entièrement de la définition numérique des objets concernés.

Or cette définition est nécessairement très sensible :

- à la densité de l'échantillonnage spatial,
- aux prétraitements (*e.g.*, estimation aux noeuds d'une grille plus ou moins dense),
- à l'algorithme de calcul des isolignes.

Il existe de nombreux programmes de calcul d'isolignes dont les résultats diffèrent. Néanmoins, on peut espérer que les éléments qui traduisent la variation globale de la VR tels que les sommets, les cuvettes, les cols, les colines, les vallées, etc., soient correctement identifiés, même si leur description géométrique est fallacieuse. Nous proposons donc de définir une mesure de complexité spatiale :

1. propre aux cartes isoplèthes, *i.e.* utilisant un modèle géomatique prenant en compte de façon complète leurs propriétés,
2. opérant sur une information de nature topologique plutôt que géométrique, pour une raison de robustesse.

En proposant de mesurer la complexité topologique d'une carte isoplèthe, nous considérons implicitement qu'en écologie il faut *"imaginer une structure spatiale de façon assez qualitative, même si sa mise en évidence a demandé l'usage de moyens lourds, y compris mathématiques et informatiques"* (Legay & Debouzie 1985).

Dans ce qui suit, nous considérons que la topologie de la carte isoplèthe est décrite par le graphe T des relations de voisinage entre isolignes (Section 2.3.3) : il devient alors possible de définir une mesure de complexité topologique.

11.2.1 Première définition de la complexité topologique

Une carte isoplèthe décrit la variation spatiale d'une variable, autrement dit, une forme. Afin de constituer une famille de formes comparables entre elles, les cartes isoplèthes doivent être calculées :

- pour des valeurs standardisées, par exemple en appliquant la transformation

$$z(s_i) = \frac{z(s_i) - z_{\min}}{z_{\max} - z_{\min}} \tag{11.11}$$

pour $i = 1, \ldots, n$, avec

$$z_{\min} = \min_{s_i \in s} z(s_i) \tag{11.12}$$

$$z_{\max} = \max_{s_i \in s} z(s_i) \tag{11.13}$$

et l'ensemble des supports $s = \{s_i \mid i = 1, \ldots, n\}$,

- avec les mêmes cotes (*e.g.*, les cotes $0.0, 0.1, 0.2, \ldots, 1.0$).

En l'état actuel de la topologie, il est difficile de donner une définition mathématique précise de la complexité d'une forme (Thom 1972). Du reste, une définition mathématique présente peu d'intérêt en biométrie si elle ne peut pas se traduire par une définition opératoire, *i.e.* un algorithme. Ainsi, nous cherchons à définir la complexité topologique d'une carte isoplèthe de plus en plus précisément à partir d'une définition générale et assez intuitive de la complexité topologique d'une forme.

Soit f et g deux formes telles que f peut être déformée continuement en g. Considérons que la déformation continue de f en g la plus rapide peut être discrétisée à pas constant par k formes intermédiaires. Il est envisageable de définir une dissimilarité à valeur entière mesurant la difficulté de transformation de f en g, *e.g.* $d(f, g) = k$. Autrement dit, la transformation de f en g la plus rapide nécessite d'autant moins d'étapes intermédiaires qu'elles sont similaires, *i.e.* proches dans l'espace des formes. Il est souhaitable que la dissimilarité entre formes respecte les deux premiers axiomes d'une distance, *i.e.* $d(f, g) = d(g, f)$ (symétrie) et $d(f, g) = 0$ si f et g sont la même forme (réflexivité). S'il est possible de définir une forme de base f_0, *i.e.* une forme topologiquement la plus simple possible, alors la complexité topologique d'une forme f quelconque peut être définie comme $d(f, f_0)$ (Thom 1972). Dans ce contexte, la mesure de la complexité topologique que nous recherchons nécessite :

- le choix d'une forme de base,
- une définition opératoire de la distance entre les formes des cartes isoplèthes.

11.2.1.1 Forme de base

Le choix de la forme de base est assez naturel. La topologie la plus simple pour une carte isoplèthe est celle du gradient, ce qui correspond à un graphe T uniquement constitué d'un chemin $x_1, \ldots, x_n$ où x_1 et x_n sont les cotes extrêmes de la carte isoplèthe. Ce gradient peut être visualisé sous la forme d'une montagne à un seul sommet, ou un flanc de cette montagne, le graphe T étant le même.

11.2.1.2 Distance entre formes

La forme d'une carte isoplèthe est décrite par le graphe qui représente sa topologie (graphe T). Il est naturel de mesurer la distance entre deux formes f et g grâce à une distance d'édition entre leurs graphes T respectifs. Néanmoins, il semble que la conception d'un algorithme opérationnel de distance entre graphes soit encore un problème ouvert (Miclet 1984).

Au lieu de considérer le graphe T qui décrit la topologie des isolignes, il est également possible de manipuler le graphe orienté décrivant la topologie des domaines compris entre les isolignes[7] (*e.g.*, Sircar & Cebrian 1991). Le problème se trouve simplifié dans la mesure où l'on fait l'économie des relations Vois puisque deux domaines adjacents ne peuvent pas être compris entre des isolignes de mêmes cotes α et β. Même dans le cas fort improbable d'une structure en forme de toit parfait, l'isoligne correspondant au sommet du toit est en fait dédoublée en deux isolignes de même géométrie, mais de sens opposés, encadrant

[7] L'ensemble des domaines compris entre les isolignes est un ensemble muni d'une relation d'ordre partiel, *i.e.* un ensemble partiellement ordonné ou *poset* (*partially ordered set*).

un domaine de surface nulle. Ainsi, le graphe de la topologie des domaines est sans cycle et constitue donc un arbre doublement orienté dans lequel il ne subsiste plus que des relations Inf et Sup. Or les relations Inf et Sup sont réciproques de sorte qu'il suffit de connaître l'une pour en déduire l'autre. La topologie d'une carte isoplèthe peut donc être représentée plus simplement par un arbre A simplement orienté qui représente les relations Inf (par exemple).

L'arbre A n'est généralement pas une arborescence. Bien qu'il soit possible de fixer arbitrairement un sommet comme racine (*e.g.*, Mark 1978), cette pratique *ad hoc* nous semble injustifiable *a priori*. En outre, il n'est pas souhaitable que les sommets de cet arbre soient valués par les cotes. En effet, si la forme d'une carte f est identique à celle d'une carte g, alors $d(f, g) = 0$ même si les cotes de f diffèrent de celles de g (*e.g.*, par une translation τ sur l'échelle des cotes). A notre connaissance, le calcul d'une distance d'édition entre des arbres sans racine, orientés et non étiquetés, constitue un problème totalement ouvert.

11.2.2 Seconde définition de la complexité topologique

Le calcul de la distance d'édition entre arbres sans racine, orientés et non étiquetés, étant actuellement hors de notre portée, il faut envisager une mesure de la complexité topologique d'une carte isoplèthe définie dans un autre espace de représentation. Nous considérons successivement l'espace de représentation formé :

- par les matrices d'adjacence des arbres A,
- par les tables des demi-degrés des sommets des arbres A.

11.2.2.1 Matrice d'adjacence

Soit $m_G^+(x, y)$ le nombre d'arcs d'un graphe orienté G ayant x comme sommet initial et y comme sommet final (Section 2.1.2.4, p. 17). A tout graphe orienté G comportant n sommets $\{x_i \mid i = 1, \ldots, n\}$ est associée une matrice d'adjacence d'éléments $a_{ij} = m_G^+(x_i, x_j)$ (Berge 1970). Nous cherchons donc une mesure de complexité C opérant directement sur la matrice d'adjacence $\mathbf{A}$ associée à un arbre A. Comme les sommets de A sont non étiquetés, le résultat de C doit être invariant par permutation des lignes et des colonnes de $\mathbf{A}$. A notre connaissance, seules les valeurs propres et le permanent d'une matrice satisfont à ce prérequis.

Valeurs propres Les valeurs propres de $\mathbf{A}$ sont invariantes par permutation des lignes et des colonnes. En géographie, il a été proposé d'utiliser la valeur propre principale $\lambda_{\max}$ comme une mesure du degré de connectivité globale d'un graphe (*cf.* les références citées par Boots & Royle 1991). Actuellement, nous n'avons pas véritablement d'argument en faveur de l'utilisation de la valeur propre principale de $\mathbf{A}$ comme mesure de complexité topologique de l'arbre associé A. En outre, $\lambda_{\max}$ peut être nulle pour différentes topologies de A, et par conséquent s'avérer dans l'impossibilité de les discriminer.

Permanent Le permanent d'une matrice carrée d'éléments a_{ij} s'écrit (Bouvier & George 1992, p. 566) :

$$\text{per}\,[\mathbf{A}] = \sum_{\sigma} a_{1\sigma(1)} a_{2\sigma(2)} \cdots a_{n\sigma(n)} \tag{11.14}$$

où σ décrit les permutations des indices $i = 1, \ldots, n$. Le permanent présente la propriété d'être invariant pour toute permutation des lignes ou des colonnes (Kaufman 1968, p. 95).

L'utilisation du permanent comme mesure de complexité topologique n'est supportée par aucun argument. En outre, dans le cas des arbres A, si l'absence d'arc entre deux sommets x et y ($m_G^+(x, y) = 0$) n'est pas codée par une valeur non nulle, alors $\text{per}\,[\mathbf{A}] = 0$ pour tout arbre A.

En conséquence, nous ne savons pas définir de façon cohérente la complexité topologique d'une carte isoplèthe dans l'espace de représentation des matrices d'adjacence.

11.2.2.2 Table des demi-degrés des sommets

L'arbre A décrivant la topologie d'une carte isoplèthe permet d'identifier très facilement des éléments tels que les sommets, les cuvettes, les vallées, les cols et d'autres structures plus complexes telles que les cratères, les îles d'un lac, etc. Nous proposons de simplifier le problème en considérant uniquement la présence de ces éléments tout en négligeant leur organisation relative. Cette simplification revient à considérer les caractéristiques des sommets de A pris individuellement plutôt que l'arbre en entier. En fait, notre démarche peut être vue comme le développement formel de l'approche visuelle de McCarty & Salisbury (1961, *op. cit.* Monmonier 1974). En effet, dans leur mesure composite de la complexité d'une carte isoplèthe, McCarty & Salisbury (1961) font intervenir le nombre de sommets et de cuvettes présents dans la carte.

Chaque sommet x de A peut être caractérisé par ses demi-degrés extérieur $d_G^+(x)$ et intérieur $d_G^-(x)$ (Section 2.1.2.4, p. 16). A tout arbre A il est donc possible d'associer une table d^-/d^+ décrivant ses sommets[8], par exemple pour une carte isoplèthe f_0 décrivant un gradient par 5 domaines successifs :

d^-	d^+
0	1
1	1
1	1
1	1
1	0

Considérer les tables d^-/d^+ à la place des arbres représente une certaine perte d'information. En effet, au-delà de $n = 4$, la relation de l'ensemble des arbres vers l'ensemble des tables n'est pas bijective mais surjective. Pour s'en convaincre, il suffit de considérer un cas de petite taille, *e.g.* $n = 5$ (Fig. 11.4.c). On constate qu'il existe plusieurs arbres possibles pour une même table.

Afin d'apprécier la simplification combinatoire que représente le passage des arbres A aux tables d^-/d^+, il faut comparer le dénombrement des arbres sans racine, orientés et

[8] Ici, l'ordre des sommets dans la table n'a aucune importance puisque les arbres ne sont pas étiquetés.

non étiquetés $\tau(n)$ au dénombrement des tables $t(n)$. Le dénombrement $\tau(n)$ est donné par Riordan (1958, p. 138), mais à notre connaissance, le dénombrement combinatoire des tables $t(n)$ n'a pas été publié[9]. Pour dénombrer des structures discrètes sans recourir à l'analyse combinatoire, il est possible de les construire grâce à un algorithme, et de les compter (*e.g.*, Evans *et al.* 1967). A cet effet, nous avons conçu deux algorithmes : le premier utilise un mécanisme de *backtracking* assez compliqué mais efficace, le second est un algorithme naïf, fiable mais inefficace. Les résultats montrent la simplification considérable que représente le passage d'un arbre A à sa table d^-/d^+ (Tab. 11.2). Outre cette simplification combinatoire, la table d^-/d^+ permet de définir des classes d'équivalence topologique constituées par tous les sommets de A ayant les mêmes demi-degrés intérieur et extérieur.

n	1	2	3	4	5	6	7	8	9	10	11
τ	1	1	3	8	27	91	350	1376	5743	24635	108968
t	1	1	3	8	21	52	124	284	629	1352	2829

Tableau 11.2: Dénombrement combinatoire des arbres orientés, sans racine, et non étiquetés (τ), et des tables des demi-degrés correspondantes (t), en fonction du nombre de sommets (n).

La mesure de la complexité topologique d'une carte isoplèthe peut être reformulée en termes de complexité de son arbre A. Un tel arbre apparaît d'autant plus complexe qu'il comporte des sommets de degrés élevés. Dans une telle approche, la forme de base f_0 (gradient) constitue la forme la plus simple, ce qui est cohérent avec la première définition (Section 11.2.1.1). L'examen de la table de f_0 pour $n = 5$ suggère que la ressemblance des demi-degrés est un bon indicateur de la simplicité topologique. Cette ressemblance peut être appréciée à l'intérieur de chaque colonne d^- et d^+, pour tous les sommets, et entre les deux colonnes, pour chaque sommet. En outre, la symétrie[10] des demi-degrés contribue également à la simplicité topologique. Il y a symétrie des demi-degrés lorsque l'ensemble $\mathcal{M}^{-/+}$ des mots $d_i^- d_i^+$ est égal à l'ensemble $\mathcal{M}^{+/-}$ des mots $d_i^+ d_i^-$ avec $i = 1, \ldots, n$. Par exemple, dans le cas de f_0 pour $n = 5$, il y a symétrie des demi-degrés puisque les ensembles $\mathcal{M}^{-/+} = \{01, 11, 11, 11, 10\}$ et $\mathcal{M}^{+/-} = \{10, 11, 11, 11, 01\}$ sont composés des mêmes mots. La complexité topologique $c(f)$ peut donc s'écrire comme $c(f) = w(f) + b(f) - s(f)$ avec :

- $w(f)$ un terme de dissemblance intra-colonne, inter-ligne,
- $b(f)$ un terme de dissemblance inter-colonne, intra-ligne,
- $s(f)$ un terme de symétrie entre demi-degrés.

Le terme $w(f)$ peut être calculé pour tous les couples (i, j) avec $i \neq j \in \{1, \ldots, n\}$ comme :

$$w(f) = \sum_{i<j} \left| d_i^- - d_j^- \right| + \left| d_i^+ - d_j^+ \right| \tag{11.15}$$

[9] Les premiers termes de ce dénombrement figurent désormais à l'entrée A007835 sur le serveur de séquences d'entiers du Dr. Neil J. A. Sloane (AT&T Fellow) (www.research.att.com/~njas/sequences/).

[10] Dans leur mesure composite de la complexité d'une carte isoplèthe, McCarty & Salisbury (1961, *op. cit.* Monmonier 1974) font également intervenir un degré de symétrie visuel de la carte.

Le terme $b(f)$ peut être calculé pour les n sommets comme :

$$b(f) = \sum_i \left| d_i^- - d_i^+ \right| \tag{11.16}$$

Enfin, le terme de symétrie $s(f)$ peut se calculer comme :

$$s(f) = \begin{cases} 1 & \text{si } \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \delta\left(n_{ij}\left(d^-, d^+\right), n_{ij}\left(d^+, d^-\right)\right) = n^2 \\ 0 & \text{sinon} \end{cases} \tag{11.17}$$

avec

$$n_{ij}(d, d') = \sum_{k=1}^{n} \delta(i, d_k)\, \delta(j, d'_k) \tag{11.18}$$

et le delta de Kronecker

$$\delta(\alpha, \beta) = \begin{cases} 1 & \text{si } \alpha = \beta \\ 0 & \text{sinon} \end{cases} \tag{11.19}$$

La fonction $n_{ij}(d, d')$ donne le nombre d'occurrences du mot ij dans la table en considérant les mots formés par la juxtaposition des colonnes d et d'. Appliqué à f_0 pour $n = 5$, les calculs donnent $w(f_0) = 2(n-1)$, $b(f_0) = 2$ et $s(f_0) = 1$. La complexité de la forme de base s'écrit donc $c(f_0) = 2n - 1$, $\forall n$. Comme il est souhaitable d'avoir une complexité topologique nulle pour la forme de base f_0, nous définissons finalement la complexité topologique d'une carte isoplèthe f comme $C(f) = c(f) - c(f_0)$ soit :

$$C(f) = \sum_{i<j} \left| d_i^- - d_j^- \right| + \left| d_i^+ - d_j^+ \right| + \sum_i \left| d_i^- - d_i^+ \right| - s(f) - 2n + 1 \tag{11.20}$$

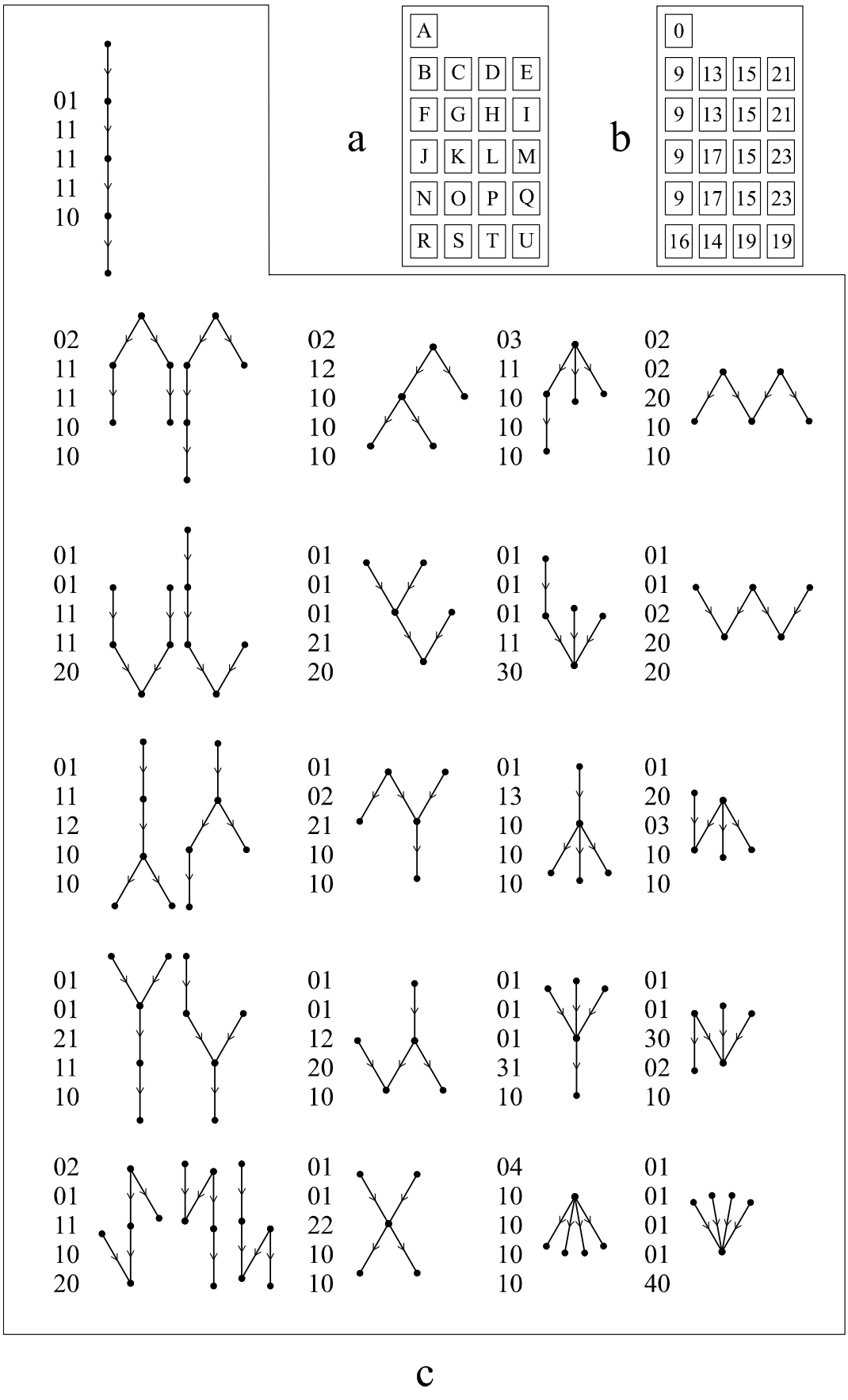
Pour n grand, le terme de symétrie contribue peu à la mesure de la complexité et peut éventuellement être négligé. Appliquée au cas $n = 5$ (Fig. 11.4), la complexité topologique (11.20) donne le classement :

$$A < B, F, J, N < C, G < S < D, H, L, P < R < K, O < T, U < E, I < M, Q \tag{11.21}$$

conforme à l'appréciation intuitive que l'on peut se faire de la complexité topologique des micro-paysages qui sont associés aux arbres (non figurés). La discrimination opérée dans un cas d'aussi petite taille plaide également en faveur de la mesure de complexité (11.20).

11.2.3 Comparaison entre les définitions

Bien que la première définition soit formellement plus séduisante que la seconde, elle n'est pas actuellement opérationnelle. En effet, nous ne connaissons pas d'algorithme calculant la distance d'édition entre arbres sans racine, orientés et non étiquetés. Toutefois, pour $n = 4$, il est possible de calculer les distances d'édition à la main étant donné le petit nombre de sommets et d'arbres mis en jeu (Fig. 11.5). Par ailleurs, nous avons conçu un algorithme calculant une telle distance à partir des tables d^-/d^+.



Figure˜11.4: Complexité topologique C des arbres sans racine, orientés, et non étiquetés, pour $n = 5$ sommets. (a) Identification des différentes tables des demi-degrés des sommets. (b) Valeurs de C pour les différentes tables. (c) Enumération des arbres et des tables.

La relation entre les arbres A et les tables des demi-degrés de leurs sommets étant bijective jusqu'à $n = 4$, il s'avère donc possible de calculer automatiquement la complexité topologique correspondant à la première définition, mais uniquement pour $n = 4$ (Fig. 11.5).

Bien que de portée limitée, la comparaison entre les deux définitions pour le cas $n = 4$ peut donner une idée de leur intérêt relatif. La première définition donne le classement $A < B, C, D, E < F, G, H$, tandis que la seconde définition donne le classement $A < B, C, D, E < F, G < H$. Le pouvoir discriminant de la seconde définition s'avère donc supérieur à celui de la première pour $n = 4$. Il est vraisemblable que ce résultat peut se généraliser à $n \gg 4$, et qu'au fur et à mesure que n augmente, la seconde définition discrimine plus de formes que la première. En conséquence, la seconde définition de la complexité topologique nous semble actuellement très satisfaisante.

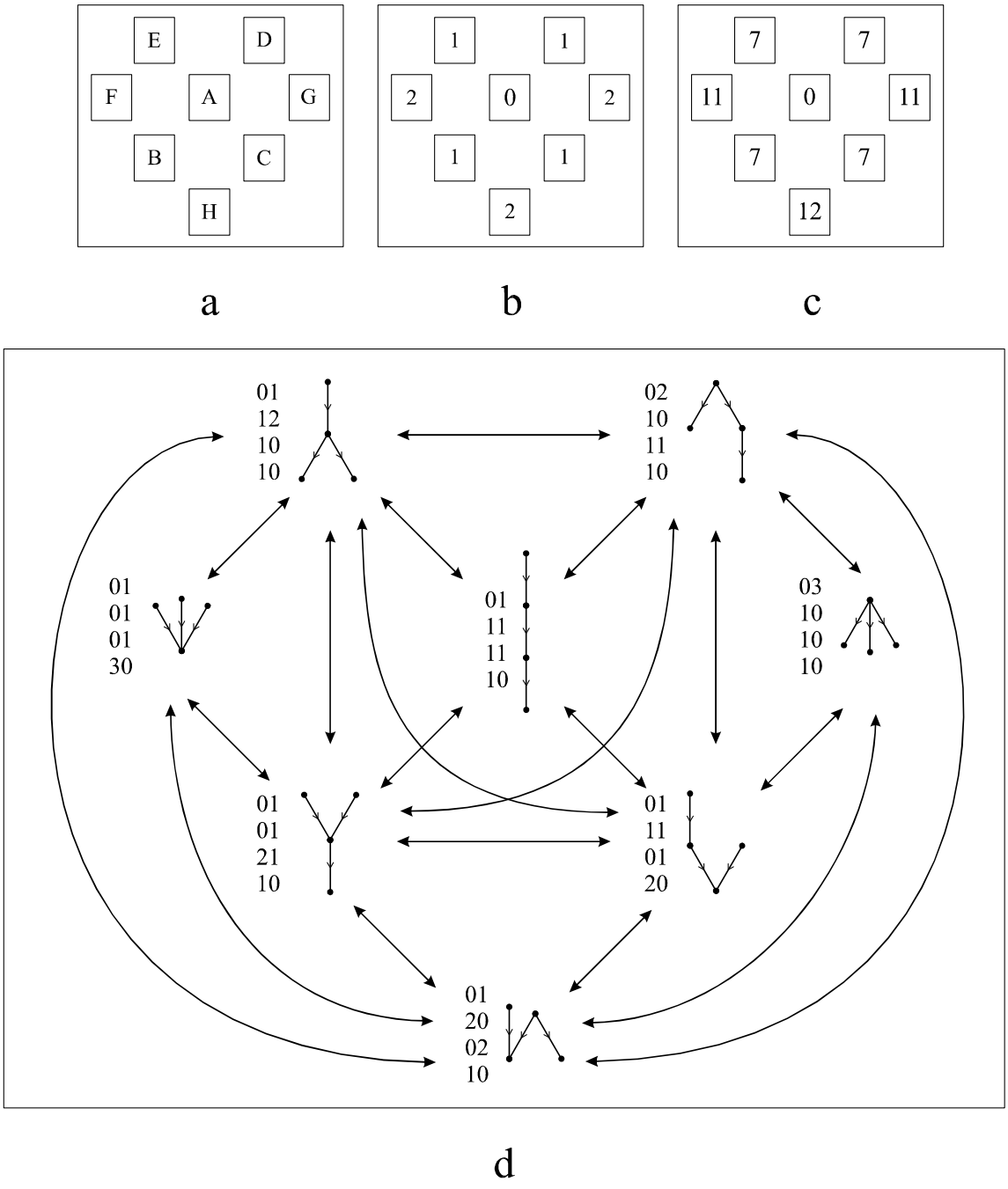
11.2.4 Exemple

L'intérêt pratique de la définition de la complexité topologique que nous avons retenue peut être apprécié grâce à une étude de cas de taille réelle. Nous simulons six modèles de FAST-2 sur une grille 30×30. Les variogrammes théoriques suivent les modèles périodique, gaussien, cubique, sphérique, pentasphérique et exponentiel, avec un paramétrage identique $\theta = (1, 7999, 10)$. Afin de comparer directement l'effet du type de modèle de variogramme théorique, toutes choses étant égales par ailleurs, le générateur de nombres pseudo-aléatoires (Annexe B) est initialisé avec la même graine avant chaque simulation, et les données simulées sont standardisées.

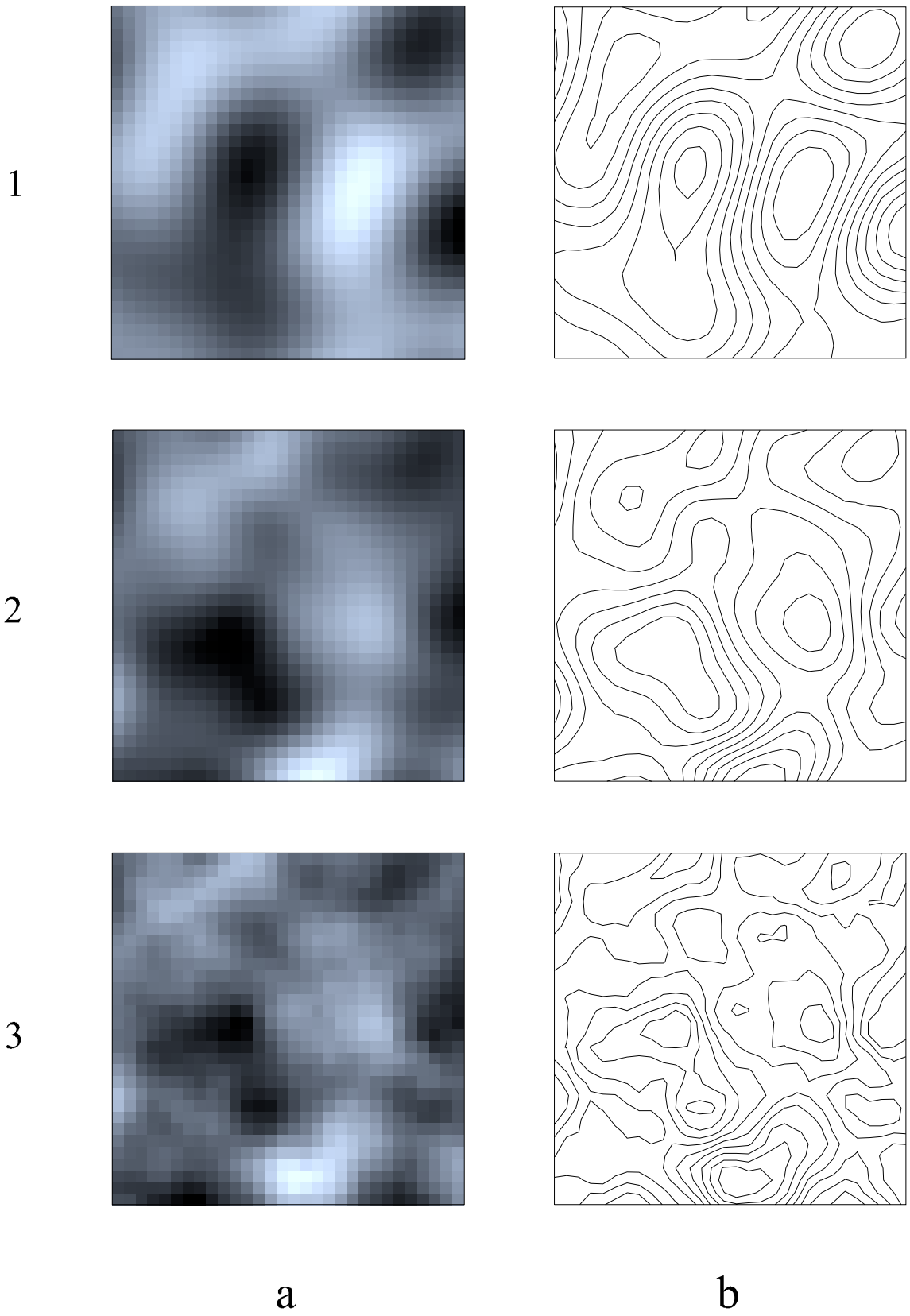
Les cartes isoplèthes sont calculées pour les cotes $0.0, 0.1, \ldots, 1.0$, directement à partir des grilles 30×30, sans aucun lissage qui risquerait de produire des intersections entre isolignes (Ripley 1981, p. 76). Les isolignes comportant moins de 5 points ne sont pas prises en compte. Le classement donné par la mesure de complexité topologique C (11.20) est conforme à l'appréciation visuelle (Tab. 11.3, Fig. 11.6 & 11.7). L'examen visuel des cartes isoplèthes suggère également que le nombre d'isolignes pourrait constituer une première approximation de la complexité topologique (Tab. 11.3). Néanmoins, le nombre d'isolignes ne correspond à aucune définition précise de la complexité topologique, et le bon accord avec le classement donné par C relève trop de la contingence pour pouvoir constituer un indice fiable. En effet, il est facile d'imaginer des situations où des cartes isoplèthes sont de complexités topologiques très différentes tout en ayant le même nombre d'isolignes.

	Périodique	Gaussien	Cubique	Sphérique	Pentasphérique	Exponentiel
C	367	376	975	2169	2709	4613
n	29	26	37	50	54	67

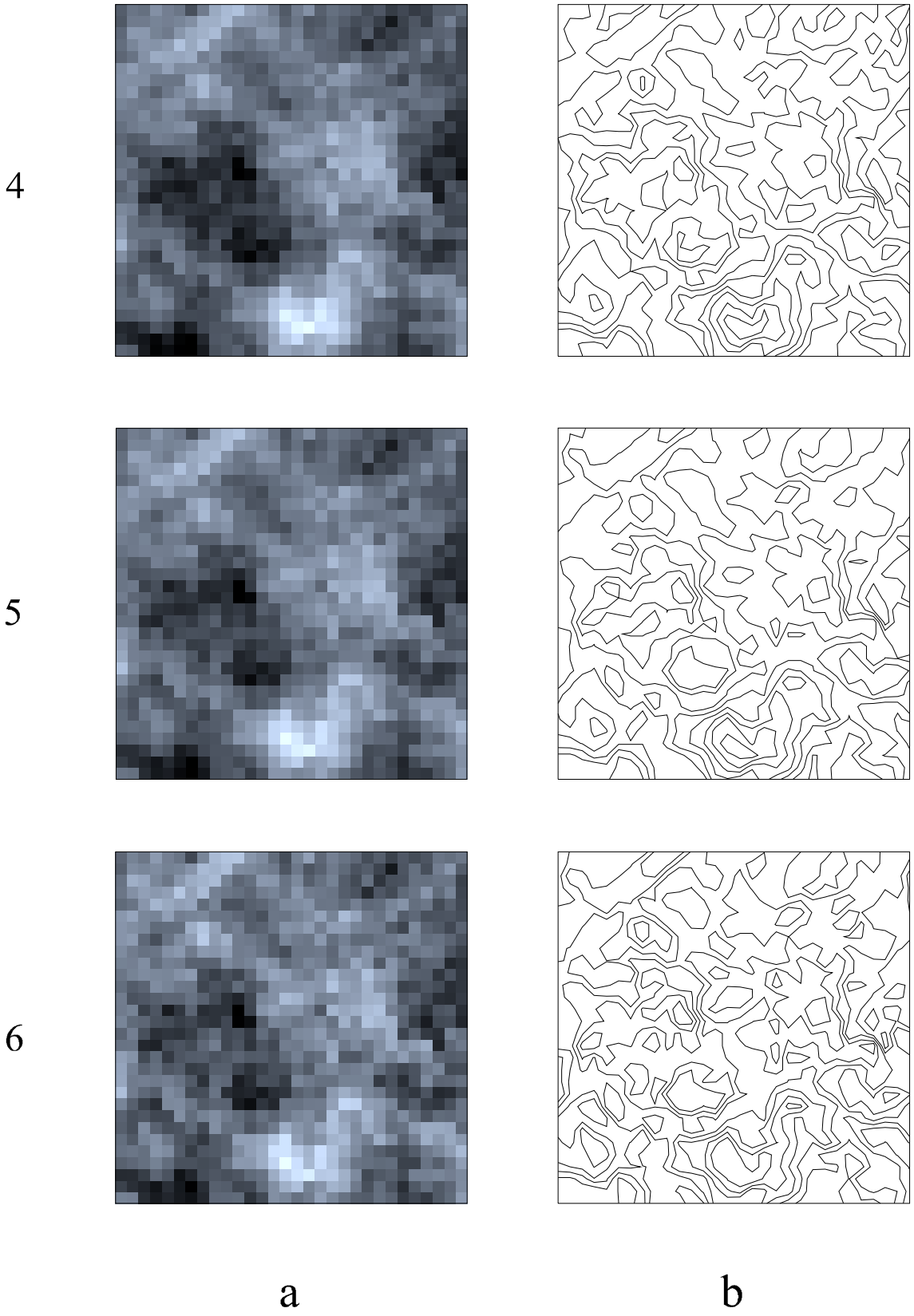
Tableau 11.3: Complexité topologique C des cartes isoplèthes comportant n isolignes, et calculées d'après des données simulées pour six modèles de variogrammes.



Figure˜11.5: Complexité topologique des arbres sans racine, orientés, et non étiquetés, pour $n = 4$ sommets. (a) Identification des différentes tables des demi-degrés des sommets. (b) Valeurs de la complexité topologique selon la première définition. (c) Valeurs de la complexité topologique selon la seconde définition. (d) Graphe d'édition des arbres et des tables correspondantes.



Figure˜11.6: Variables régionalisées simulées sur une grille 30×30 pour trois modèles de variogrammes dont le comportement à l'origine est de type parabolique. (1) Modèle périodique. (2) Modèle gaussien. (3) Modèle cubique. (a) Image 30×30. (b) Carte isoplèthe.



Figure˜11.7: Variables régionalisées simulées sur une grille 30×30 pour trois modèles de variogrammes dont le comportement à l'origine est de type linéaire. (4) Modèle sphérique. (5) Modèle pentasphérique. (6) Modèle exponentiel. (a) Image 30×30. (b) Carte isoplèthe.

11.3 Complexité des images

Nous considérons dans cette section des images correspondant à des VR quantitatives (images en niveaux de gris) ou binaires (images en noir et blanc). Ces images sont des grilles de valeurs qui peuvent être acquises par télédétection, par échantillonnage systématique sur le terrain, ou bien issues d'un calcul (*e.g.*, une simulation). Bien que ces situations soient différentes, pour simplifier, dans tous les cas nous parlerons de *données*.

Il existe plusieurs façons de mesurer la complexité d'une image, selon qu'il est fait appel à un modèle (déterministe ou stochastique) ou bien à une description directe. Dans le cas du recours à un modèle, ce n'est plus la complexité des données qui est directement mesurée mais celle du modèle. Considérons plusieurs images définies à la même résolution et modélisées au sens d'une certaine procédure ; la comparaison de leur complexité s'effectue en comparant la complexité de leurs modèles respectifs. Dans un cadre déterministe, il peut s'agir d'ajuster une surface de tendance à chaque image et de considérer le degré du polynôme ajusté comme un indice de la complexité du modèle, et donc de l'image elle-même. Cependant, cette approche assez simpliste ne s'avère pas fiable en pratique (Monmonier 1974). La comparaison entre modèles stochastiques peut également être envisagée (Chiarello 1994). Pour classer un ensemble de modèles stochastiques selon leur complexité, il nous semble que le problème consiste à :

1. mesurer à quel point deux modèles $\mathcal{M}_1$ et $\mathcal{M}_2$ s'écartent de l'*équivalence stochastique*[11],
2. identifier parmi $\mathcal{M}_1$ et $\mathcal{M}_2$ quel est le modèle dont la complexité est la plus grande.

Dans ce qui suit, nous ne faisons par référence explicitement à la comparaison entre modèles mais plutôt à des procédures descriptives qui peuvent opérer directement sur les données afin de mesurer une "dimension fractale" — dans un sens à préciser — ou bien une complexité définie comme une distance par rapport à une forme de base, dans le cadre d'une décomposition hiérarchique de l'image sous la forme d'un *quadtree*.

11.3.1 Fractales

Les fractales peuvent être divisées en *fractales géométriques* ou *déterministes* telles que la courbe ou le flocon de von Koch (Section 11.1.1, Fig. 11.1), et en *fractales aléatoires* ou *stochastiques* (Saupe 1988).

11.3.1.1 Fractales géométriques

Une première approche consiste à voir l'image de façon géométrique comme une population de parallélépipèdes rectangles : la base de chaque parallélépipède rectangle est formée par le pixel, et sa hauteur est proportionnelle à la valeur associée au pixel. Dans ce contexte, il est envisageable de recourir aux définitions opératoires de la dimension fractale géométrique, par exemple en utilisant la méthode des boîtes étendue au domaine

[11] La notion d'*équivalence stochastique* est définie dans Prokhorov (1995).

tridimensionnel (x, y, z) afin de déterminer une dimension fractale $2 < D < 3$ (Sugihara & May 1990, Vedyushkin 1994). Une première difficulté provient du fait que les unités des coordonnées (x, y) et l'unité de la variable z sont généralement différentes (Maurer 1994, p. 70), un contre-exemple étant fourni par la topographie. En outre, l'utilisation de la méthode des boîtes suppose l'isotropie parce que les côtés orthogonaux de la grille sont supposés avoir la même relation vis-à-vis de la variabilité spatiale selon les deux axes (Milne 1991).

11.3.1.2 Fractales stochastiques

La plus simple des fractales stochastiques correspond au mouvement Brownien unidimensionnel. Le mouvement Brownien est une FAI-0 stricte dont les incréments ont une distribution gaussienne et sont *statistiquement auto-similaires* au sens où (Voss 1988, Saupe 1988) :

$$Z(t_0 + h) - Z(t_0) \tag{11.22}$$

et

$$\frac{1}{r^H}\left[Z(t_0 + rh) - Z(t_0)\right] \tag{11.23}$$

ont la même distribution conjointe, pour tout t_0 et $r > 0$, avec $H = 0.5$. Autrement dit, $Z(h)$ et $\frac{1}{r^H} Z(rh)$ sont deux FAI-0 *stochastiquement indistingables*[12].

Il existe une relation de proportionnalité entre la variance des incréments et l'écart entre les supports h (Saupe 1988, Wen & Sinding-Larsen 1997) :

$$\mathrm{E}\left[\{Z(t_0 + h) - Z(t_0)\}^2\right] \propto h^{2H} \tag{11.24}$$

Par définition du variogramme d'une FAI-0, la relation de proportionalité (11.24) peut s'écrire :

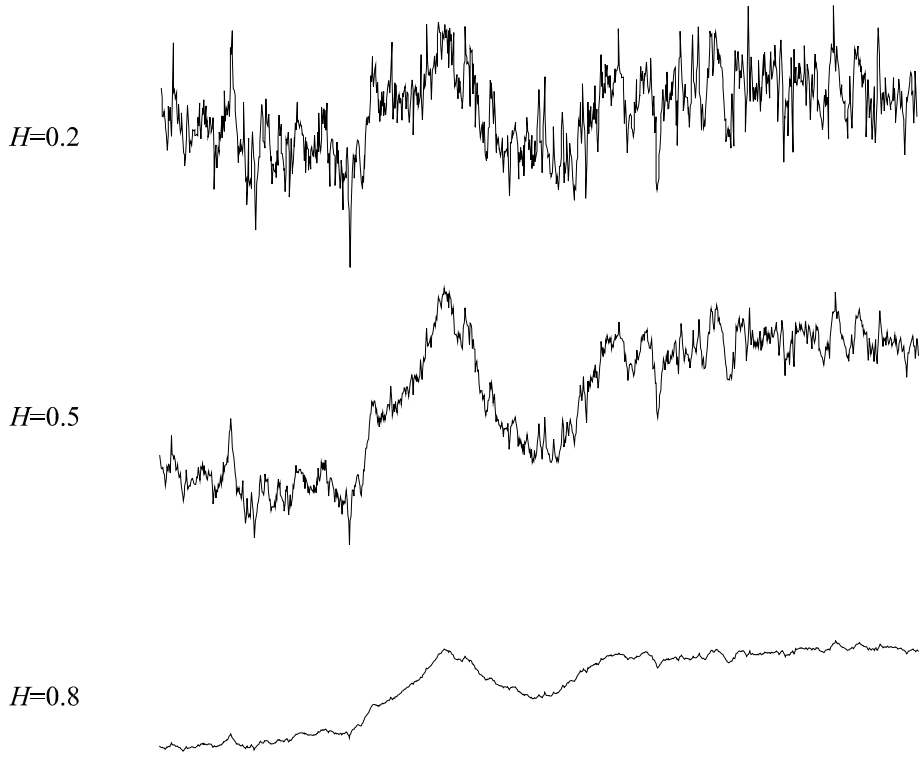
$$\gamma(h) \propto h^{2H} \tag{11.25}$$

La généralisation des valeurs de H à l'intervalle $[0, 1]$ définit les FAI-0 du type fBm (*fractional Brownian motion*) (Saupe 1988). Les FAI-0 de type fBm sont évidemment généralisables à l'espace bidimensionnel.

En ce qui concerne la structure d'autocorrélation des incréments des FAI-0 de type fBm, trois situations peuvent être distinguées selon la valeur de H (Saupe 1988, p. 84, Sugihara & May 1990, Marshall *et al.* 1998) :

- pour $H < 0.5$ (anti-persistance), autocorrélation négative des incréments (*e.g.*, Fig. 11.8, $H = 0.2$),
- pour $H = 0.5$ (mouvement Brownien), pas d'autocorrélation des incréments (*e.g.*, Fig. 11.8, $H = 0.5$),
- pour $H > 0.5$ (persistance), autocorrélation positive des incréments (*e.g.*, Fig. 11.8, $H = 0.8$).

[12] La propriété d'*indistingabilité stochastique* est définie dans Shiryaev (1995).



Figure˜11.8: Réalisations de fBm unidimensionnels définis pour $H = 0.2$, $H = 0.5$, et $H = 0.8$.

Corrélativement, une valeur $H \to 0$ correspond à une FA très irrégulière tandis qu'à l'autre extrême, une valeur $H \to 1$ correspond à une FA très régulière (Voss 1988). Notons que la limite inférieure $H = 0$ correspond à un effet de pépite pur tandis que la limite supérieure $H = 1$ est compatible avec le fait que le variogramme ne croît pas plus vite que h^2 (Section 4.2.2.2, p. 74).

L'exposant H est nommé *exposant de Hurst*[13] (Milne 1991, Hastings & Sugihara 1993, p. 25, Wen & Sinding-Larsen 1997) ou *codimension fractale* (Hewett 1993, Eghball 1999) car H est relié à la dimension fractale D par (Wen & Sinding-Larsen 1997) :

$$D = D_{top} + 1 - H \tag{11.26}$$

où D_{top} est la dimension topologique de l'espace des supports, soit $D = 3 - H$ dans un espace bidimensionnel (exemples dans Voss 1988, Ramstey & Raffy 1989, Milne 1992, Roach & Fowler 1993, Jaggi *et al.* 1993, He *et al.* 1994, Zhang & Selinus 1997, Bonetto & Ladaga 1998, Bellehumeur & Legendre 1998). Ainsi, une surface très irrégulière ($H \to 0$) tend à remplir complètement un volume ($D_{top} = 3$) et présente une dimension fractale $D \to 3$, tandis qu'une surface très lisse, sans aspérité, sans relief marqué, tend vers le plan ($D_{top} = 2$) et présente une dimension fractale $D \to 2$ (Goodchild 1980, Fig. 3, Bonetto & Ladaga 1998). La situation extrême de l'effet de pépite pur ($H = 0$) conduit à une dimension fractale $D = 3$, ce qui correspond bien à la situation de complexité maximale.

L'exposant de Hurst peut être estimé en traçant le logarithme de $\gamma(h)$ en fonction du logarithme de h afin de linéariser la relation de proportionalité (11.25). La pente

[13] L'exposant H est désigné de façon erronée comme la *mesure d'Hausdorff* par Shook & Gray (1996).

de la droite ajustée au graphe log-log est $m = 2H$, d'où l'on déduit H, puis D par la relation (11.26). Il est également possible d'estimer la dimension fractale stochastique D en utilisant une méthode spectrale (*e.g.*, Wen & Sinding-Larsen 1997, Critten 1997), mais le variogramme constitue l'outil le plus pratique (Carr & Benzer 1991).

11.3.1.3 Critique générale

Force est de constater que la variété des applications des fractales et des modèles correspondants rend le sujet particulièrement confus (Gardner 1998, p. 26). Cette confusion est en partie causée par le fait que Mandelbrot n'a pas donné de définition précise de ce qu'est une fractale mais plusieurs définitions — non équivalentes — de la notion de dimension (Mendès-France 1987). Ainsi, au moins dix types de dimensions fractales ont été proposés dans la littérature (Takayasu 1990, *op. cit.* Wen & Sinding-Larsen 1997). Roach & Fowler (1993) considèrent à juste titre que l'absence de définition unique des fractales constitue un des problèmes majeurs de l'analyse fractale.

L'état de confusion causé par les différentes définitions opératoires de la dimension fractale est entretenu par la divergence des résultats obtenus. Par exemple, Carr & Benzer (1991) ne voient aucune raison d'attendre une corrélation entre la dimension fractale géométrique et la dimension fractale stochastique, tandis que Vedyushkin (1994) affirme que toutes les méthodes sont mathématiquement équivalentes mais qu'en pratique elles peuvent donner des valeurs de D différentes pour les mêmes données. En outre, d'après Loehle & Li (1996), il n'existe pas de théorie statistique de l'estimation de la dimension fractale qui permettrait, notamment, de tester si deux valeurs de D sont significativement différentes.

La confusion provient également d'un certain flou dans la littérature concernant la signification de l'auto-similarité et de l'auto-affinité, exactes ou statistiques. Milne (1991) note par exemple que la distinction entre l'auto-affinité et l'auto-similarité statistique est "subtile", sans préciser en quoi consiste cette subtilité. En fait, l'auto-similarité et l'auto-affinité exactes caractérisent des objets géométriques qui peuvent être décrits comme le résultat de l'union de copies d'eux-mêmes, à différentes échelles. Selon que le changement d'échelle est isotrope ou anisotrope, il convient de parler d'auto-similarité ou d'auto-affinité (Voss 1988, pp. 59-62, Hastings & Sugihara 1993, p. 16). L'auto-affinité est par conséquent plus générale que l'auto-similarité. L'auto-similarité et l'auto-affinité sont qualifiées de *statistiques* lorsque la ressemblance à différentes échelles d'un objet avec lui même n'est qu'approximative, et d'*exactes* dans le cas contraire.

Le cas de l'auto-similarité (ou de l'auto-affinité) exacte est assez clair dans la mesure où il s'agit d'un concept géométrique qui n'a pas de contrepartie dans le monde réel. Nous avons critiqué le recours à la géométrie fractale pour mesurer la complexité de la plupart des objets écologiques modélisés par des polylignes parce qu'ils n'étaient pas auto-similaires (Section 11.1.1.3). Nous étendons cette critique aux surfaces modélisées par des images et faisons peu de cas de l'utilisation de la géométrie fractale dans ce contexte (*cf.* Jaggi *et al.* 1993). En revanche, le concept de fractale stochastique apparaît adéquat parce qu'il permet de mesurer la complexité sans avoir à faire référence à des objets géométriques bien définis (Burrough 1981, Roach & Fowler 1993). Par exemple, Palmer (1988) utilise le variogramme en écologie végétale en insistant sur le fait que la structure spatiale de la végétation est fractale tout en ne présentant pas la propriété d'auto-similarité.

Dans le cas des fractales stochastiques, il convient de préciser s'il s'agit : (a) d'estimer la dimension fractale théorique d'une FA à partir d'une VR vue comme une réalisation de cette FA, (b) de mesurer la complexité d'une VR quelconque.

Dans le cas (a), un modèle fractal stochastique du type fBm est compatible avec des données présentant un variogramme qui croît de façon monotone (Phillips 1985). En revanche, ce modèle n'est plus approprié lorsque le variogramme ne croît pas de façon monotone mais plutôt par paliers successifs (Burrough 1983).

Dans le cas (b), la complexité spatiale définie par D *via* l'exposant de Hurst H ne constitue pas autre chose qu'un résumé du variogramme expérimental, traduisant à la fois l'intensité de l'autocorrélation spatiale et la régularité spatiale de la VR, conformément à nos axiomes 1 et 2 (p. 331). Il n'est donc pas surprenant que la dimension fractale puisse être vue comme un indice de dépendance spatiale de la VR (*e.g.,* Palmer 1988).

En pratique, le calcul de la dimension fractale d'après le variogramme pose quelques difficultés (Palmer 1988, Cullinan & Thomas 1992), et dans certains cas, il arrive même que D dépasse la dimension topologique supérieure (Palmer 1988). Il est évident que le calcul de D ne pose pas de problème dans le cas d'une réalisation d'une FAI-0 de type fBm. En revanche, des difficultés surviennent lorsqu'il s'agit de traiter des variogrammes rencontrés en pratique, qui présentent généralement un ou plusieurs paliers, parce que dans ce cas le graphe log-log n'est pas linéaire (*e.g.*, Burrough 1983, Fig. 6, Jaggi *et al.* 1993, Fig. 9, Leduc *et al.* 1994, Fig. 3, Maurer 1994, Fig. 4.6, p. 74).

En conséquence, Valdez-Cepeda & Olivares-Sáenz (1998) affirment que la dimension fractale ne peut pas être calculée si le variogramme ne suit pas un modèle puissance (modèle sans seuil). En fait, lorsque l'estimation de la dimension fractale s'avère mal définie, le variogramme fournit néanmoins l'information nécessaire pour caractériser la complexité, mais il est impossible de la quantifier globalement au moyen d'une seule valeur de D (Phillips 1985). Dans ce contexte, un variogramme de transition peut être vu grossièrement comme la succession d'un modèle de type puissance — parabolique dans le cas des modèles périodique, gaussien et cubique, et linéaire dans le cas des modèles pentasphérique, sphérique et exponentiel — d'une zone de transition plus ou moins brutale, puis d'un effet de pépite pur au-delà de la portée de l'autocorrélation spatiale. L'existence d'une structure emboîtée conduit à répéter la succession précédente et traduit bien la rupture d'échelle du phénomène régionalisé sous-jacent.

Dans le cas classique d'un variogramme présentant un seuil, la partie approximativement linéaire du graphe log-log s'étend jusqu'à la zone de transition (*e.g.*, Bellehumeur & Legendre 1998, Fig. 1). En conséquence, Phillips (1985) trace le graphe log-log uniquement pour les distances en deçà de la portée du variogramme. Dans cette optique, nous proposons de limiter le tracé du graphe log-log aux toutes premières classes de distances afin de résumer le comportement à l'origine du variogramme, ce qui revient à quantifier l'intensité de l'autocorrélation (Sarnelle *et al.* 1993) et la régularité spatiale de la VR, autrement dit, la complexité spatiale à petite échelle (Burrough 1981).

En proposant cette pratique, nous nous inscrivons dans la démarche qui consiste à synthétiser de plus en plus la structure spatiale de la VR, en calculant successivement la nuée variographique (Section 7.1.2), les h-scattergrammes (Section 7.1.3), le variogramme $\gamma(h)$ (Section 7.1.3, p. 192), et enfin l'exposant de Hurst[14].

[14] D'autres indices ont été proposés par Linden & van Doren (1986, *op. cit.* Guillobez & Arnaud 1998).

11.3.2 Approche hiérarchique

Au lieu de partitionner la variance de l'image en classes de distances à l'aide du variogramme, puis de résumer le comportement à l'origine du variogramme sous la forme d'un indice tel que l'exposant de Hurst, il est possible d'envisager une décomposition hiérarchique de la variance.

Dans le contexte spatial, la décomposition hiérarchique de la variance est une approche ancienne puisqu'elle constitue la base des techniques de quadrats contigus utilisées en écologie végétale depuis Greig-Smith (1952). Palmer (1988) note très justement que ce type de technique présente des similitudes avec le variogramme bien que la décomposition de la variance soit différente. En effet, dans le cas des techniques de quadrats contigus, la variance est exprimée en fonction du niveau d'agrégation des quadrats tandis que le variogramme exprime classiquement la variance en fonction de la distance entre points (Ver Hoef *et al.* 1993). Toutefois, en adoptant certaines conventions, il est possible de définir un variogramme qui soit une fonction du niveau d'agrégation des quadrats, et d'établir ainsi des liens étroits entre les deux approches (Ver Hoef *et al.* 1993).

Indépendamment de son utilisation en écologie végétale, l'approche hiérarchique est appliquée à une image carrée en contrôlant la construction d'un *quadtree* par les tests F de l'ANOVA hiérarchique associée (Csillag & Kabos 1996). Sans discuter de la validité statistique des tests F de l'ANOVA hiérarchique dans ce contexte (*cf.* Pielou 1969, p. 105, Chessel 1978, Ripley 1981, p. 109), il est évident que la variance constitue une des statistiques les plus appropriées pour servir de critère d'homogénéité lors de la construction du *quadtree* d'une image, le test F pouvant éventuellement être utilisé d'un point de vue heuristique.

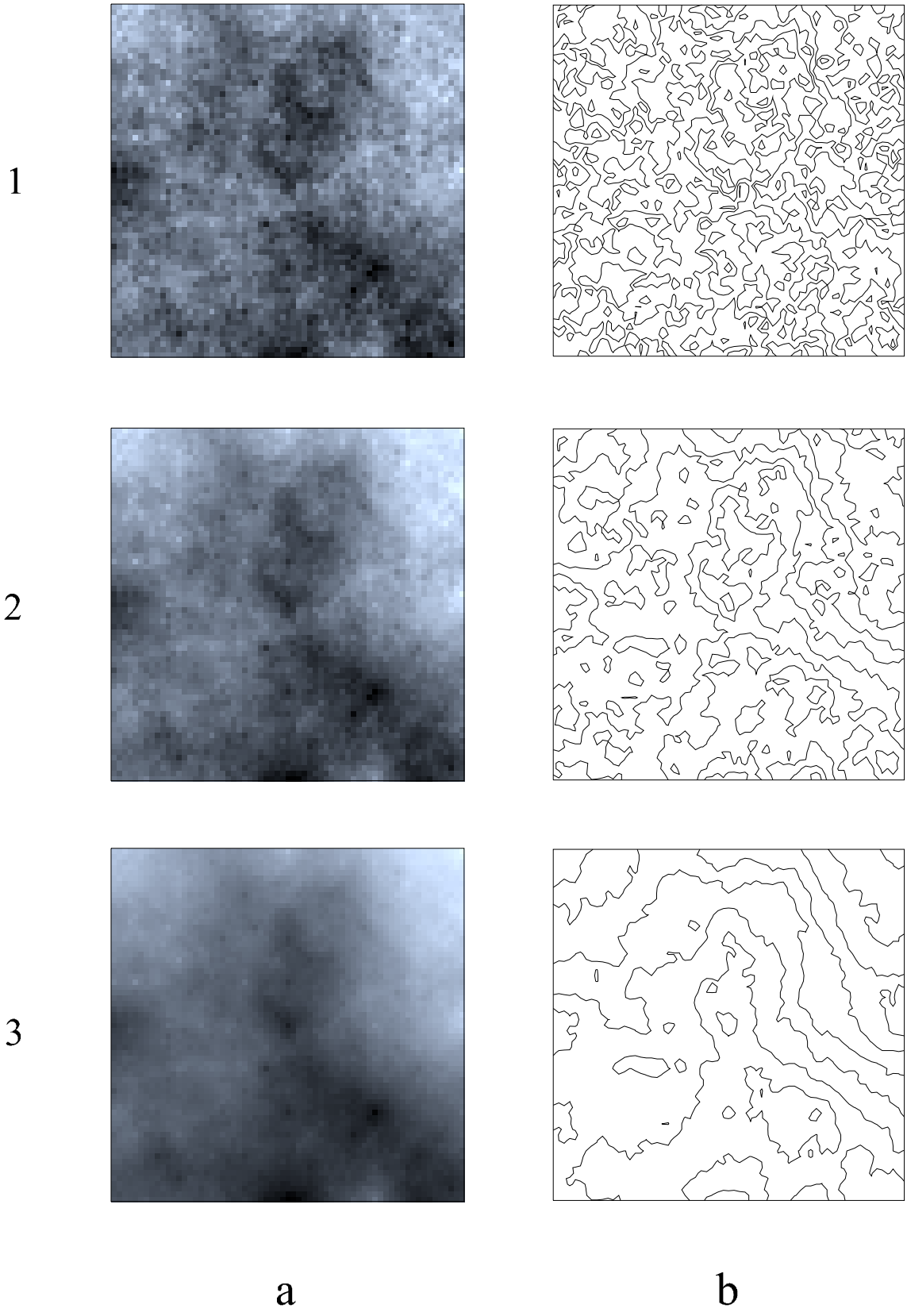
L'association d'un *quadtree* à une image permet de changer d'espace de représentation, de la même façon que nous l'avons exposé dans la Section 10.3.1. Dans cet espace de représentation, au moins deux définitions de la complexité peuvent être envisagées :

- la distance du *quadtree* par rapport à une forme de base, dans le même esprit que ce qui est exposé dans la Section 11.2.1, la distance la plus naturelle étant la distance d'édition de Selkow (1977) décrite dans la Section 10.3.2,
- l'entropie du *quadtree*, au sens d'une mesure de complexité de tous les chemins allant de la racine jusqu'aux feuilles (Green 1973).

Bien que ces deux approches soient plus compliquées que le calcul de l'exposant de Hurst à partir du variogramme, elles méritent toutefois d'être explorées plus avant.

11.3.3 Exemple

Considérons l'approximation d'une réalisation d'une FA de type fBm bidimensionnel par la méthode de déplacement décrite par Saupe (1988, pp. 96-101). Nous générons une image de résolution 65×65 et de pas $\Delta \simeq 0.015625$ pour trois valeurs de l'exposant de Hurst : $H = 0.2$, $H = 0.5$ et $H = 0.8$ (Fig. 11.9.1a, 11.9.2a & 11.9.3a).



Figure˜11.9: Variables régionalisées simulées sur une grille 65×65 pour trois valeurs de l'exposant de Hurst. (1) $H = 0.2$. (2) $H = 0.5$. (3) $H = 0.8$. (a) Image 65×65. (b) Carte isoplèthe.

Le variogramme local est calculé pour 45 classes de distances de pas Δ et de tolérance $\varepsilon = \Delta/2$. Le variogramme montre un comportement à l'origine convexe pour $H = 0.2$ (Fig. 11.10.1a), très légèrement concave pour $H = 0.5$ (Fig. 11.10.2a), et concave pour $H = 0.8$ (Fig. 11.10.3a).

Afin de quantifier la régularité spatiale de la VR dans chaque cas, seules les 10 premières classes de distances[15] sont utilisées afin de tracer $\log \gamma(h)$ en fonction de $\log h$. Une droite est ajustée au graphe log-log en minimisant les moindres carrés ordinaires (Fig. 11.10.1b, 11.10.2b & 11.10.3b). La pente m de la droite de régression augmente progressivement avec H (Tab. 11.4). L'exposant de Hurst observé H_{obs} estime correctement la valeur théorique H, avec toutefois une surestimation pour $H = 0.2$, et une sous-estimation pour $H = 0.5$ et $H = 0.8$, ce qui peut être attribué à la méthode ou, plus vraisemblablement, à l'erreur de fluctuation (Tab. 11.4). Corrélativement, la dimension fractale estimée $2 < D_{obs} < 3$ diminue à mesure que la régularité spatiale de la VR (et de la FA) augmente (Tab. 11.4).

H	m	H_{obs}	D_{obs}
0.2	0.552	0.276	2.724
0.5	0.949	0.475	2.525
0.8	1.375	0.688	2.312

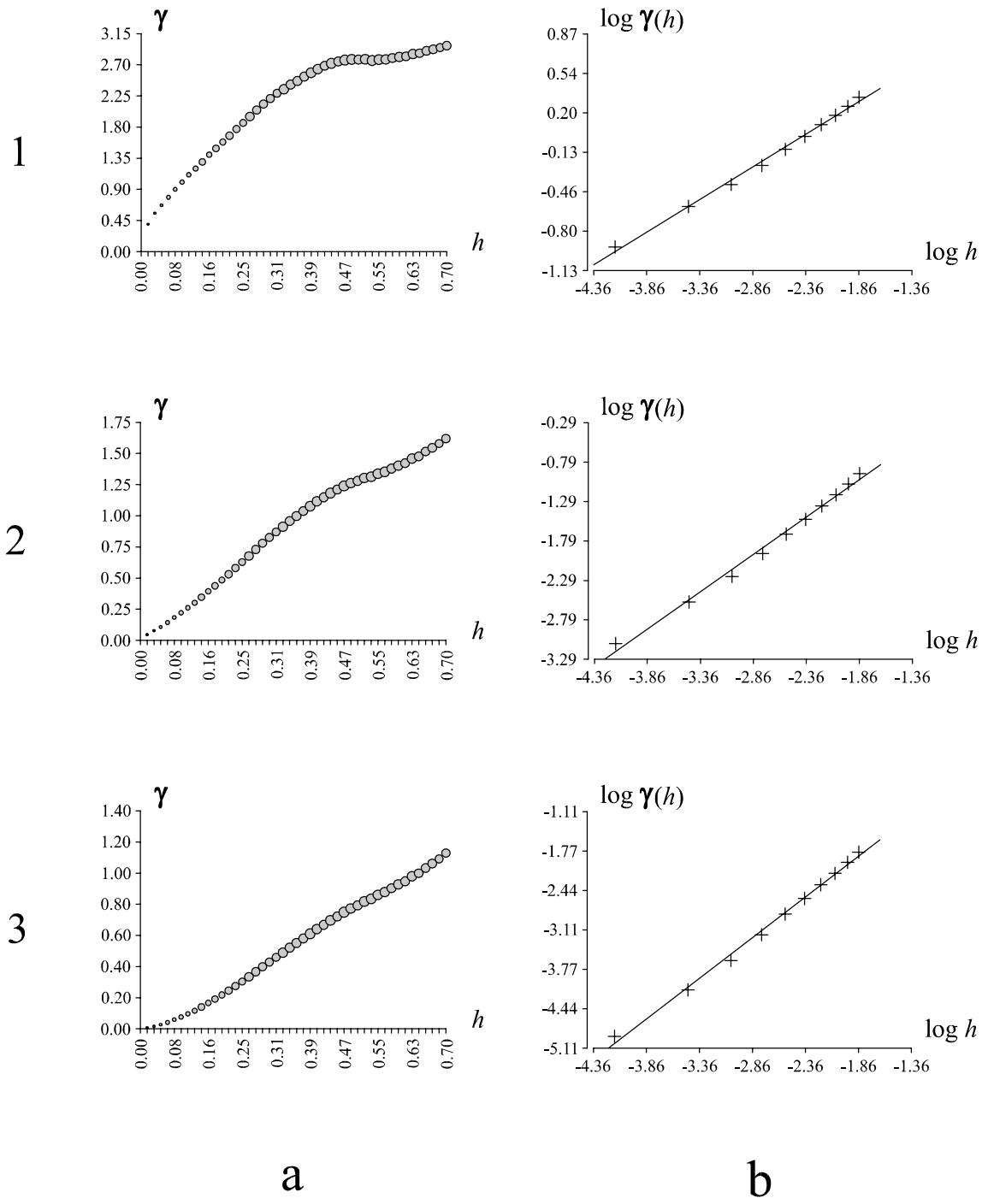
Tableau 11.4: Calcul de la dimension fractale de trois réalisations de fBm bidimensionnels définis par $H = 0.2$, $H = 0.5$, et $H = 0.8$. m : pente de la droite de régression du graphe log-log établi à partir des 10 premiers points du variogramme local. H_{obs}, D_{obs} : valeurs observées de l'exposant de Hurst et de la dimension fractale correspondante.

Les mêmes calculs sont effectués à partir des cinq premières classes des variogrammes locaux des six images considérées dans la Section 11.2.4. Le classement donné par D_{obs} (Tab. 11.5) se révèle parfaitement conforme à celui de la complexité topologique des cartes isoplèthes associées aux images (Tab. 11.3, Fig. 11.6 & 11.7). Cet accord constitue une forme de validation mutuelle des deux approches de mesure de la complexité spatiale.

Modèle	m	H_{obs}	D_{obs}
Périodique	1.655	0.828	2.172
Gaussien	1.571	0.786	2.214
Cubique	1.322	0.661	2.340
Sphérique	0.824	0.412	2.588
Pentasphérique	0.776	0.388	2.612
Exponentiel	0.599	0.300	2.700

Tableau 11.5: Calcul de la dimension fractale de réalisations de FAST-2 bidimensionnelles pour six modèles de variogrammes. m : pente de la droite de régression du graphe log-log établi à partir des 5 premiers points du variogramme local. H_{obs}, D_{obs} : valeurs observées de l'exposant de Hurst et de la dimension fractale correspondante.

[15] Dans une étude de Monte-Carlo portant sur la précision de l'estimation de la dimension fractale thérorique de fBm unidimensionnels, Wen & Sinding-Larsen (1997) recommandent de ne pas utiliser plus de dix classes de distances à partir de l'origine.



Figure˜11.10: Mesure de la complexité spatiale de trois variables régionalisées simulées sur une grille 65×65 pour trois valeurs de l'exposant de Hurst. (1) $H = 0.2$. (2) $H = 0.5$. (3) $H = 0.8$. (a) Variogramme local. (b) Graphe log-log établi à partir des 10 premiers points du variogramme.

Chapitre 12

Conclusion

> *"... studies point to increased interest in geostatistical methods in ecology, but they leave important considerations, explanations, and caveats unadressed. One missing feature is a description of the underlying theory and assumptions of geostatistics. These fondations must be appreciated in order that the transfer of the techniques be accomplished effectively and correctly."* (Rossi *et al.* 1992)
>
> *"Il n'y a plus pour le chercheur* [...] *ni privilège d'une discipline, ni hiérarchie entre disciplines, mais plutôt organisation de ces disciplines, de leurs techniques et de leurs pratiques, en vue d'atteindre un objectif* [...] *Finalement, ce qui est devenu un facteur décisif dans bien des domaines, c'est l'indisciplinarité"* (Legay 1986a)

Appréhender de façon globale le traitement des variables régionalisées (VR) en écologie constitue un défi qu'il est impossible de relever seul. Même en nous limitant aux apports de la géomatique et de la géostatistique, il nous a été impossible de traiter le sujet de façon exhaustive. Avant d'envisager de poursuivre le travail entrepris dans cette thèse, il nous reste à :

- mettre en évidence les principaux thèmes abordés, ainsi que ceux que nous avons nécessairement dû négliger,
- porter un regard général sur le statut de la géostatistique, les apports mutuels de la géomatique et de la géostatistique, le rôle de l'informatique et le type de biométrie illustré par ce mémoire.

12.1 Bilan et perspectives

Les quatre sujets que nous jugeons prioritaires en ce qui concerne le traitement des VR en écologie statistique sont :

- l'échantillonnage,
- le traitement univarié,
- le traitement multivarié,
- l'inférence des processus.

12.1.1 Echantillonnage

L'échantillonnage tel qu'il est considéré par la géostatistique met essentiellement l'accent sur les aspects physiques de l'échantillonnage et fait appel à une théorie reposant sur une hiérarchie d'erreurs (Cox *et al.* 1997). La théorie utilisée est celle de l'échantillonnage des matériaux particulaires proposée par Pierre Gy (*cf.* David 1977, Deverly 1984a, 1984b, Cressie 1991, p. 323, Myers 1997, pp. 183-240). De prime abord, ce type de préoccupation est assez déconnecté des questions posées par l'échantillonnage en écologie. En conséquence, les apports de la géostatistique dans ce domaine ne sont pas tous directement identifiables dans la littérature géostatistique consacrée à l'échantillonnage (*e.g.*, Deverly 1984a, 1984b, Myers 1997).

12.1.1.1 Calcul de la précision de la moyenne

Le calcul de la variance d'estimation globale σ_E^2 pour différents motifs d'échantillonnage peut apparaître comme le premier grand apport de la géostatistique (Petitgas 1991). Toutefois, en biologie, l'utilisation de σ_E^2 s'est répandue uniquement au sein de la communauté des halieutes, sous l'impulsion de l'ICES (*International Council for the Exploration of the Sea*), et la littérature concernée se trouve essentiellement concentrée au sein de l'*ICES Journal of Marine Science*. L'application de la géostatistique à l'estimation des ressources halieutiques à partir de campagnes acoustiques a fait l'objet de deux thèses au centre de géostatistique de l'ENSMP, à Fontainebleau (Petitgas 1991, Guilbin 1997). Cependant, le défaut commun à ces applications est qu'elles ne replacent pas σ_E^2 dans un contexte statistique général. En effet, tout se passe comme si l'interprétation de σ_E^2 allait de soi.

Jusque récemment, en écologie on considérait que l'estimation de la variance par la formule classique s^2/n ne pouvait pas être utilisée dans le cas de données autocorrélées (*e.g.*, Simard *et al.* 1992, Dutilleul *et al.* 1993). De Gruijter & ter Braak (1990) ont insisté sur le fait que les résultats de la théorie de l'échantillonnage aléatoire étaient valides même en présence d'autocorrélation spatiale, et par conséquent, la formule classique s^2/n peut être utilisée dans le cas de l'EAS d'une population spatialement structurée (Legendre 1993). Une source majeur de confusion provient de l'imprécision quant au cadre inférentiel adopté pour calculer la précision d'une estimation spatiale. En confondant, d'une part, la différence entre l'indépendance spatiale des données et l'indépendance stochastique des variables aléatoires, et d'autre part, le paradigme statistique sous-jacent (randomisation du dispositif, modélisation par une variable aléatoire unique ou par une fonction aléatoire), il est impossible de discuter sérieusement du problème statistique de la précision des estimations spatiales (Chapitre 5). Ainsi, Petitgas (1993) indique bien que deux méthodes sont disponibles pour calculer la variance de la moyenne dans un contexte d'autocorrélation spatiale — *i.e.* la théorie de l'échantillonnage aléatoire et la géostatistique — mais ne fournit aucun éclaircissement quant à l'interprétation des différentes variances considérées. Nous avons essayé d'éclaircir ces questions dans les Chapitres 4, 5 & 6. En particulier, dans Aubry & Debouzie (1999a), nous avons contribué au transfert effectif de la variance σ_E^2 en écologie en discutant de son interprétation par rapport à la variance d'échantillonnage dans le cas de l'échantillonnage systématique. Puis, dans Aubry & Debouzie (1999b), nous avons étendu cette discussion à d'autres types d'échantillonnage et à une variance géostatistique conditionnée par les valeurs (Chapitres 4 & 6).

12.1.1.2 Calcul de l'efficacité d'un dispositif

La géostatistique propose également des formules permettant de comparer l'efficacité des principaux dispositifs d'échantillonnage. Toutefois, les formules nécessaires sont déjà présentes dans Zubrzycki (1958), autrement dit, avant la naissance officielle de la géostatistique de Matheron, en 1962 (Chapitre 5).

Dans le cas d'un dispositif pour lequel aucune formule n'est disponible, le recours à une méthode de Monte-Carlo permet néanmoins de calculer son efficacité (Chapitre 5). D'une façon plus générale, n'importe quelle stratégie d'échantillonnage peut être évaluée à l'aide d'une méthode de Monte-Carlo mettant en oeuvre la simulation de fonctions aléatoires (FA) (Chapitre 5).

12.1.1.3 Optimisation de l'échantillonnage

Au-delà du calcul de l'efficacité des dispositifs d'échantillonnage, le recours aux FA permet d'optimiser des motifs d'échantillonnage en analysant l'effet du nombre et de la répartition spatiale des supports dans le domaine d'étude — voire même l'effet de leur surface ou de leur géométrie — en fonction de la structure d'autocorrélation spatiale de la population (Chapitre 8).

Switzer (1979) affirme que l'optimisation du motif d'échantillonnage constitue habituellement un exercice futile alors que cette démarche est considérée comme particulièrement utile dans le cas des échantillonnages réalisés dans une discipline telle que la foresterie (Houllier 1992). En fait, l'optimisation de l'échantillonnage ne présente d'intérêt pratique qu'à partir du moment où la signification des critères optimisés est bien comprise ; c'est pourquoi, après la description de chaque procédure d'optimisation, nous avons essayé de redonner **du sens** aux résultats obtenus (Chapitre 8).

A taille d'échantillon fixée, le motif d'échantillonnage se révèle très important, que ce soit pour l'estimation globale, la cartographie, ou l'estimation du variogramme. Dans toutes ces problématiques, nos résultats confirment l'intérêt de l'échantillonnage systématique sur une grille, du moins lorsque la densité d'échantillonnage est en adéquation avec le degré de variabilité spatiale de la VR et que celle-ci est relativement homogène (Chapitre 8).

En ce qui concerne les multivariables régionalisées (MVR), l'optimisation de l'échantillonnage spatial n'en est encore qu'à ses balbutiements et fait généralement appel à des techniques de partitionnement spatial en plus de la géostatistique (*e.g.*, McBratney & Webster 1983b, Legendre *et al.* 1989, Odeh *et al.* 1990). Il est évident que l'optimisation de l'échantillonnage des MVR pose des difficultés techniques supérieures à celles rencontrées dans le cas des VR, mais les difficultés théoriques sont également plus importantes. En effet, il nous semble difficilement envisageable d'optimiser l'échantillonnage de plusieurs VR simultanément, *i.e.* à partir d'un seul et même motif d'échantillonnage, à moins que ces VR présentent une forte association spatiale. A l'issue d'une hypothétique procédure d'optimisation de l'échantillonnage d'une MVR, le résultat risque fort d'être un compromis conduisant à une estimation médiocre pour la totalité des VR.

12.1.1.4 Changement de support

Nous n'avons pas étudié de façon détaillée la relation entre la variabilité de la VR et la taille des supports, *i.e.* la problématique dite du *changement de support* (Isaaks & Srivastava 1989, pp. 458-488, Zhang *et al.* 1990, Cressie 1991, p. 66, pp. 284-289, Bellehumeur *et al.* 1997, Bellehumeur & Legendre 1997). Ce sujet a une longue histoire puisqu'il remonte au moins aux articles de Mercer & Hall (1911) et de Fairfeld-Smith (1938) (*op. cit.* Cressie 1991, p. 284).

12.1.2 Traitement univarié

Dans le traitement univarié des VR en écologie, il est possible d'identifier quatre étapes :

1. Calcul d'indices et ajustement de distributions discrètes. Cette approche considère la distribution statistique des valeurs, mais pas la localisation ni la géométrie des supports : il ne s'agit donc pas à proprement parler du traitement d'une VR.
2. Calcul et test statistique d'indices et de fonctions d'autocorrélation spatiale. L'analyse de l'autocorrélation spatiale permet une analyse assez fine de la structure spatiale d'une VR, mais elle pose des problèmes d'interprétation en présence de tendances, et ne fournit pas une représentation directe de la variabilité spatiale de la VR. L'analyse de l'autocorrélation spatiale est étroitement liée à la mesure de la complexité spatiale.
3. Cartographie de la VR par interpolation. La cartographie permet d'obtenir une représentation approximative de la structure spatiale de la VR, mais les méthodes d'interpolation présentent l'inconvénient de sous-estimer la variabilité spatiale réelle, ce qui peut conduire à des erreurs d'interprétation.
4. Simulation stochastique de la structure spatiale d'une VR. La simulation permet d'approximer la variabilité spatiale de la VR, en respectant sa structure d'autocorrélation spatiale, sa distribution statistique et, éventuellement, d'autres caractéristiques.

Examinons quels sont les apports réels de la géomatique et de la géostatistique en ce qui concerne les étapes 2 à 4.

12.1.2.1 Mesure de l'autocorrélation et de la complexité

On s'accorde à reconnaître que l'analyse de l'autocorrélation spatiale a été introduite en écologie par Jumars *et al.* (1977) et indépendamment en biologie évolutive par Sokal & Oden (1978a, 1978b). Les statistiques introduites par Jumars *et al.* (1977) et Sokal & Oden (1978a, 1978b) sont issues de la géographie (Cliff & Ord 1973), domaine largement dominé par les tests d'autocorrélation spatiale (Ripley 1988b), et actuellement, ce sont encore les géographes qui produisent le plus de mesures et de tests d'autocorrélation spatiale (Anselin 1988, Getis 1989, Hubert & Arabie 1991, Deichman & Anselin 1994, Ord & Getis 1995, Anselin 1995, Tiefelsdorf & Boots 1997, Simon 1997). En introduisant l'utilisation des outils d'analyse de l'autocorrélation spatiale en écologie, l'objectif clairement affiché par Jumars *et al.* (1977) était d'exploiter de façon efficace une information négligée par

les méthodes traditionnelles (indices et ajustement de lois discrètes), *i.e.* la localisation spatiale des données.

Dans le Chapitre 3, nous avons abordé l'analyse de la dépendance spatiale, à l'aide de mesures d'autocorrélation proprement dite, ou de mesures de variance ou de covariance, les trois types de mesures étant étroitement liés. Dans cette perspective, le variogramme défini par la géostatistique ne constitue pas réellement une nouveauté dans la mesure où il correspond, à un facteur de standardisation près, au c de Geary. En outre, le variogramme est étroitement lié aux méthodes proposées en écologie végétale depuis l'article fondateur de Greig-Smith (1952). En effet, la représentation graphique de la variabilité en fonction du niveau d'agrégation de quadrats ou de la distance entre quadrats, qu'il s'agisse de la TTLV (*Two-Term Local Variance*) de M.O. Hill (1973), ou de la PQV (*Paired-Quadrat Variance*) de Ludwig & Goodall (1978), présente de profondes analogies avec le variogramme (Ver Hoef *et al.* 1993, Brockman & Murray 1997).

En fait, le principal apport de la géostatistique contemporaine a été de proposer des fonctions dites *non ergodiques* (Isaaks & Srivastava 1988, 1989), qui ont été rapidement adoptées en écologie des populations (Rossi *et al.* 1992, Liebhold *et al.* 1993). La géostatistique définit encore d'autres fonctions telles que le *rodogramme*, ou le *madogramme* (Journel 1988, Deutsch & Journel 1992, p. 42, Walker *et al.* 1997) et différents types de variogrammes relatifs (Isaaks & Srivastava 1989, pp. 163-170, Deutsch & Journel 1992, pp. 41-42, Myers 1997, pp. 297-300) dont l'intérêt reste à évaluer en écologie.

On constate donc que la géostatistique, tout comme la géographie, contribue à augmenter le nombre des mesures de la dépendance spatiale. Quant à elle, la géomatique contribue à diversifier les pratiques en proposant de calculer divers graphes de voisinage et diverses distances. Actuellement, les chercheurs se trouvent donc confrontés à une surabondance d'outils et de pratiques[1], parmi lesquels il convient de faire un choix : le Chapitre 3 a été entièrement consacré à cette problématique.

La mesure de la complexité spatiale s'effectue en pratique essentiellement en termes d'autocorrélation spatiale et de dimension fractale (revue dans Turner *et al.* 1991, Cullinan & Thomas 1992, Cooper *et al.* 1997). Le Chapitre 11 a permis, à la fois d'éclaircir le statut de l'analyse fractale en écologie, le lien entre la dimension fractale stochastique et le variogramme, puis le lien entre la complexité spatiale, la dépendance spatiale et la régularité spatiale. En mettant à part le calcul de la dimension fractale pour résumer le comportement à l'origine du variogramme, il nous est apparu que l'analyse fractale offrait peu d'intérêt en pratique. En dehors de l'interprétation des variogrammes emboîtés, la géostatistique était jusqu'à présent assez peu concernée par la problématique de la complexité spatiale, au contraire de l'écologie statistique. En conséquence, nous avons proposé :

- de calculer la complexité géométrique d'une carte choroplèthe dans le cadre de la géométrie probabiliste plutôt que dans celui de la géométrie fractale, en faisant appel à des opérateurs fournis par la géomatique,
- des définitions opératoires de la complexité topologique d'une carte isoplèthe, faisant intervenir des modèles et des opérateurs géomatiques,
- des définitions opératoires de la complexité d'une image, faisant intervenir son *quadtree*.

[1] Il est évidemment totalement inepte d'affirmer que les mesures et études de l'autocorrélation spatiale sont peu développées en écologie (*e.g.*, Koenig & Knups 1998).

Seules nos propositions concernant la complexité topologique d'une carte isoplèthe ont été complètement évaluées. Nos autres propositions méritent d'être explorées plus avant parce que la mesure de la complexité spatiale est importante, non seulement afin d'apprécier la relation qu'elle entretient avec la diversité spécifique (Baudry & Baudry-Burel 1982, Scheiner 1992, Chiarello 1994, He *et al.* 1994), mais également avec la stabilité des écosystèmes (Legendre *et al.* 1989, Keitt 1997).

12.1.2.2 Cartographie

La géostatistique de Matheron est déjà pressentie par Sokal & Oden (1978a) comme une approche prometteuse pour l'analyse des structures spatiales en écologie. En ce qui concerne l'analyse structurale de la répartition des espèces, Matheron lui-même considérait qu'il y aurait un *"immense domaine à découvrir, avec la théorie des fonctions aléatoires en particulier"*, et *"une moisson de résultats nouveaux et inattendus à faire sur la répartition des espèces"* (Matheron, comm. pers. *in* Cartan 1978).

De fait, le krigeage est aujourd'hui largement utilisé comme procédure de cartographie de VR biotiques ou abiotiques (Chapitre 6, Annexe G). Toutefois, si la cartographie par krigeage présente l'avantage essentiel de tenir compte de la structure d'autocorrélation spatiale lors de l'interpolation, elle présente également l'inconvénient de donner une image faussée de la variabilité spatiale réelle de la VR cartographiée. La cartographie par krigeage doit donc être considérée comme un bon moyen de reproduire l'allure générale de la structure spatiale d'une VR, mais une reproduction plus fidèle de la variabilité spatiale nécessite de recourir à la simulation stochastique.

12.1.2.3 Simulation stochastique

Depuis quelques années, l'attention en géostatistique s'est déplacée du krigeage vers la simulation stochastique, et en particulier la simulation conditionnelle (Deutsch & Journel 1992).

Bien que la simulation conditionnelle ait déjà été mentionnée dans la littérature écologique, notamment par Liebhold *et al.* (1993) et surtout par Journel (1994b), les exemples d'applications en écologie restent excessivement rares (*e.g.*, Rossi *et al.* 1993, Mowrer 1997, Aubry & Debouzie 1999b). Nelson *et al.* (1999) considèrent qu'il est probable que peu de chercheurs en écologie utiliseront la simulation conditionnelle dans un futur proche, à cause des difficultés techniques qu'elle soulève. Or c'est justement cette voie qui nous semble la plus prometteuse parce qu'elle permet de proposer des cartes stochastiques (Chapitres 4 & 6).

Les développements actuels en matière de simulation conditionnelle tendent à reformuler le problème en termes d'optimisation combinatoire, et utilisent la méthode du recuit simulé (*cf.* Deutsch & Journel 1992, pp. 154-160, Deutsch & Cockerham 1994, Goovaerts 1996, 1997, pp. 409-420, Fang & Wang 1997, Groleau & Marcotte 1997, Goovaerts 1998b), voire même les algorithmes génétiques (Bergeret & Besse 1997) : ces heuristiques ont été exposées dans le cadre de l'optimisation de l'échantillonnage (Chapitre 8).

12.1.3 Traitement multivarié

De façon analogue au traitement univarié, dans le traitement multivarié des VR en écologie il est possible d'identifier quatres étapes :

1. Description, mesure et test de la structure de corrélation entre plusieurs VR, au moyen de méthodes classiques, a-spatiales.
2. Description, mesure et test de l'association spatiale entre plusieurs VR.
3. Cartographie d'une VR en tirant partie de la connaissance d'autres VR qui lui sont associées.
4. Simulation stochastique d'une MVR.

L'ensemble de ces étapes ne pouvait pas être traité de façon satisfaisante dans le cadre de ce mémoire compte tenu de la richesse des problématiques sous-jacentes. Nous nous contentons d'indiquer quelques pistes et de rappeler les éléments de réponse apportés dans les Chapitres 9 & 10, sans aborder la question de la simulation stochastique des MVR (*cf.* Deutsch & Journel 1992, pp. 121-123, Verly 1993, Almeida & Journel 1994, Goovaerts 1997, pp. 390-393, 400-403) .

12.1.3.1 Description, mesure et test de la structure de corrélation

A notre connaissance, la comparaison de plusieurs méthodes destinées à traiter le problème de l'autocorrélation dans le test de la corrélation entre deux VR localisées dans un espace bidimensionnel n'a jamais été publiée[2]. Les résultats obtenus en simulant des FA montrent que, parmi les méthodes mentionnées, les seules méthodes que l'on puisse recommander sont les méthodes paramétriques incorporant explicitement la structure d'autocorrélation spatiale des VR, *i.e.* le test de Monte-Carlo spatial — faisant intervenir la simulation non conditionnelle d'une FA — et le test paramétrique modifié (Chapitre 9). Cependant, ces méthodes ne peuvent pas être appliquées de façon universelle parce qu'elles nécessitent de pouvoir estimer de façon fiable le variogramme d'au moins une des deux VR mises en jeu.

Switzer (1983) indique que les défauts potentiels des procédures d'analyse multivariée qui n'exploitent pas les caractéristiques spatiales des données sont parfois apparents lorsque les sorties sont cartographiées. Par exemple, cet auteur mentionne le fait que les cartes isoplèthes des scores de l'ACP ou des valeurs des fonctions discriminantes peuvent montrer des composantes de haute fréquence fallacieuses. L'auteur mentionne également les fortes structures spatiales des résidus issus d'une régression multiple.

L'impact de l'autocorrélation spatiale sur une méthode d'analyse multivariée classique telle que l'ACP ne semble pas avoir été évalué, bien que la question soit posée par Haining (1991). A notre connaissance, la comparaison de plusieurs méthodes destinées à traiter le problème de l'autocorrélation spatiale dans la régression simple ou multiple entre VR localisées dans un espace bidimensionnel n'a pas non plus été publiée[3]. Il est crucial de réaliser cette étude compte tenu de l'importance des techniques de régression en écologie.

[2] En ce qui concerne les séries temporelles halieutiques, voir Pyper & Peterman (1998a, 1998b).

[3] En ce qui concerne les séries temporelles de conductance stomatique, voir Strachan & Harvey (1996).

12.1.3.2 Description, mesure et test de l'association spatiale

Le concept de corégionalisation et les fonctions structurales croisées proposées par la géostatistique sont introduits dans le Chapitre 10. Cependant, nous n'avons pas mentionné l'existence d'un analogue multivarié du variogramme (Mackas 1984, Bourgault & Marcotte 1991, Bourgault 1992). Nous n'avons pas non plus évalué l'intérêt pratique de la géostatistique multivariable qui est pourtant actuellement bien développée (*cf.* Matheron 1982, Galli & Wackernagel 1987, Goulard 1988, Wackernagel & Butenuth 1989, Goovaerts 1992, 1993, 1994a, Goovaerts & Sonet 1993, Wackernagel 1993, 1998), et qui entre en concurrence avec des méthodes élaborées en dehors de la géostatistique (*e.g.*, Wartenberg 1985a, 1985b, Thioulouse *et al.* 1995). En ce qui concerne le domaine du partitionnement spatial, l'application de la géostatistique en est restée au stade des balbutiements les plus élémentaires (*e.g.*, Oliver & Webster 1989, Bourgault *et al.* 1992, Charmet *et al.* 1994).

Nous nous sommes surtout intéressés à la mesure de l'association spatiale entre des cartes binaires et entre des cartes quantitatives, modélisées sous la forme d'images. Dans ce contexte, nous avons fait appel à la structure de *quadtree* largement employée en géomatique, en utilisant une coopération entre l'approche structurelle et l'approche statistique, démarche qui s'est avérée très satisfaisante (Chapitre 10).

12.1.3.3 Cartographie exploitant de l'information auxilaire

La cartographie d'une VR peut avantageusement tirer partie de la connaissance d'une ou plusieurs VR qui lui sont associées. La géostatistique propose ainsi une extension du krigeage prenant en compte la covariation spatiale entre les VR, nommée *cokrigeage* (*cf.* Journel & Huijbregts 1978, pp. 324-343, Myers 1982, 1983, Issaks & Srivastava 1989, pp. 400-416, Carr & Myers 1990, Rosenbaum & Söderström 1996, Goovaerts 1997, pp. 203-258, Long & Myers 1997, Goovaerts 1998a, Zhang *et al.* 1999). Nous n'avons pas évalué l'intérêt pratique du cokrigeage en écologie, pas plus que celui des méthodes concurrentes telles que le krigeage dans des strates, le krigeage simple avec moyenne locale variable, le krigeage avec dérive externe, le krigeage combiné avec la régression ou avec une analyse factorielle en mode Q (Ver Hoef & Cressie 1993, Simard *et al.* 1993, Hudson 1993, Hudson & Wackernagel 1994, Asli & Marcotte 1995, Knotters *et al.* 1995, Odeh *et al.* 1995, Bourennane *et al.* 1996, Goovaerts 1997, pp. 185-202, Juang & Lee 1998b, Goovaerts 1999).

12.1.4 Inférence des processus

Il est important de distinguer deux types distincts d'inférence basée sur les échantillons : celle qui consiste à décrire les caractéristiques d'une population finie et celle qui concerne les causes qui ont produit ces caractéristiques (Hansen 1987). Ce second type d'inférence est particulièrement important mais également beaucoup plus difficile que le premier (Hansen 1987).

Dans ce mémoire, nous n'avons traité que le premier type d'inférence (Chapitres 5, 6, 7 & 8) en évacuant dès l'introduction le problème de l'inférence des processus à partir des structures spatiales. Ce type de problématique part du constat que toute structure spatiale est l'aboutissement d'une histoire, le résultat d'un processus (Matheron 1970).

Relier les structures spatiales aux processus est un défi fondamental en écologie et en biologie évolutive. Il s'agit d'une tâche particulièrement difficile parce qu'il est possible que différents mécanismes produisent un même type de structure spatiale.

Certains chercheurs considèrent que l'un des principaux intérêts de la géostatistique en écologie est précisément de permettre des inférences quant aux processus générateurs des structures spatiales observées (*e.g.*, Simard *et al.* 1992). En biologie évolutive, ce type d'avis est partagé par Sokal et ses collaborateurs qui affirment que l'analyse de l'autocorrélation spatiale permet d'inférer les processus de microévolution dans les populations naturelles (*cf.* Sokal & Oden 1978b, Sokal 1979, Sokal 1986, Sokal & Jacquez 1991, Sokal & Oden 1991, Sokal *et al.* 1997). Cependant, Slatkin & Arter (1991) contestent la possibilité d'inférer les processus microévolutifs à partir d'une analyse de l'autocorrélation spatiale. La question fondamentale est de savoir s'il est effectivement possible d'obtenir de l'information sur le processus qui a produit la structure spatiale observée à partir d'une analyse de l'autocorrélation spatiale ou de la cartographie du phénomène (Cliff & Ord 1981). Legendre *et al.* (1997) notent à juste titre que les structures spatiales ne contiennent aucune interprétation en elles-mêmes, car c'est bien d'interprétation dont il s'agit.

En fait, autant dans l'inférence des caractéristiques d'une population nous pouvons choisir entre l'inférence *design-based* et l'inférence *model-based*, autant dans l'inférence des causes qui déterminent ces caractéristiques il n'y a pas d'alternative à l'utilisation de modèles (Hansen *et al.* 1983, Hansen 1987). Dans ce contexte, l'inférence des processus à partir des structures spatiales ne peut se faire que de façon indirecte, en comparant les structures spatiales issues de modèles explicatifs à la structure spatiale réelle (*e.g.*, Schotzko & Knudsen 1992, Chiarello 1994, Heil & van Deursen 1996). La mesure de l'accord entre les deux jeux de structures doit permettre d'effectuer un classement entre plusieurs modèles concurrents afin de retenir un ou plusieurs modèles plausibles. Dans cette démarche, le concept de carte stochastique (Chapitre 6) et les mesures d'association spatiale (Chapitre 10) ont un rôle important à jouer.

12.2 Ecologie statistique et variables régionalisées

Les VR écologiques peuvent présenter toutes les formes de distribution statistique, depuis la loi exponentielle négative, jusqu'à la loi normale, en passant par la loi lognormale. En pratique, les distributions présentent souvent une asymétrie positive, parfois très forte (*e.g.*, le diamètre des arbres). Les problèmes posés par les distributions asymétriques en géostatistique ne diffèrent pas de ceux rencontrés de façon générale en statistique, et l'on sait que la géostatistique linéaire opère avec le maximum d'efficacité dans un contexte gaussien (Chauvet 1985).

Une solution relativement satisfaisante consiste à transformer les données par anamorphose de façon à se rapprocher d'une distribution gaussienne, puis à appliquer la transformation réciproque aux résultats de la procédure géostatistique (Chauvet 1985) (Annexe E).

Une autre solution consiste à "disséquer" la VR et à manipuler plusieurs indicatrices (*e.g.*, Murray 1996). La prise en compte de la forme de la distribution statistique de la VR nécessite donc de recourir à la géostatistique non linéaire, même lorsqu'il s'agit de résoudre des problèmes purement linéaires (Chauvet 1985).

Une deuxième caractéristique des VR écologiques — surtout des VR biotiques — est de présenter une régularité spatiale relativement faible[4]. Cette faible régularité spatiale peut être due à une échelle d'observation inadéquate (*e.g.*, maille d'une grille d'échantillonnage trop large) et/ou au comportement agrégatif des organismes et à leur faible dispersion spatiale (*e.g.*, les pucerons). L'effet de pépite est assez fréquent et la pépite relative d'un variogramme expérimental peut être très élevée, ce qui limite alors considérablement l'intérêt des outils géostatistiques.

Enfin, lorsque la VR est définie sur un domaine du plan non connexe (*e.g.*, les positions d'arbres, un habitat fragmenté), les procédures géostatistiques doivent être amendées par des procédures géomatiques afin d'opérer la restriction spatiale correspondante, qu'il s'agisse de kriger ou de simuler.

Dans ce qui suit, nous abordons le statut de la géostatistique, son intégration avec la géomatique et la statistique, le rôle de l'informatique, et enfin le type de biométrie illustré par ce mémoire.

12.2.1 Géostatistique et statistique

Les géostatisticiens ont souvent tendance à insister sur le fait que la géostatistique est originale et diffère profondément de la statistique classique. En effet, Matheron (1963) présente la géostatistique comme une nouvelle science et ne fait référence à aucun auteur, et trente ans plus tard, Chauvet (1994) affirme que *"les méthodes géostatistiques se développent* [...] *dans des directions qui leur sont propres, et qui n'ont plus que de lointains rapports avec la statistique classique"*.

D'abord, il convient de préciser que la géostatistique n'apparaît pas *ex nihilo*. David (1977) fait remonter l'utilisation du variogramme à Langsaetter en 1926, dans le domaine de la foresterie. Cressie (1988b, 1989) précise que le variogramme apparaît dans la littérature à plusieurs reprises, sous la plume de Kolmogorov en physique, en 1941, de Jowett dans le domaine des séries temporelles, en 1952, de Yaglom dans le domaine des probabilités, en 1957, et de Gandin en météorologie, en 1963. De même, le krigeage a souvent été redécouvert, dans différents domaines (Cressie 1990). L'effet de pépite dû à des erreurs d'observation était déjà décrit par Matérn (1960, pp. 55-59), ainsi que la régularisation (Matérn 1960, pp. 59-62), et même une procédure analogue au cokrigeage (Matérn 1960, pp. 108-109). Nous nous sommes également aperçus que la notion de variable régionalisée, de dépendance spatiale exprimée sous une forme analogue au variogramme, ainsi que la problématique de la variance d'extension étaient déjà présentes dans Osborne (1942). En dehors de la théorie des FAI-k (Matheron 1973), il faut reconnaître que la géostatistique apporte peu de choses nouvelles du point de vue des mathématiques théoriques, le concept de FA ayant été formalisé dans les années 1930 par Lévy, Kolmogorov et Khintchine, et les outils théoriques utilisés en géostatistique linéaire étant déjà en place dans les années 1940 suite aux travaux de Cramér, Wiener et Bochner (Chauvet 1994). Mais, bien que les FA soient classiquement utilisées dans le domaine des séries temporelles, les FA multidimensionnelles étaient restées virtuellement inconnues des praticiens (Journel 1986a). L'originalité de la géostatistique réside avant tout dans le rapprochement effectué entre des

[4] Assez curieusement, David (1977, p. 91) affirme que les données biologiques ou écologiques considérées par les biométriciens sont plus continues que celles rencontrées dans le domaine minier.

problèmes techniques très concrets et un arsenal de méthodes mathématiques (Chauvet 1994).

Ensuite, et c'est sans doute le plus important, il faut reconnaître que la géostatistique fait partie de la statistique, même si le vocabulaire qu'elle utilise et la forme des procédures qu'elle propose peuvent laisser penser le contraire. Le recours aux FA consiste en fait à choisir un certain modèle de superpopulation. Or, les modèles de superpopulations sont utilisés en statistique au moins depuis Cochran (1946), afin de comparer des dispositifs d'échantillonnage ou des estimateurs entre eux. L'utilisation des modèles de superpopulations pour développer des estimateurs optimaux est plus récente (*cf.* Smith 1976), mais est également bien connue de la statistique. Le développement de l'estimateur du krigeage sous un certain type de FA s'inscrit donc parfaitement dans la pratique statistique. En outre, le krigeage n'est pas autre chose qu'une régression linéaire multiple selon les moindres carrés généralisés (Journel 1986a, Olea 1992, Chauvet 1994, p. 67, Goovaerts 1999), et en tant que tel, il peut être développé dans le cadre de la théorie de la régression (Corsten 1985, 1989, Stein & Corsten 1991). L'*universalité* définie en géostatistique équivaut strictement à l'absence de ξ-*biais* en statistique. De même, la notion d'*incrément* de la géostatistique est analogue à celle de *contraste* en statistique. Enfin, le modèle *universel* de la géostatistique est analogue au *modèle linéaire généralisé* de la statistique. Brus & de Gruijter (1997) vont même jusqu'à écrire, fort justement, que dans l'utilisation qu'elle fait des modèles stochastiques paramétriques (les FA), la géostatistique est plus proche de la statistique classique que ne l'est la théorie de l'échantillonnage probabiliste.

En conséquence, comme nous avons essayé de le montrer tout au long de ce mémoire, il n'est plus question aujourd'hui de considérer que la géostatistique est séparée du tronc commun de la statistique appliquée. Il reste que ce "séparatisme" originel n'est pas sans avoir eu des conséquences fâcheuses sur l'intelligibilité de la géostatistique.

12.2.2 Intelligibilité de la géostatistique

Actuellement, il existe une grande confusion en ce qui concerne la signification même du terme *géostatistique*. Par exemple, Nicholson & Mather (1996) incluent l'analyse des semis de points (*point pattern analysis*) dans les techniques géostatistiques, et Chou & Soret (1996) mentionnent le test de Mantel partiel en tant que méthode géostatistique. Il existe fréquemment une confusion entre les statistiques spatiales (au sens anglo-saxon du terme) et la géostatistique (au sens français du terme) (*e.g.*, Cardina *et al.* 1995), et entre la géostatistique et l'analyse de l'autocorrélation spatiale (*e.g.*, Kitron *et al.* 1996). Parmi les non spécialistes, le sens de *géostatistique* tend à glisser vers celui de *statistique spatiale* au sens large.

La géostatistique (au sens strict) est du reste assez mal comprise. Le non spécialiste peut tout d'abord être assez désappointé de constater que, hormis le test de Kolmogorov-Smirnov, la géostatistique n'utilise aucun test (Olea 1992), et ne fait pas non plus référence à la notion de degré de liberté (Merks 1992). C'est en réalité parfaitement cohérent avec le fait que la géostatistique au sens strict (*cf.* Journel 1983b) consiste à manipuler des modèles de superpopulations sous la forme de FA, ce qui n'a, à l'origine, rien à voir avec la théorie des tests. Cependant, il nous semble que le défaut d'intelligibilité qui caractérise la géostatistique provient en partie de ce qu'elle n'identifie pas les différentes interprétations du concept de superpopulation. Il est par exemple incompréhensible de chercher à estimer

les paramètres d'une FA lorsque celle-ci est utilisée pour modéliser un phénomène unique à la fois dans le temps et l'espace, et nous avons souvent insisté sur le rôle que nous faisions jouer au modèle de superpopulation dans ce type de situation. En revanche, cette démarche est compréhensible dès lors que le phénomène est susceptible de se répéter dans le temps (*e.g.*, les précipitations, la croissance annuelle de l'herbe dans une prairie paturée, etc.), ou éventuellement dans l'espace.

En négligeant de préciser la signification de la modélisation d'un phénomène par une FA, et en considérant comme superflu de fournir des interprétations claires des différentes variances définies en géostatistique, on aboutit forcément à une profonde incompréhension, mêlée de ressentiment, comme peuvent en témoigner l'article de Philip & Watson (1986), puis la lettre ouverte de Philip & Watson (1987) adressée à Matheron. Les réponses de Srivastava (1986) et de Matheron (1986, 1987) n'ont cependant contribué en rien à faire progresser l'intelligibilité de la géostatistique, à en juger par l'article très agressif (et en vérité assez inepte) de Merks (1992). Avec ce mémoire, nous espérons avoir contribué à rendre la géostatistique davantage intelligible.

12.2.3 Géomatique, géostatistique et statistique

Nelson *et al.* (1999) notent fort justement que l'analyse spatiale n'a pas toujours besoin de faire appel à la géostatistique. En particulier, les Chapitres 10 & 11 ont montré que des méthodes reposant sur des concepts, des objets et des opérateurs géomatiques, étaient complémentaires des outils géostatistiques. Selon Jumars *et al.* (1977), lorsque les échantillons sont coûteux à acquérir et que leur analyse est bon marché, la meilleure approche consiste à utiliser plusieurs méthodes applicables de façon valide au problème à traiter. Nous nous inscrivons résolument dans cette approche qui multiplie les points de vue sur l'objet étudié et enrichit donc considérablement l'analyse.

Les trois champs méthodologiques que nous avons considérés dans ce mémoire, *i.e.* la géomatique, la géostatistique au sens strict et la statistique, peuvent et doivent être utilisés de façon intégrée. Malheureusement, de même que la géostatistique s'est longtemps développée en dehors de la statistique générale, en élaborant les SIG, la géomatique s'est également développée de façon assez indépendante de l'analyse spatiale statistique, ce qui se traduit actuellement par une demande croissante d'intégration des deux (Arbia 1993, Haining *et al.* 1996). Il n'est pas surprenant de constater que dans de nombreuses problématiques, les chercheurs font appel simultanément aux SIG et à la géostatistique (*cf.* Liebhold *et al.* 1993, Höck *et al.* 1993, Nelson *et al.* 1994, Atkinson 1996, Orum *et al.* 1997, McBratney & Odeh 1997, Myers *et al.* 1997, White *et al.* 1998, Nelson *et al.* 1999).

En pratique, les SIG sont le plus souvent utilisés pour stocker les données, les exporter vers un programme statistique, puis les résultats sont importés et visualisés (*e.g.*, Kadmon & Danin 1997). L'interfaçage entre les principaux SIG et logiciels statistiques du marché semble plus fort qu'il y a quelques années (Croft & Kessler 1996), mais encore faut-il que les outils statistiques utilisés soient valides en présence d'autocorrélation spatiale, ce qui n'est pas le cas d'une procédure aussi élémentaire que le test de la corrélation linéaire entre deux variables (Chapitre 9). A notre connaissance, l'intégration des outils statistiques dédiés au traitement des VR concerne essentiellement l'analyse de l'autocorrélation spatiale et l'interpolation par krigeage. La simulation conditionnelle semble encore loin d'être concernée, bien qu'elle soit fortement recommandée (Journel 1996).

Au sein des SIG, l'analyse de l'autocorrélation spatiale est généralement effectuée grâce au I de Moran (Ding & Fotheringham 1992, Shen 1994, Hansen 1994), mais l'utilisation des langages de macro-commandes des SIG — vraisemblablement interprétés et non pas compilés — et de programmes externes rend les traitements assez inefficaces[5].

Les systèmes d'interpolation spatiale construits en reliant des logiciels géostatistiques aux SIG restent d'actualité parce que les quelques procédures géostatistiques présentes dans les SIG commercialisés sont très rudimentaires[6] (Varekamp *et al.* 1996, Gilbert 1997). Toutefois, un tel assemblage hétéroclite est à la fois inefficace et source d'erreurs, notamment à cause de l'incompatibilité des formats de fichiers et des multiples manipulations effectuées dans des environnements logiciels différents.

Si l'intérêt de l'incorporation des outils géostatistiques au sein des SIG est évident, réciproquement, les structures hiérarchiques utilisées par la géomatique peuvent enrichir les logiciels géostatistiques, notamment les *kd-trees* (*cf.* Ooi 1990, pp. 28-32) afin d'augmenter l'efficacité du krigeage en voisinage glissant (Aubry 1996b), et les *quadtrees* pour kriger selon une résolution variable (Mason *et al.* 1994), ou renouveler la problématique de la décomposition hiérarchique de la variabilité spatiale (Csillag & Kabos 1996).

12.2.4 Le rôle de l'informatique

Il est possible d'identifier trois facettes essentielles du rôle actuel de l'informatique en écologie statistique. En premier lieu, l'informatique propose des modèles de structuration des connaissances et de représentation symbolique ou numérique de l'information qui définissent des cadres de réflexion fortement organisés propices à l'intégration des connaissances (Houllier 1992, Chevenet 1994). La notion de *brainware* de Kitagawa (1974), forgée sur le modèle des termes *software* et *hardware*, désigne précisément l'aspect conceptuel de l'informatique. Cet aspect a été illustré essentiellement dans le Chapitre 2 consacré aux modèles géomatiques. Le second apport de l'informatique réside dans l'implémentation des algorithmes sous la forme de logiciels largement diffusables : cet aspect ne fait pas à proprement partie du travail de recherche mais plutôt de sa valorisation. Enfin, le troisième apport concerne la puissance de calcul autorisée par les ordinateurs actuels. Ce dernier aspect est particulièrement important parce que le recours à des techniques coûteuses en temps de calcul, dites *méthodes intensives*, est d'un intérêt considérable pour l'écologie statistique spatiale.

Le coût des méthodes intensives auxquelles nous avons fait référence dans ce mémoire (tests de permutation ou de randomisation, intégration de Monte-Carlo, simulation de Monte-Carlo, optimisation combinatoire), reste encore assez élevé pour certains types ou tailles de problèmes. Par exemple, dans une application écologique, Rossi *et al.* (1993) simulent des réalisations conditionnelles sur une grille comptant 13 020 noeuds en utilisant la méthode de simulation séquentielle, méthode assez coûteuse en temps de calcul. Rossi *et al.* (1993) génèrent seulement $L = 100$ réalisations, ce qui nécessite tout de même 18 heures de calcul sur un PC à base de processeur Intel 80386.

[5] Shen (1994) signale qu'il lui faut 8 heures de calcul pour analyser l'autocorrélation spatiale d'un ensemble de 101 zones !

[6] A cet égard, le cas de ArcInfo — leader mondial dans le domaine des SIG — est exemplaire, le krigeage y étant implémenté pratiquement comme une boîte noire.

Selon que l'on agit au niveau du logiciel ou du matériel, il existe évidemment deux solutions — non exclusives — afin d'augmenter l'efficacité des méthodes intensives :

- concevoir et implémenter des algorithmes plus efficaces, par exemple, dans le cas de la simulation des FA, employer des techniques utilisant la transformée de Fourier rapide (Davis 1987b, Dietrich & Newsan 1993, 1995, Chan & Wood 1997), et dans le cas des tests de Monte-Carlo, procéder à un échantillonnage séquentiel sous H_0 (Besag & Clifford 1991),
- tabler sur l'augmentation constante de la puissance de calcul des ordinateurs.

La seconde solution ne demande aucun effort, tout juste un peu de patience. Pour s'en convaincre, il suffit de remarquer que, dans le cadre de l'optimisation combinatoire de l'échantillonnage par recuit simulé (Chapitre 8), Sacks & Schiller (1988) utilisent un super-ordinateur Cray X-MP. Par ailleurs, Epperson (1995) effectue une analyse d'autocorrélation de structures spatiales simulées par un modèle génétique au moyen d'un super-ordinateur Cray C90. Or, au cours de notre thèse, tous nos calculs ont été effectués sur un simple PC à base de processeur Intel Pentium Pro cadencé à 200 MHz, et au moment où nous rédigeons ces lignes, des PC à base de processeurs Intel Pentium III cadencés à 600 MHz sont déjà disponibles dans le commerce.

Comme l'illustre déjà le travail exposé dans ce mémoire, l'augmentation de la puissance de calcul a un impact très profond sur la pratique de l'analyse et de la modélisation des structures spatiales en écologie. En effet, les méthodes de Monte-Carlo sont très souvent utilisées dans le domaine spatial parce qu'il s'agit d'un domaine riche en problèmes pour lesquels les résultats analytiques sont très difficiles à obtenir, même lorsque le problème est très simple à décrire (Besag & Clifford 1989). En outre, avec les méthodes de Monte-Carlo, les effets de frontière qui doivent être considérés dans toute approche analytique sont évités (Besag & Clifford 1989). Néanmoins, Clifford (1998) dénonce fort à propos les problèmes posés par les méthodes intensives en statistique, et notamment l'impossibilité dans laquelle se trouve le lecteur d'un travail de recherche de juger la validité des résultats, obtenus par un programme qu'il ne lui est pas donné de vérifier ni d'utiliser. Ce problème renvoie directement à la diffusion des logiciels (des exécutables, voire même des sources), notamment par le biais d'Internet.

12.2.5 Biométrie indisciplinaire

De façon assez surprenante, Dodge (1993, p. 37) définit la biométrie comme *"une partie de la biostatistique qui traite particulièrement de la statistique des naissances, des décès, des mariages et des divorces"*. Plus classiquement, le dictionnaire encyclopédique Larousse définit la biométrie comme *"l'application à la biologie des méthodes générales de la statistique"*. En fait, la biométrie ne peut se contenter d'être un domaine d'application de la statistique, parce que les méthodes statistiques ne suffisent pas pour traiter les données observées (Legay 1986b). Plus généralement, toute définition de la biométrie faisant référence à un seul champ disciplinaire bien identifié, tel que la biostatistique, la biomathématique ou la bio-informatique, est réductrice et éloignée de son statut actuel : la biométrie est, au sens large, une science de transfert (Legay 1986b).

Ce qui caractérise la biométrie contemporaine c'est davantage un type d'attitude face aux problématiques biologiques plutôt qu'un ensemble de méthodes, dont la cardinalité ne cesse d'augmenter au fil des années. La démarche qui prévaut en biométrie dépasse la conception classique selon laquelle *"le schéma du naturaliste se trouvera considérablement enrichi grâce à la formulation mathématique beaucoup plus précise"* (Matheron 1970). La biométrie cherche à préciser les objets d'étude, les concepts, les questions, les démarches méthodologiques, en utilisant tous les moyens mis à sa disposition par les disciplines bien établies telles que la mathématique, la statistique et l'informatique (Legay 1986b). La biométrie ne devrait privilégier aucune discipline, méthode ou concept, en fonction d'une hiérarchie discutable, ou pire, en fonction de la mode en vigueur. La biométrie ne devrait pas non plus se limiter à la conception théorique de méthodes mathématiques et à leur mise en oeuvre quasi-automatique (Houllier 1992). L'objet de la biométrie n'est pas nécessairement de faire preuve de virtuosité mathématique, statistique ou informatique, mais de proposer des solutions opérationnelles, qui répondent réellement aux questions posées par les biologistes. Cette démarche impose au biométricien "de terrain" de se comporter autant comme un théoricien, que comme un expérimentateur et un pédagogue, et d'avoir une vision la plus vaste possible des techniques disponibles.

Nous estimons que le travail exposé dans ce mémoire illustre bien l'absence de discipline privilégiée, l'articulation entre différentes méthodes, en vue d'atteindre certains objectifs. En particulier, l'utilisation conjointe des approches structurelle et statistique pour traiter le problème de l'association entre cartes binaires illustre l'assertion selon laquelle l'analyse des données constitue un domaine de choix pour une étude des interactions entre le symbolique et le numérique (Kodratoff & Diday 1991). Cependant, il est assez inconfortable de naviguer systématiquement à l'interface de plusieurs disciplines — encore très compartimentées — et cette démarche "indisciplinaire" (Legay 1986a) présente le risque de ne satisfaire à aucun des critères d'évaluation des parties en présence.

En ce qui concerne les méthodes exposées dans ce mémoire, nous avons essayé de respecter un compromis raisonnable entre leurs aspects purement techniques, leur intérêt pratique pour les utilisateurs, et une certaine dose de pédagogie, même si cela doit générer un sentiment d'insatisfaction chez les spécialistes de chacune des disciplines abordées.

Annexe A

Abréviations

ACP	Analyse en Composantes Principales
AIC	*Akaike Information Criterion*
ANCOVA	*Analysis of Covariance*
ANOVA	*Analysis Of Variance*
BBS	*Breeding Bird Survey*
BIC	*Bayesian Information Criterion*
BLUE	*Best Linear Unbiased Estimator*
BLUP	*Best Linear Unbiased Predictor*
CAH	Classification Ascendante Hiérarchique
CBC	*Christmas Bird Count*
CL	Combinaison Linéaire
CLA	Combinaison Linéaire Autorisée
CONDOR	*Committee on the Next Decade of Operations Research*
DBH	*Diameter at Breast Height*
DT	*Delaunay Triangulation*
EAS	Echantillonnage Aléatoire Simple
EE	Echantillonnage Emboîté
EHT	Estimateur de Horvitz-Thompson
EMST	*Euclidean Minimum Spanning Tree*
ENSMP	Ecole Nationale Supérieure des Mines de Paris
EP	Erreur de Prédiction
ES	Echantillonnage Systématique
FA	Fonction Aléatoire
FAI-k	Fonction Aléatoire Intrinsèque d'ordre k
FAST-2	Fonction Aléatoire Stationnaire à l'ordre 2
fBm	*fractional Brownian motion*
FFT	*Fast Fourier Transform*
FIFO	*First-In First-Out*

GA	*Genetic Algorithms*
GCV	*Generalized Cross-Validation*
GG	*Gabriel Graph*
GLS	*Generalized Least Squares*
GPS	*Global Positioning System*
GRASP	*Greedy Randomized Adaptative Search Procedure*
GSLIB	*Geostatistical Library*
HBDS	*Hypergraph Based Data Structure*
ICES	*International Council for the Exploration of the Sea*
IMSL	*International Mathematics and Statistics Library*
ISV	*Integral of the Semi-Variogram*
IWD	*Inverse Weighted Distance*
KD	Krigeage Disjonctif
KI	Krigeage d'Indicatrices
KL	Krigeage Lognormal
KM	Krigeage Modifié
KO	Krigeage Ordinaire
KP	Krigeage de Probabilités
KS	Krigeage Simple
KT	Krigeage avec un modèle de Tendance
KU	Krigeage Universel
LAUO	Linéarité, Autorisation, Universalité, et Optimalité
LS	*Least Squares*
MAE	*Mean Absolute Error*
MINQ	*Minimum Norm Quadratic*
MINQ(U,I)	*Minimum Norm Quadratic, Unbiased, Invariant*
MIVQ	*Minimum Variance Quadratic*
MIVQ(U,I)	*Minimum Variance Quadratic, Unbiased, Invariant*
ML	*Maximum Likelihood*
MSE	*Mean Squared Error*
MST	*Minimum Spanning Tree*
MVR	Multi-Variable Régionalisée
NEM	*Neighborhood Expectation-Maximisation*
NLLF	*Negative Log-Likelihood Function*
OLS	*Ordinary Least Squares*
OPT	Motif d'échantillonnage Optimal
PC	*Personal Computer*
PQV	*Paired-Quadrat Variance*
PR	Phénomène Régionalisé
PRM	Performance Relative Moyenne

RAPD	*Random Amplified Polymorphic DNA*
REML	*Restricted Maximum Likelihood*
RNG	*Relative Neigbourhood Graph*
SA	*Simulated Annealing*
SCO	*Compact Schools*
SIG	Système d'Information Géographique
SSM	*Small Schools*
STR	Echantillonnage aléatoire Stratifié
SV	*Semi-Variogram*
TS	*Tabu Search*
TTLV	*Two-Terms Local Variance*
UTM	*Universal Transverse Mercator*
VA	Variable Aléatoire
VR	Variable Régionalisée
WLS	*Weighted Least Squares*

Annexe B

Générateurs de nombres aléatoires

Les méthodes intensives utilisées tout au long de notre travail de recherche peuvent se classer en différentes catégories en fonction de leur objectif et de leur principe :

- simulation de Monte-Carlo en géométrie probabiliste,
- simulation de fonctions aléatoires en géostatistique,
- intégration de Monte-Carlo,
- test de randomisation,
- test de Monte-Carlo[1],
- réplication d'un dispositif d'échantillonnage,
- optimisation combinatoire.

D'une façon générale, le recours aux méthodes intensives en statistique ne cesse de se développer, qu'il s'agisse de tester des hypothèses (Edgington 1987, Noreen 1989, Manly 1991, 1993, Good 1994), de réduire le biais d'estimation et de calculer des intervalles de confiance (Manly 1991, 1993, Efron & Tibshirani 1993) ou même d'étudier les procédures statistiques elles-mêmes (*e.g.*, Kowalski 1972, Bivand 1980). L'analyse des données en écologie n'échappe évidemment pas à cette tendance (revue dans Crowley 1992). En particulier, les méthodes de Monte-Carlo sont très souvent utilisées dans le domaine spatial parce qu'il s'agit d'un domaine riche en problèmes pour lesquels les résultats analytiques sont très difficiles à obtenir, même lorsque le problème est très simple à décrire (Besag & Clifford 1989). Bien que le terme *méthode de Monte-Carlo* ait été proposé à la fin des années 1940 par Metropolis & Ulam (1949), le principe des méthodes de Monte-Carlo remonte au moins à Buffon, et celui des simulations de Monte-Carlo en statistique au dernier quart du $19^{ème}$ siècle (Stigler 1991). Depuis l'article fondateur de Metropolis & Ulam (1949), le nombre d'applications des méthodes de Monte-Carlo ne cesse d'augmenter[2], parallèlement à la montée en puissance des moyens de calcul. Il en va naturellement de même pour toutes les méthodes intensives.

[1] Certains auteurs ne font pas la distinction entre les tests de Monte-Carlo et les tests de randomisation (*e.g.*, Manly 1991, 1993). Nous qualifions de *test de Monte-Carlo* toute procédure de test faisant intervenir explicitement un modèle de distribution statistique, paramétrique ou non paramétrique (Noreen 1989).

[2] En 1981, Rubinstein estime que plus de 3000 articles sur la simulation et les méthodes de Monte-Carlo ont été publiés au cours des 15 années précédentes.

Les méthodes intensives nécessitent de grandes quantités de nombres aléatoires. En 1955, la *RAND Corporation* a publié une table de 10^6 nombres aléatoires produits à partir d'une source physique (Rubinstein 1981, p. 20, Ripley 1987, p. 15). Le recours à une telle table est actuellement obsolète, à la fois parce qu'il est courant d'avoir besoin de plus de 10^6 nombres aléatoires, et parce qu'il est plus rapide de calculer directement des nombres aléatoires plutôt que de les chercher dans une table. Dodge (1996) a toutefois proposé de remplacer les 10^6 nombres aléatoires de la *RAND Corporation* par les 10^6 premières décimales de π, ou de stocker plusieurs milliards de décimales de π sur un CD-ROM. En dehors de la question théorique concernant le caractère aléatoire de la séquence des décimales de π — sujet qui n'est pas nouveau puisqu'il est mentionné dans Bouvier & George (1992, p. 354) — la proposition de Dodge (1996) est tout à fait marginale et ne présente aucun intérêt en pratique. Nous considérons donc uniquement la production de nombres aléatoires au moyen d'un algorithme. Il apparaît immédiatement une contradiction en ce qu'il est impossible de générer des nombres aléatoires au sens strict à partir d'un algorithme déterministe (Sedgewick 1991, p. 527). En toute rigueur, il convient donc de parler de *nombres pseudo-aléatoires* et de *simulation déterministe du hasard* (Maurin 1975).

L'objet de cette annexe n'est pas de traiter en détail de la production des nombres pseudo-aléatoires parce qu'il s'agit d'un domaine extrêmement technique, qui a produit une littérature très abondante[3]. Nous nous contentons ici de justifier le choix du générateur que nous utilisons dans notre travail de recherche, ce qui nécessite de considérer successivement le choix du type de générateur, de ses paramètres et de la valeur initialisant le générateur ou *graine*.

B.1 Choix du type de générateur

Une suite de nombres, supposés aléatoires, est caractérisée par la distribution de leurs valeurs et par la façon dont ces valeurs se succèdent (Maurin 1975). En ce qui concerne la distribution, fondamentalement nous cherchons à produire des nombres aléatoires U répartis de façon uniforme dans l'intervalle $]a, b[$, soit selon la fonction densité de probabilité :

$$f(U) = \begin{cases} (b-a)^{-1} & \text{si } a < U < b \\ 0 & \text{sinon} \end{cases} \tag{B.1}$$

En fait, il suffit de savoir générer de façon uniforme des nombres aléatoires U compris dans l'intervalle[4] $]0, 1[$ pour pouvoir ensuite se ramener à une distribution uniforme dans l'intervalle $]a, b[$ et plus généralement, à n'importe quelle forme de distribution. Le problème consiste donc à choisir un algorithme qui permette d'obtenir des valeurs $U \in]0, 1[$ qui satisfont certains critères statistiques.

Cependant, il convient de hiérarchiser les propriétés qui doivent absolument être vérifiées, en fonction de l'application envisagée. En effet, un "mauvais" générateur pour

[3] Sowey (1978) répertorie environ 150 références bibliographiques consacrées à la production et au test de nombres aléatoires pour la seule période 1972-1976. Sowey (1986) estime qu'il se publie sur ce sujet en moyenne 30 articles par an depuis 1970.

[4] La question de savoir si l'intervalle est ouvert ou fermé en 0 dépend des applications et de l'implémentation du générateur (*e.g.*, McLeod 1985). Ici, nous considérons un intervalle ouvert en 0.

une application peut constituer un "bon" générateur pour une autre. Par exemple, les méthodes d'intégration dites de *quasi-Monte-Carlo*[5] peuvent utiliser des nombres de Fibonacci (Ripley 1987, pp. 189-193). Or on sait que les nombres produits par la récursion de Fibonacci $X_{i+1} = (X_i + X_{i-1}) \,(\mathrm{mod}\, 1)$ constituent de piètres nombres pseudo-aléatoires (Rubinstein 1981, p. 24, Ripley 1987, pp. 15-16) et il convient plutôt de parler de nombres *quasi-aléatoires* (Sedgewick 1991, p. 528). Il apparaît donc que des domaines d'application différents tels que, par exemple, la simulation stochastique et l'analyse numérique, n'ont pas nécessairement les mêmes exigences en ce qui concerne le caractère aléatoire des nombres utilisés (Ripley 1983).

Les moyens à mettre en oeuvre dans un générateur sont d'autant plus lourds que les exigences statistiques sont plus nombreuses (Maurin 1975). Il existe théoriquement un ensemble infini d'épreuves concevables pour juger de la perfection du caractère aléatoire d'une séquence de nombres (Maurin 1975). Aucun générateur ne peut passer avec succès tous les tests imaginables. La question qu'il faut se poser n'est pas de savoir si le générateur utilisé satisfait au plus grand nombre de tests possibles, quitte à être inefficace et d'implémentation délicate, mais surtout s'il se révèle approprié pour l'application envisagée (Maurin 1975). En général, les propriétés souhaitées sont l'uniformité de la distribution des valeurs, leur indépendance et une période élevée[6].

Plutôt que de faire confiance aveuglément à un générateur dont la complexité semble être une garantie de qualité, il nous semble préférable de procéder à une analyse critique des résultats en fonction du type de générateur et/ou de ses paramètres. Ce type d'analyse constitue finalement la seule garantie de maîtriser le processus de simulation déterministe du hasard mis en oeuvre au moyen d'un algorithme ou d'un *package* logiciel particulier. Par exemple, Baker (1997) utilise le générateur `ran1` décrit par Press *et al.* (1989, pp. 219-220) dans une analyse statistique de type Monte-Carlo, et obtient des résultats qui présentent une périodicité anormale. Baker (1997) constate que l'implémentation du générateur n'est pas adaptée à son application et montre comment la modifier. Un autre type de problème se pose avec le générateur du *package* géostatistique GSLIB (Deutsch & Journel 1992) qui semble inapproprié pour simuler de nombreuses réalisations parce qu'il serait sujet à la corrélation sérielle (Mowrer 1997). Dans l'étude des séries temporelles par simulation de processus autorégressifs d'ordre 1, Paulsen (1984) considère que, à moins d'utiliser un "mauvais" générateur, le choix du générateur a peu d'impact sur les résultats. Cette conclusion ne peut toutefois pas être généralisée à tous les types d'applications des générateurs de nombres pseudo-aléatoires.

Il nous semble judicieux de choisir un type de générateur de préférence simple — donc rapide et d'implémentation aisée — et dont les propriétés sont bien connues, ce qui est le cas des générateurs congruentiels (Ripley 1983). Toutefois, Coquillard & Hill (1997, p. 166) considèrent qu'un *"scientifique averti doit éviter ce type de générateur ou bien sélectionner avec précaution les paramètres"*. Comme nous proposons d'utiliser un générateur congruentiel, il convient effectivement de choisir ses paramètres avec le plus grand soin.

---

[5] Les méthodes d'intégration de *quasi-Monte-Carlo* exploitent des résultats de la théorie des nombres, assez difficiles d'accès pour un non spécialiste (*cf.* Niederreiter 1978, Fang *et al.* 1994).

[6] Par exemple, pour simuler l'invasion du nord-ouest de la Méditerranée par l'algue *Caulerpa taxifolia*, Hill *et al.* (1998) utilisent un générateur de période $2^{128} \simeq 3.40 \times 10^{38}$ et sont ainsi certains de pouvoir réaliser 10^5 exécutions indépendantes de leur modèle.

B.2 Choix des paramètres du générateur

Nous avons choisi d'utiliser un générateur congruentiel (Rubinstein 1981, p. 21, Ripley 1987, p. 20) :

$$X_{i+1} \equiv (aX_i + c) \,(\mathrm{mod}\, m) \tag{B.2}$$

avec $i = 1, \ldots, n$ et n la longueur de la séquence de nombres pseudo-aléatoires à générer. L'expression (B.2) définit la classe des *générateurs congruentiels mixtes* qui font intervenir à la fois un multiplicateur a, un incrément c et un modulo m, avec a, c, m des entiers non négatifs. En imposant $c = 0$, l'expression (B.2) définit alors la classe des *générateurs congruentiels multiplicatifs*.

Etant donnée une valeur initiale (ou *graine*) X_0, une suite $(X_1, X_2, \ldots, X_n)$ de n nombres pseudo-aléatoires est produite en appliquant séquentiellement (B.2). Des valeurs $U_i \in]0, 1[$ sont obtenues simplement en calculant $U_i = X_i/m$ (Rubinstein 1981, Ripley 1987). La période p de la séquence des U_i ne peut pas excéder m : lorsque $p = m$, la période est qualifiée de *période complète* (*full period*) (Rubinstein 1981).

La définition opératoire d'un générateur congruentiel mixte (B.2) nécessite de choisir les constantes a, c et m. Le choix de m est généralement effectué de sorte que l'opération mod s'effectue efficacement, ce qui revient à choisir $m = r^\beta$ pour un ordinateur travaillant en base r et avec un mot machine de taille β (Rubinstein 1981, p. 23, 25, Ripley 1987, p. 22). Il est classique de choisir $m = 2^{32}$, le nombre premier de Mersenne[7] $m = 2^{31} - 1$, ou encore le nombre premier $m = 2^{16} + 1$ (Ripley 1983). Ce choix étant effectué, il reste encore beaucoup de liberté pour choisir a et c.

Les propositions de paramètres (m, a, c) doivent être évaluées au sens de certains critères à respecter. Les deux critères essentiels sont l'uniformité de la distribution des valeurs et leur indépendance. Si m est suffisamment grand, alors les valeurs U_i sont approximativement distribuées de façon uniforme (Ripley 1987, p. 23). Il reste donc essentiellement à tester l'indépendance des valeurs de la séquence des U_i. Pour ce faire, il existe de nombreux tests que l'on peut classer en :

- tests empiriques, effectués à partir d'une séquence donnée (Chassé & Debouzie 1973, Maurin 1975, pp. 22-26, Rubinstein 1981, pp. 30-33, Ripley 1987, pp. 43-44),
- tests théoriques, effectués directement à partir de l'algorithme et de ses paramètres.

Les tests les plus utiles concernent la structure de treillis induite par la séquence des U_i. Ainsi, le test empirique le plus simple et le plus efficace pour identifier les générateurs qui ne satisfont pas au critère d'indépendance consiste à représenter dans $[0, 1]^2$ le nuage de points de coordonnées (U_i, U_{i+1}) (Maurin 1975, p. 24, Ripley 1983, 1987, pp. 23-24, Sharp & Bays 1992). Si le nuage de points dégénère en lignes régulièrement espacées, c'est qu'il existe des relations fonctionnelles entre tirages successifs, et par conséquent, le critère d'indépendance n'est pas satisfait (Maurin 1975, pp. 25-26, Sharp & Bays 1992). Ce simple test visuel est suffisant pour effectuer un premier tri parmi un ensemble de générateurs. Cependant, une analyse plus fine est nécessaire afin de classer des générateurs qui présentent un nuage d'apparence aléatoire dans $[0, 1]^2$. Il convient alors d'examiner les nuages de points de coordonnées $(U_i, \ldots, U_{i+k-1})$ dans un espace $[0, 1]^k$. Au-delà de

[7] Un nombre de Mersenne est un entier de la forme $M_n = 2^n - 1$, avec n entier (Bouvier & George 1992, p. 470).

$k = 3$, il devient évidemment impossible d'utiliser un test visuel. En revanche, grâce aux tests théoriques, il est possible de calculer des constantes qui résument la structure du nuage dans $[0,1]^k$ (Atkinson 1980, Ripley 1983, 1987, pp. 33-41). Il s'avère que la valeur de l'incrément c joue peu de rôle dans la distribution multidimensionnelle des U_i au contraire du multiplicateur a (Ripley 1983, 1987). En outre, il est souhaitable d'obtenir un générateur dont la période soit la plus élevée possible, *i.e.* au moins 2^{30} (Ripley 1987, p. 26). Les choix importants concernent donc essentiellement a et m.

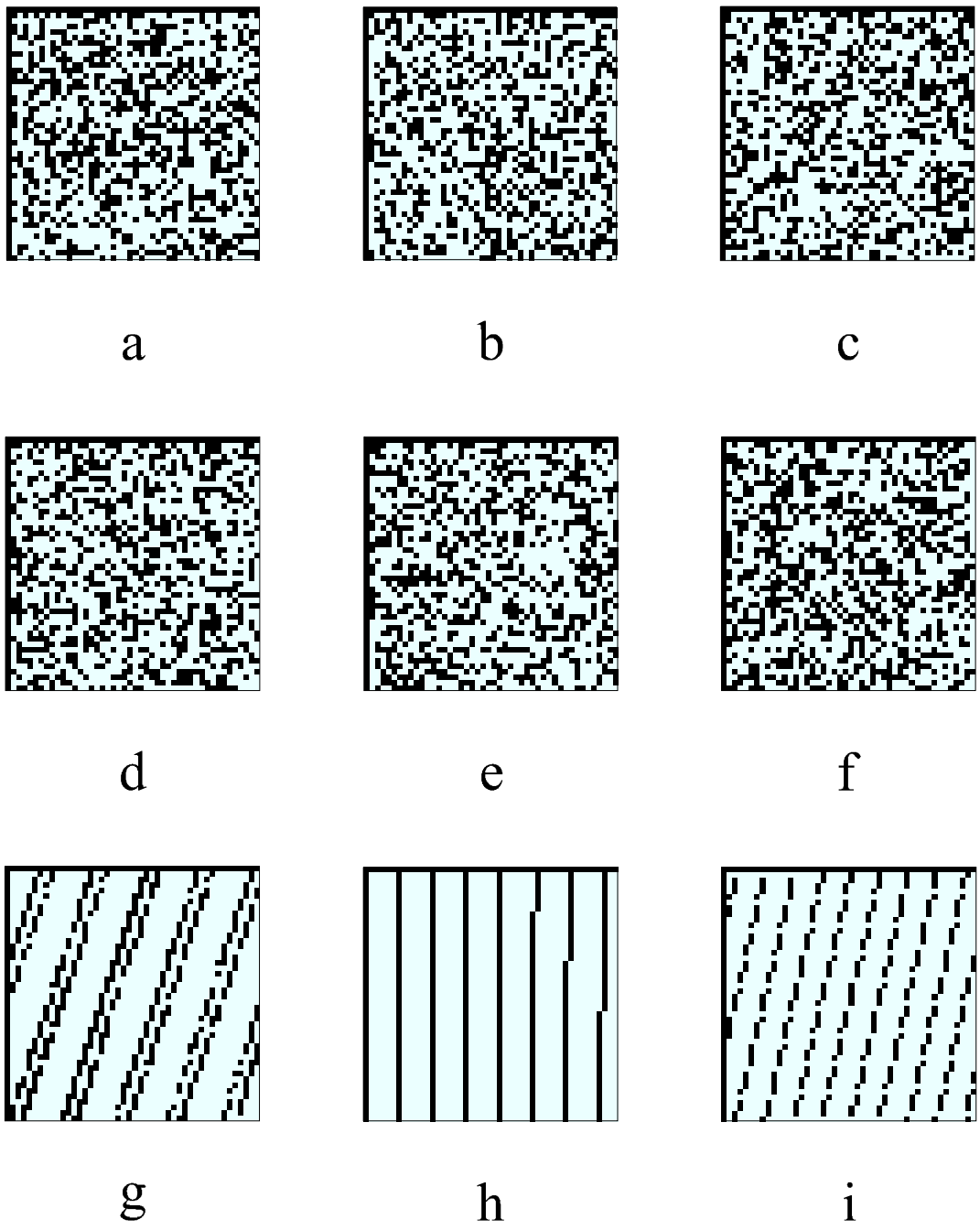
De nombreux générateurs sont comparés au moyen du test visuel par Sharp & Bays (1992) pour $k = 2$ et $k = 3$, au moyen de tests théoriques par Atkinson (1980) pour $2 \leq k \leq 5$, et par Ripley (1983, 1987, p. 39) pour $2 \leq k \leq 4$. Dans ce qui suit, nous comparons neuf générateurs au moyen du test visuel pour $k = 2$, puis les générateurs qui satisfont à ce test préliminaire sont classés au moyen de tests théoriques effectués pour $2 \leq k \leq 8$.

B.2.1 Test visuel

Dans un premier temps, nous avons choisi de comparer visuellement dans $[0,1]^2$ les neuf générateurs suivants :

a. Le générateur ($m = 2^{32}$, $a = 69069$, $c = 1$), utilisé par DEC pour les compilateurs de ses ordinateurs VAX (Ripley 1987, p. 38).
b. Une variante du précédent ($m = 2^{32}-209$, $a = 69069$, $c = 1$), proposée par Marsaglia (1985, *op. cit.* Altman 1988).
c. Le générateur ($m = 2^{31} - 1$, $a = 16807$, $c = 0$), utilisé sur l'ordinateur IBM 360, ainsi que dans le *package* mathématique et statistique IMSL (Rubinstein 1981, p. 26), mentionné notamment par Coquillard & Hill (1997, p. 166).
d. Le générateur ($m = 2^{32}$, $a = 71365$, $c = 1$), proposé par Marsaglia (1972, *op. cit.* Atkinson 1980).
e. Le générateur ($m = 2^{32}$, $a = 100485$, $c = 1$), proposé par Marsaglia (1972, *op. cit.* Atkinson 1980).
f. Le générateur de Wichmann & Hill (1982) qui combine trois générateurs congruentiels, utilisé notamment par Manly (1991, 1993), et dont Zeisel (1986) a montré qu'il pouvait s'exprimer comme le générateur unique ($m = 27817185604309$, $a = 16555425264690$, $c = 0$).
g. Le générateur RANDU ($m = 2^{31}$, $a = 65539$, $c = 0$), utilisé sur les ordinateurs IBM 360/370 et pdp11 (Ripley 1983, 1987, p. 23).
h. Le générateur ($m = 2^{16} + 1$, $a = 75$, $c = 0$), utilisé pour le BASIC d'un des tous premiers micro-ordinateurs, le zx81 de Sinclair (Ripley 1987, p. 38).
i. Le générateur ($m = 2^{18}$, $a = 3125$, $c = 0$), recommandé par Chassé & Debouzie (1973) et utilisé notamment par Houllier (1986) pour simuler une loi multinormale.

Les générateurs (a) à (f) satisfont au test visuel puisque les nuages de points dans $[0,1]^2$ présentent un aspect aléatoire (Fig. B.1.a à B.1.f). En revanche, les générateurs (g) à (i) s'avèrent très mauvais comme en témoignent les alignements de points (Fig. B.1.g à B.1.i). Les générateurs (g) et (h) sont bien connus pour être particulièrement mauvais (Ripley 1983, 1987), et le générateur (i) s'avère tout aussi peu recommandable.



Figure˜B.1: Test visuel pour les neuf générateurs (a) à (i) (détails dans le texte). Chaque nuage comporte 10^5 points (U_i, U_{i+1}). Après capture d'écran du nuage de points, seule la partie supérieure gauche du carré $[0,1]^2$ est représentée afin d'observer les détails. L'effet d'escalier apparent sur les lignes obliques est dû à la résolution le l'écran (*aliasing*).

B.2.2 Tests théoriques

Les tests théoriques ont été développés pour les générateurs de période complète, mais peuvent être étendus aux générateurs multiplicatifs pour lesquels m est premier et $p = m^d - 1$ (Ripley 1983).

Il existe une incertitude en ce qui concerne la période du générateur (f). En effet, Wichmann & Hill (1982) considèrent que $p \simeq 2.78 \times 10^{13}$, ce qui correspond à la valeur de m proposée par Zeisel (1986) et laisse penser que la période est complète, mais par la suite, Wichmann & Hill (1984) affirment que la période vaut "seulement" $p \simeq 6.95 \times 10^{12}$. En conséquence, il n'est pas certain que les tests théoriques puissent s'appliquer au générateur (f).

Sans entrer dans les détails des tests théoriques, les trois constantes que nous calculons sont l_k/l_1 qui doit être proche de 1, l_k qui doit être la plus faible possible et ν qui doit être la plus élevée possible (Ripley 1983). Les calculs ont été effectués à partir de l'algorithme décrit dans Ripley (1987, pp. 220-226), en utilisant des flottants codés sur 10 octets. Les résultats permettent de classer les générateurs (a) à (f) par ordre croissant de qualité : c < a < b < e < d < f (Tab. B.1).

	k	2	3	4	5	6	7	8
	l_k/l_1	1.06	1.29	1.30	1.25	4.71	3.31	2.27
a	l_k	1.61	72.54	464.96	1363.76	6484.14	8718.96	9421.71
	ν	65139.93	1439.63	229.79	83.61	15.56	13.04	13.04
	l_k/l_1	1.06	1.05	1.66	2.11	1.45	2.85	2.42
b	l_k	1.61	64.29	552.26	1729.79	3425.96	8577.64	10630.09
	ν	65202.47	1650.01	188.96	66.60	34.91	12.04	11.05
	l_k/l_1	7.60	3.39	2.07	1.67	1.67	1.84	2.32
c	l_k	5.95	159.21	704.32	1730.64	3528.76	6925.85	9870.05
	ν	16807.00	638.90	147.25	66.63	29.92	16.55	12.65
	l_k/l_1	1.18	1.12	1.18	1.37	2.29	2.31	1.71
d	l_k	1.66	69.08	441.57	1478.93	3379.68	8132.74	8924.74
	ν	60648.65	1652.60	246.65	85.70	35.55	13.27	13.27
	l_k/l_1	1.34	1.10	1.37	1.14	3.12	1.70	1.65
e	l_k	1.77	69.96	468.84	1381.98	4343.51	5736.39	8392.53
	ν	56811.83	1543.21	256.79	86.57	24.37	20.45	13.64
	l_k/l_1	1.57	2.07	1.41	1.20	1.37	2.18	1.39
f	l_k	0.02	4.55	54.31	235.17	775.91	1867.92	2790.69
	ν	4218961.34	22583.42	1845.67	481.66	158.03	59.14	40.35

Tableau B.1: Résultats des tests théoriques pour les générateurs (a) à (f) dans les dimensions $k = 2$ à $k = 8$ (détails dans le texte). La constante l_k/l_1 doit être proche de 1, la constante l_k doit être la plus faible possible et la constante ν doit être la plus élevée possible. Les valeurs de l_k doivent être multipliées par un facteur 10^{-5}.

Compte tenu de l'incertitude attachée aux tests théoriques du générateur (f), nous avons finalement retenu le générateur (d). Sa période est égale à $m = 2^{32}$, soit $p \simeq 4.29 \times 10^9$, et bien qu'elle soit inférieure à celle du générateur (f), elle nous semble largement suffisante pour la plupart des applications[8].

Cependant, s'il s'agit de traiter les nombres pseudo-aléatoires comme des chaînes de bits, le classement des générateurs serait différent. Ainsi, les générateurs (a) et (b) ne satisfont pas aux tests du caractère aléatoire des bits, au contraire du générateur (c) qui s'avère acceptable (Altman 1988). En conséquence, nous recommandons le générateur (d) lorsqu'il s'agit d'utiliser des nombres pseudo-aléatoires $U \in]0, 1[$, mais nous ne garantissons pas un comportement aléatoire au niveau des bits eux-mêmes.

B.3 Choix de la graine du générateur

A la suite des tests visuels et théoriques, nous avons choisi d'utiliser le générateur congruentiel mixte :

$$X_{i+1} \equiv (71365 \cdot X_i + 1) \left(\mathrm{mod}\, 2^{32}\right) \tag{B.3}$$

Nous avons implémenté ce générateur en Delphi (avatar du Pascal) de façon simple, sans garantie d'optimalité en termes de temps d'exécution, comme suit :

```
{------------------------------------------------------------------------}
{ Méthode congruentielle mixte => résultat dans ]0,1[                    }
{------------------------------------------------------------------------}

Function Alea(var seed : EXTENDED) : EXTENDED;

    const modulo : EXTENDED = 4294967296.0;

    var deno, seed8, delta, ipart : EXTENDED;

begin
    seed8:=seed*71365.0+1.0;
    delta:=seed8-modulo;
    if (delta>0) then begin
        deno:=delta/modulo;
        ipart:=int(deno);
        seed8:=seed8-(ipart+1.0)*modulo;
    end;
    seed:=seed8;
    Result:=seed8/modulo;
end;
```

Le type EXTENDED est un type de flottant codé sur 10 octets, comportant 19 ou 20 chiffres significatifs, et couvrant une portée s'étendant de 3.4×10^{-4932} à 1.1×10^{4932} (Borland 1996).

[8] MacLaren (1992) considère cependant qu'il ne faut pas utiliser de séquences dépassant $p^{2/3}$ valeurs sous peine d'obtenir une uniformité excessive par rapport à une séquence aléatoire stricte.

La séquence des nombres pseudo-aléatoires produite par la fonction Alea est entièrement déterminée par la graine du générateur seed. Cette graine peut être choisie de façon à obtenir une séquence qui satisfait au critère d'uniformité. Il existe plusieurs tests envisageables pour tester l'uniformité (Maurin 1975, pp. 5-22, Rubinstein 1981, pp. 26-30). L'approche la plus courante consiste à utiliser le test du χ^2 de conformité en divisant $]0,1[$ en de nombreux sous-intervalles, ou le test de Kolmogorov-Smirnov à partir de la fonction de répartition empirique de la séquence $(U_1,\ldots,U_n)$ (Ripley 1987, p. 44, Coquillard & Hill 1997, pp. 172-174).

Nous avons choisi de calculer le χ^2 de conformité en divisant $]0,1[$ en 100 sous-intervalles, pour 50 séquences de 10^5 nombres pseudo-aléatoires générées par la fonction Alea, à partir de 50 valeurs de seed, elles-mêmes générées aléatoirement entre 1 et 2000. Parmi les 10 séquences présentant la plus faible valeur du χ^2 de conformité, la meilleure graine obtenue est $X_0 = 603$, suivie par la graine $X_0 = 1374$ (Tab. B.2).

En conséquence, pour générer une séquence de nombres pseudo-aléatoires en utilisant la fonction Alea, nous utilisons la graine $X_0 = 603$, et lorsqu'il s'avère nécessaire de produire une seconde séquence, indépendante de la première, nous utilisons aussi la graine $X_0 = 1374$. Les autres valeurs figurant dans le Tableau B.2 sont utilisées pour apprécier la sensibilité d'une procédure au choix de la graine.

	1	2	3	4	5	6	7	8	9	10
X_0	603	1374	115	278	1896	768	756	643	1221	828
χ^2_{obs}	66.08	72.35	78.78	79.34	83.55	84.20	84.76	85.41	85.51	86.10

Tableau B.2: Valeurs du χ^2 de conformité observé (χ^2_{obs}) en fonction de la graine X_0 utilisée pour initialiser la fonction Alea (détails dans le texte).

Annexe C

Théorie des isolignes

Les propriétés des isolignes constituent un thème d'étude ancien (Cayley 1859, Maxwell 1870) qui a refait surface avec l'avènement des ordinateurs et la cartographie automatique (*e.g.*, Morse 1968, 1969). Dans cette annexe nous proposons quelques éléments de théorie des isolignes élaborés à partir du modèle proposé par Bouillé (1975) (Section 2.3.3), sans toutefois prétendre à la rigueur mathématique. Ces éléments de théorie des isolignes permettent de construire des algorithmes traitant la topologie d'une carte isoplèthe en mode vecteur de façon relativement efficace.

Soit I un ensemble d'isolignes défini sur le géoïde. Toute isoligne $a \in I$ est une courbe fermée séparant la surface du géoïde en un domaine intérieur $\overset{\circ}{a}$ et en un domaine complémentaire $\complement a$. Soit $\mathcal{C}$ l'ensemble des cotes des isolignes de I ordonné par les relations "prédécesseur" ($\prec$) et "successeur" ($\succ$) dans $\mathbb{R}$, notées fonctionnellement $\alpha = \text{pred}\,(\beta)$ et $\beta = \text{succ}\,(\alpha)$, avec $\alpha, \beta \in \mathcal{C}$. Soit z une fonction $z : I \rightarrow \mathcal{C}$ qui à toute isoligne $a \in I$ associe la cote correspondante $\alpha \in \mathcal{C}$. Plusieurs relations topologiques peuvent être définies entre les isolignes de I.

Définition 1 *Le prédicat* $\text{Inc}(a, b)$ *est vrai si* a *inclut* b *au sens :*

$$\text{Inc}(a, b) \Leftrightarrow \overset{\circ}{b} \subset \overset{\circ}{a} \tag{C.1}$$

Le prédicat Inc *définit une relation d'inclusion topologique réflexive, antisymétrique et transitive.*

Définition 2 *La relation* a Jonc b *est vérifiée s'il est possible de joindre* $a \in I$ *et* $b \in I$ *sans traverser aucune autre isoligne*[1] $c \in I$*. La relation* Jonc *est une relation topologique réflexive, symétrique et non transitive.*

Définition 3 *La fonction* $\text{Inf}\,(a) = b$ *est définie comme la restriction de la relation* Jonc *:*

$$\text{Inf} : I \rightarrow I \Leftrightarrow \text{Jonc} : A \rightarrow B \tag{C.2}$$

avec $A, B \subset I$*, telle que pour tout* $a \in A$ *et tout* $b \in B$ *on ait* $\beta = \text{pred}\,(\alpha)$ *avec* $z\,(a) = \alpha$ *et* $z\,(b) = \beta$*. La fonction* Inf *définit une relation topologique antiréflexive, antisymétrique et non transitive.*

[1] Morse (1969) définit deux isolignes comme adjacentes s'il est possible de les relier par une ligne ne traversant aucune autre isoligne.

Définition 4 *La fonction* $\mathrm{Sup}\,(a) = b$ *est définie comme la réciproque :*

$$\mathrm{Sup} : I \to I \Leftrightarrow \mathrm{Inf}^{-1} : I \to I \tag{C.3}$$

autrement dit, comme la restriction :

$$\mathrm{Sup} : I \to I \Leftrightarrow \mathrm{Jonc} : A \to B \tag{C.4}$$

avec $A, B \subset I$*, telle que pour tout* $a \in A$ *et tout* $b \in B$ *on ait* $\beta = \mathrm{succ}\,(\alpha)$ *avec* $Z\,(a) = \alpha$ *et* $Z\,(b) = \beta$*. La fonction* Sup *définit une relation topologique antiréflexive, antisymétrique et non transitive.*

Définition 5 *La relation* $a\,\mathrm{Vois}\,b$ *est définie comme la restriction de la relation* $a\,\mathrm{Jonc}\,b$ *:*

$$\mathrm{Vois} : I \to I \Leftrightarrow \mathrm{Jonc} : H \to H \tag{C.5}$$

avec $H \subset I$*, telle que pour tout* $h \in H$ *on ait* $z\,(h) = \alpha$*. Autrement dit, la relation* $a\,\mathrm{Vois}\,b$ *est vérifiée si la relation* $a\,\mathrm{Jonc}\,b$ *est vérifiée et si* $z\,(a) = z(b)$*. La relation topologique* Vois *est réflexive, symétrique et non transitive.*

A partir des définitions précédentes, nous proposons de formaliser les propriétés essentielles des isolignes de I grâce à un certain nombre de théorèmes.

Théorème 1 *Pour toutes les isolignes* $a, b \in I$ *telles que* $a \cap b \neq \emptyset$ *on a, soit* $\overset{\circ}{a} \cap \overset{\circ}{b} = \overset{\circ}{a} \Leftrightarrow \mathrm{Inc}\,(b, a)$*, soit* $\overset{\circ}{a} \cap \overset{\circ}{b} = \overset{\circ}{b} \Leftrightarrow \mathrm{Inc}\,(a, b)$*. Autrement dit, aucune isoligne ne traverse une autre isoligne.*

Corollaire 1 *Soient deux isolignes* $a, b \in I$ *:*

$$\neg\,\mathrm{Inc}\,(a, b) \;\wedge\; \neg\,\mathrm{Inc}\,(b, a) \Leftrightarrow \overset{\circ}{a} \cap \overset{\circ}{b} = \emptyset \tag{C.6}$$

Corollaire 2 *Soient deux isolignes* $a, b \in I$ *:*

$$\mathrm{Inc}\,(a, b) \Rightarrow \neg\,\mathrm{Inc}\,(b, a) \tag{C.7}$$

La relation Inc *étant une relation antisymétrique, l'implication (C.7) est simple car de* $\neg\,\mathrm{Inc}\,(b, a)$ *on ne peut pas déduire que* $\mathrm{Inc}\,(a, b)$ *est vrai : il est possible d'avoir* $\mathrm{Inc}\,(a, b)$ *ou bien* $\neg\,\mathrm{Inc}\,(a, b)$*.*

Théorème 2 *Soient deux isolignes* $a, b \in I$*, avec* $z\,(a) = \alpha$*,* $z\,(b) = \beta$ *telles que* $\alpha = \beta$*, ou bien* $\alpha = \mathrm{pred}\,(\beta)$*, ou bien* $\alpha = \mathrm{succ}\,(\beta)$ *:*

$$a\,\mathrm{Jonc}\,b \Leftrightarrow @ \in I - \{a, b\} \mid \mathrm{Inc}\,(c, a) \neq \mathrm{Inc}\,(c, b) \tag{C.8}$$

Corollaire 3 *Soient deux isolignes* $a, b \in I$*, avec* $z\,(a) = \alpha$*,* $z\,(b) = \beta$ *telles que* $\alpha = \beta$*, ou bien* $\alpha = \mathrm{pred}\,(\beta)$*, ou bien* $\alpha = \mathrm{succ}\,(\beta)$*, si* $\overset{\circ}{a} \cap \overset{\circ}{b} = \emptyset$ *alors :*

$$a\,\mathrm{Jonc}\,b \Leftrightarrow @ \in I - \{a, b\} \mid \mathrm{Inc}\,(c, a) \;\wedge\; \neg\,\mathrm{Inc}\,(c, b) \;\vee\; \mathrm{Inc}\,(c, b) \;\wedge\; \neg\,\mathrm{Inc}\,(c, a) \tag{C.9}$$

Corollaire 4 *Soient deux isolignes* $a, b \in I$*, avec* $z(a) = \alpha$*,* $z(b) = \beta$ *telles que* $\alpha = \beta$*, ou bien* $\alpha = \mathrm{pred}(\beta)$*, ou bien* $\alpha = \mathrm{succ}(\beta)$*, si* $\mathrm{Inc}(a, b)$ *alors :*

$$a \,\mathrm{Jonc}\, b \Leftrightarrow @ \in I - \{a, b\} \mid \mathrm{Inc}(a, c) \;\wedge\; \mathrm{Inc}(c, b) \tag{C.10}$$

Désignons par $\mathrm{Ferm}(a)$ un prédicat vrai lorsque son argument $a \in I$ est une isoligne fermée. Un ensemble d'isolignes I sera dit défini dans une *topologie globale* s'il constitue un ensemble topologiquement cohérent au sens des théorèmes fondamentaux (1) et (2), et s'il respecte le théorème suivant :

Théorème 3 *Si* I *est un ensemble d'isolignes défini dans une topologie globale :*

$$a \in I \Rightarrow \mathrm{Ferm}(a) \tag{C.11}$$

A partir d'un ensemble d'isolignes I défini dans une topologie globale, une carte isoplèthe est obtenue en appliquant un polygone D de géométrie quelconque sur I. Les isolignes de I qui ne traversent pas D restent fermées, tandis que les isolignes de I qui traversent D sont ouvertes : ceci entraîne une restriction des relations topologiques qu'elles entretenaient auparavant avec les autres isolignes de I, *i.e.* dans la topologie globale. La *restriction topologique* opérée par le polygone D conduit donc à un ensemble d'isolignes J défini dans une *topologie locale*.

Notons la restriction topologique $\mathrm{Rest} : I \rightarrow J$ par le polygone D sous la forme d'un opérateur préfixé :

$$I' \,\mathrm{Rest}\,(D) = J \tag{C.12}$$

Il est également possible de définir une *extrapolation topologique* $\mathrm{Ext} : J \rightarrow \{I_1, I_2, \ldots, I_n\}$ conduisant à n ensembles d'isolignes topologiquement cohérents. De même que la restriction topologique, l'extrapolation topologique est notée sous la forme d'un opérateur préfixé :

$$J' \,\mathrm{Ext}\,(D) = I_k \tag{C.13}$$

où I_k désigne un élément de $\{I_1, I_2, \ldots, I_n\}$. L'extrapolation topologique n'a pas vocation à reconstituer l'ensemble I de départ, autrement dit, l'opérateur $\mathrm{Ext}(\cdot)$ n'est pas l'opérateur réciproque de $\mathrm{Rest}(\cdot)$, de sorte que l'application successive des deux opérateurs conduit aux résultats suivants :

$$I' \,\mathrm{Rest}\,(D)' \,\mathrm{Ext}\,(D) = \{I_1, I_2, \ldots, I_n\} \;;\quad I \in \{I_1, I_2, \ldots, I_n\} \tag{C.14}$$

$$J' \,\mathrm{Ext}\,(D)' \,\mathrm{Rest}\,(D) = K \Rightarrow K = J \tag{C.15}$$

Théorème 4 *Soient deux isolignes* $a \in I$ *et* $b \in J$ *telles que* $a' \,\mathrm{Rest}\,(D) = b$ *:*

$$a = b \Leftrightarrow \mathrm{Ferm}(b) \tag{C.16}$$

Théorème 5 *Soient deux isolignes* $a \in I$ *et* $b \in J$ *telles que* $b' \,\mathrm{Ext}\,(D) = a$ *:*

$$b = a \Leftrightarrow \mathrm{Ferm}(b) \tag{C.17}$$

Les théorèmes (4) et (5) expriment le fait que l'invariance par la restriction topologique ou par l'extrapolation topologique ne concerne que les isolignes fermées dans la topologie locale.

Théorème 6 *Soient les isolignes* $a, b \in I$ *et* $i, j \in J$ *telles que* $a \,\mathrm{Jonc}\, b$, $a' \,\mathrm{Rest}\,(D) = i$ *et* $b' \,\mathrm{Rest}\,(D) = j$ *:*

$$a = i,\ b = j,\ \mathring{a} \cap \mathring{b} \neq \emptyset \Rightarrow i \,\mathrm{Jonc}\, j \tag{C.18}$$

Théorème 7 *Soient les isolignes* $a, b \in I$ *et* $i, j \in J$ *telles que* $a \,\mathrm{Jonc}\, b$, $a' \,\mathrm{Rest}\,(D) = i$ *et* $b' \,\mathrm{Rest}\,(D) = j$ *:*

$$i \,\mathrm{Jonc}\, j \Leftrightarrow @ \in I \mid a \,\mathrm{Jonc}\, c,\ c \,\mathrm{Jonc}\, b,\ \mathrm{Card}\,[c' \,\mathrm{Rest}\,(D)] > 1 \tag{C.19}$$

Théorème 8 *Soient deux isolignes* $a, b \in I$*, avec* $z(a) = \alpha$ *et* $z(b) = \beta$ *telles que* $\alpha = \beta$*, ou bien* $\alpha = \mathrm{pred}\,(\beta)$*, ou bien* $\alpha = \mathrm{succ}\,(\beta)$ *:*

$$\begin{array}{l} a \,\mathrm{Jonc}\, b \text{ non vérifiée à cause de } F \subset I - \{a, b\} \Rightarrow \\ \exists c, d \in F \mid a \,\mathrm{Jonc}\, c \ \wedge\ b \,\mathrm{Jonc}\, d \end{array} \tag{C.20}$$

Autrement dit, si un ensemble d'isolignes F *empêche de joindre* a *et* b *au sens de* Jonc *(avec* $z(a)$ *et* $z(b)$ *telles que la relation* $a \,\mathrm{Jonc}\, b$ *pourrait être vérifiée), alors dans cet ensemble* F *on trouve une isoligne* c *telle que* $a \,\mathrm{Jonc}\, c$*, et une isoligne* d *telle que* $b \,\mathrm{Jonc}\, d$ *(éventuellement* $c = d$*).*

Corollaire 5 *Soient deux isolignes* $a, b \in I$*, avec* $z(a) = \alpha$ *et* $z(b) = \beta$ *telles que* $\alpha = \mathrm{pred}\,(\beta)$*ou bien* $\alpha = \mathrm{succ}\,(\beta)$ *:*

$$a \,\mathrm{Jonc}\, b \text{ non vérifiée à cause de } F \subset I \Rightarrow \exists f \in F \mid\ z(f) \in \{\alpha, \beta\} \tag{C.21}$$

Corollaire 6 *Soient deux isolignes* $a, b \in I$*, avec* $z(a) = \alpha$ *et* $z(b) = \beta$ *telles que* $\alpha = \beta$ *:*

$$\begin{array}{l} a \,\mathrm{Jonc}\, b \text{ non vérifiée à cause de } F \subset I \Rightarrow \\ \exists f \in F \mid z(f) \in \{\mathrm{pred}\,(\alpha), \alpha, \mathrm{succ}\,(\alpha)\} \end{array} \tag{C.22}$$

Théorème 9 *Il n'existe pas deux isolignes* $a \neq b \in I$ *telles qu'il existe* $i, j \in I$ *avec* $\mathrm{Sup}\,(a) = \mathrm{Sup}\,(b) = i$ *et* $\mathrm{Inf}\,(a) = \mathrm{Inf}\,(b) = j$*. Autrement dit, deux isolignes ne peuvent pas avoir à la fois la même isoligne* Inf *et la même isoligne* Sup*.*

Théorème 10 *Soient deux isolignes* $a \neq b \in I$ *:*

$$a \,\mathrm{Vois}\, b\ ;\quad \mathrm{Inf}\,(a) = \mathrm{Inf}\,(b)\ \ \textit{ou bien}\ \ \mathrm{Sup}\,(a) = \mathrm{Sup}\,(b) \tag{C.23}$$

Autrement dit, deux isolignes peuvent être Vois *sans pour autant avoir les mêmes isolignes* Inf *ou bien les mêmes isolignes* Sup*.*

Théorème 11 *Soient deux isolignes* $a \neq b \in I$ *:*

$$\mathrm{Inf}\,(a) = \mathrm{Inf}\,(b)\ \ \textit{ou bien}\ \ \mathrm{Sup}\,(a) = \mathrm{Sup}\,(b)\ ;\quad a \,\mathrm{Vois}\, b \tag{C.24}$$

Autrement dit, deux isolignes qui ont les mêmes isolignes Inf *ou bien les mêmes isolignes* Sup *ne sont pas nécessairement des isolignes* Vois*.*

Corollaire 7 *Soient deux isolignes* $a \neq b \in I$ *:*

$$\mathrm{Inf}(a) = \mathrm{Inf}(b) \;\; \textit{ou bien} \;\; \mathrm{Sup}(a) = \mathrm{Sup}(b) < \; a \,\mathrm{Vois}\, b \tag{C.25}$$

Autrement dit, on ne peut pas déduire toutes les relations Vois *à partir des relations* Inf *et* Sup*, et à partir des relations* Inf *et* Sup *on risque de déduire des relations* Vois *qui n'existent pas.*

Théorème 12 *Soient deux isolignes* $a, b \in I$ *telles que* $z(a) = z(b) = \alpha$ *et la relation* $a \,\mathrm{Vois}\, b$ *non vérifiée à cause de* $F \subset I - \{a, b\}$ *:*

$$\begin{array}{l} @\mathrm{Inf}(a)\,,\; @\mathrm{Sup}(a)\,,\; @\mathrm{Inf}(b)\,,\; @\mathrm{Sup}(b) \Rightarrow \\ \exists c, d \in F \mid z(c) = z(d) = \alpha,\; a \,\mathrm{Jonc}\, c,\; b \,\mathrm{Jonc}\, d \end{array} \tag{C.26}$$

Théorème 13 *Soient deux isolignes* $a, b \in I$ *telles que* $z(a) = z(b) = \alpha$ *et la relation* $a \,\mathrm{Vois}\, b$ *non vérifiée à cause de* $F \subset I - \{a, b\}$ *:*

$$\begin{array}{l} @\mathrm{Inf}(a)\,,\; @\mathrm{Sup}(a)\,,\; @\mathrm{Inf}(b)\,,\; \exists\,\mathrm{Sup}(b) \Rightarrow \\ \exists c, d \in F \mid z(c), z(d) \in \{\alpha, \mathrm{succ}(\alpha)\}\,,\; a \,\mathrm{Jonc}\, c,\; b \,\mathrm{Jonc}\, d \end{array} \tag{C.27}$$

Théorème 14 *Soient deux isolignes* $a, b \in I$ *telles que* $z(a) = z(b) = \alpha$ *et la relation* $a \,\mathrm{Vois}\, b$ *non vérifiée à cause de* $F \subset I - \{a, b\}$ *:*

$$\begin{array}{l} @\mathrm{Inf}(a)\,,\; @\mathrm{Sup}(a)\,,\; \exists\,\mathrm{Inf}(b)\,,\; @\mathrm{Sup}(b) \Rightarrow \\ \exists c, d \in F \mid z(c), z(d) \in \{\mathrm{pred}(\alpha), \alpha\}\,,\; a \,\mathrm{Jonc}\, c,\; b \,\mathrm{Jonc}\, d \end{array} \tag{C.28}$$

Théorème 15 *Soient deux isolignes* $a, b \in I$ *telles que* $z(a) = z(b) = \alpha$ *et la relation* $a \,\mathrm{Vois}\, b$ *non vérifiée à cause de* $F \subset I - \{a, b\}$ *:*

$$\begin{array}{l} @\mathrm{Inf}(a)\,,\; @\mathrm{Sup}(a)\,,\; \exists\,\mathrm{Inf}(b)\,,\; \exists\,\mathrm{Sup}(b) \;\vee\; \exists\,\mathrm{Inf}(a)\,,\; @\mathrm{Sup}(a)\,,\; @\mathrm{Inf}(b)\,,\; \exists\,\mathrm{Sup}(b) \Rightarrow \\ \exists c, d \in F \mid z(c), z(d) \in \{\mathrm{pred}(\alpha), \alpha, \mathrm{succ}(\alpha)\}\,,\; a \,\mathrm{Jonc}\, c,\; b \,\mathrm{Jonc}\, d \end{array} \tag{C.29}$$

A partir des théorèmes (12) à (15) il est possible de déduire le sous-ensemble de $\mathcal{C}$ qu'il suffit de considérer lorsque $\mathrm{Inf}(a) \neq \mathrm{Inf}(b)$ et $\mathrm{Sup}(a) \neq \mathrm{Sup}(b)$ dans la recherche des isolignes Vois effectuée en utilisant la connaissance acquise au sujet des isolignes Inf et Sup :

	$@\mathrm{Inf}(a)$ $@\mathrm{Inf}(b)$	$@\mathrm{Inf}(a)$ $\exists\,\mathrm{Inf}(b)$	$\exists\,\mathrm{Inf}(a)$ $\exists\,\mathrm{Inf}(b)$
$@\mathrm{Sup}(a)$ $@\mathrm{Sup}(b)$	α	$\mathrm{pred}(\alpha)$ α	$\mathrm{pred}(\alpha)$ α
$@\mathrm{Sup}(a)$ $\exists\,\mathrm{Sup}(b)$	α $\mathrm{succ}(\alpha)$	$\mathrm{pred}(\alpha)$ α $\mathrm{succ}(\alpha)$	$\mathrm{pred}(\alpha)$ α $\mathrm{succ}(\alpha)$
$\exists\,\mathrm{Sup}(a)$ $\exists\,\mathrm{Sup}(b)$	α $\mathrm{succ}(\alpha)$	$\mathrm{pred}(\alpha)$ α $\mathrm{succ}(\alpha)$	$\mathrm{pred}(\alpha)$ α $\mathrm{succ}(\alpha)$

Annexe D

Matrices semi-définies positives

Soit $Z(\cdot)$ une fonction aléatoire définie sur un ensemble de supports $s = \{s_i \mid i = 1, \ldots, n\}$. Considérons par exemple que $Z(\cdot)$ est isotrope et stationnaire à l'ordre 2, de covariance $C(h) = C(0) - \gamma(h)$ avec $\gamma(h)$ un modèle de variogramme borné de seuil $C(0)$. Soit $\mathbf{C}$ la matrice de covariance $n \times n$ entre les variables aléatoires $Z(u_i)$ et $Z(u_j)$ avec $i, j = 1, \ldots, n$. En vertu de l'isotropie et de la stationarité à l'ordre 2 de $Z(\cdot)$, les éléments de $\mathbf{C}$ peuvent être calculés comme $c_{ij} = C(h_{ij})$ avec h_{ij} la distance Euclidienne entre les supports s_i et s_j des variables aléatoires associées $Z(s_i)$ et $Z(s_j)$. La matrice de covariance $\mathbf{C}$ intervient notamment dans le krigeage et la simulation des fonctions aléatoires. Une caractéristique essentielle de $\mathbf{C}$ est d'être symétrique et *semi-définie positive*.

Dans cette annexe, nous énumérons un ensemble de propriétés des matrices semi-définies positives (Gower 1971, Golub & van Loan 1983, Wackernagel 1993). Certaines de ces propriétés peuvent être exploitées afin de tester le caractère semi-défini positif d'une matrice de covariance qui n'aurait pas été calculée à partir d'un des modèles de variogrammes mentionnés en Annexe F.

Définition 6 *Une matrice* $\mathbf{C} \in I\!R^{n \times n}$ *est définie positive si et seulement si :*

$$\mathbf{x}^{\mathrm{T}}\mathbf{C}\mathbf{x} > 0 \quad \mathbf{0} \neq \mathbf{x} \in I\!R^n \tag{D.1}$$

Définition 7 *Une matrice* $\mathbf{C} \in I\!R^{n \times n}$ *est indéfinie*[1] *si* $\mathbf{x}^{\mathrm{T}}\mathbf{C}\mathbf{x} > 0$ *pour certains* $\mathbf{x}$ *et* $\mathbf{x}^{\mathrm{T}}\mathbf{C}\mathbf{x} < 0$ *pour d'autres* $\mathbf{x}$.

Théorème 16 *Une matrice* $\mathbf{C} \in I\!R^{n \times n}$ *est définie positive si et seulement si ses valeurs propres sont telles que* $\lambda_i > 0$ *pour* $i = 1, \ldots, n$.

Définition 8 *Une matrice* $\mathbf{C} \in I\!R^{n \times n}$ *est semi-définie positive si et seulement si :*

$$\mathbf{x}^{\mathrm{T}}\mathbf{C}\mathbf{x} \geq 0 \quad \mathbf{x} \in I\!R^n \tag{D.2}$$

Théorème 17 *Une matrice* $\mathbf{C} \in I\!R^{n \times n}$ *est semi-définie positive si et seulement si ses valeurs propres sont telles que* $\lambda_i \geq 0$ *pour* $i = 1, \ldots, n$.

[1] Dans le cas où C est une matrice de covariance indéfinie, Schwertman & Allen (1979) proposent de trouver la matrice semi-définie positive C^0 la plus proche de C au sens des moindres carrés.

Théorème 18 *Une matrice* $\mathbf{C} \in I\!R^{n\times n}$ *est semi-définie positive si et seulement si tous les mineurs principaux* Δ_{pp} $(p = 1, \ldots, n)$ *sont non-négatifs.*

Un mineur principal est le déterminant d'une sous-matrice principale de $\mathbf{C}$. Une sous-matrice principale est obtenue en biffant k colonnes $(k = 0, \ldots, n-1)$ et les lignes les croisant sur les éléments diagonaux. La combinatoire des mineurs principaux à vérifier rend ce critère peu intéressant en pratique pour $n > 3$ (Wackernagel 1993).

Théorème 19 *Soit* $\mathbf{X} \in I\!R^{n\times n}$. *Toute matrice* $\mathbf{C} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$ *est semi-définie positive.*

Théorème 20 *Soient deux matrices* $\mathbf{A}, \mathbf{B} \in I\!R^{n\times n}$ *semi-définies positives, la somme* $\mathbf{C} = \mathbf{A} + \mathbf{B}$ *est semi-définie positive puisque* $\mathbf{x}^{\mathrm{T}}\mathbf{C}\mathbf{x} = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x} \geq 0$

Définition 9 *Soient* $\mathbf{A}, \mathbf{B} \in I\!R^{n\times n}$ *deux matrices d'éléments respectifs* a_{ij} *et* b_{ij} $(i, j = 1, \ldots, n)$. *Notons* $\mathbf{C} = \mathbf{A} \circledast \mathbf{B}$ *le produit matriciel tel que les éléments de* $\mathbf{C}$ *sont* $c_{ij} = a_{ij} \times b_{ij}$.

Théorème 21 *Si* $\mathbf{A}, \mathbf{B} \in I\!R^{n\times n}$ *sont deux matrices symétriques et semi-définies positives alors* $\mathbf{C} = \mathbf{A} \circledast \mathbf{B}$ *est également symétrique et semi-définie positive.*

Annexe E

Anamorphoses

Une opération importante consiste à transformer la distribution statistique d'un ensemble de valeurs en une autre distribution, généralement la distribution gaussienne. Différentes fonctions peuvent être utilisées afin de se rapprocher le plus possible de la loi normale (revue dans Hoyle 1973, Sokal & Rohlf 1995, pp. 413-422). Les fonctions les plus utilisées sont le logarithme et celles définies dans le cadre de la méthode de Box-Cox (*cf.* Howarth & Earle 1979, Stoline 1991, Joseph & Bhaumik 1997).

La géostatistique emploie également des fonctions d'anamorphose d'une distribution vers une autre. En toute généralité, le principe de l'anamorphose consiste à trouver une fonction $\phi(\cdot)$ transformant les valeurs prises par une variable aléatoire originelle Z en valeurs distribuées selon la distribution d'une autre variable Y. En particulier, l'anamorphose gaussienne correspond au cas où la variable Y est distribuée selon la loi normale $\mathcal{N}(0,1)$, les valeurs de Y étant nommées les *normal scores* des valeurs de Z. La définition opératoire de $\phi(\cdot)$ peut s'effectuer au moyen de polynômes d'Hermite ajustés à l'histogramme des fréquences relatives (fonction densité de probabilité empirique) des valeurs de Z (*cf.* Journel & Huijbregts 1978, pp. 472-478) ou à partir des fréquences cumulées (fonction de répartition empirique) des valeurs de Z (*cf.* Journel & Huijbregts 1978, pp. 478-479, Isaaks & Srivastava 1989, pp. 469-471, Deutsch & Journel 1992, pp. 209-213, Goovaerts 1997, pp. 266-271). Nous ne traitons ici que de l'anamorphose de la fonction de répartition.

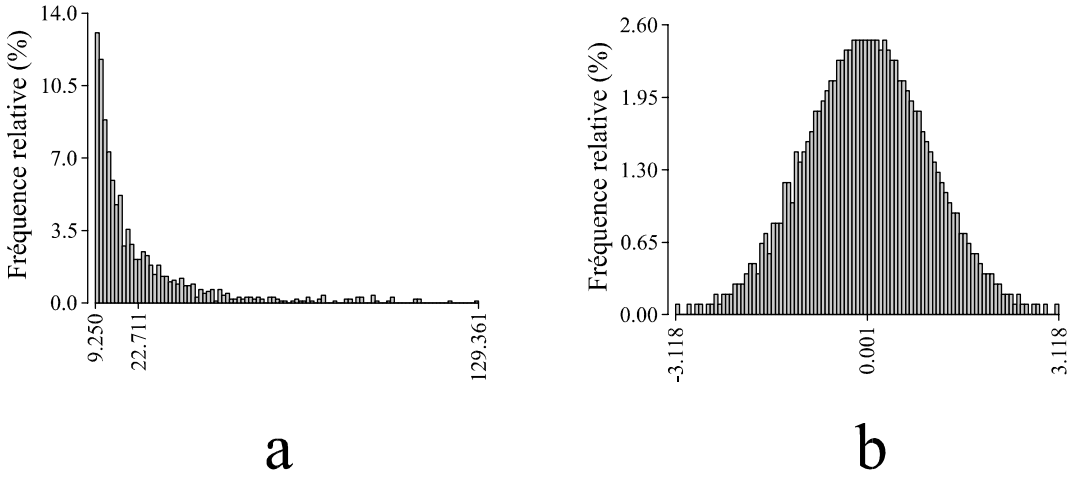
E.1 Exemple d'anamorphose

Considérons une distribution fortement asymétrique de type exponentielle négative, *e.g.* la distribution des diamètres de 1096 arbres dans une forêt tropicale[1] (Fig. E.1.a). Nous pouvons calculer la fonction de répartition empirique de Z (diamètre des arbres) notée $F(z)$ (Fig. E.2.a). A chaque valeur z comprise entre le diamètre minimal ($z_{\min} \simeq 9$ cm) et le diamètre maximal ($z_{\max} \simeq 130$ cm) nous pouvons associer une fréquence cumulée $F(z) = \alpha$. En reportant α sur la fonction de répartition de la variable Y notée $G(y)$ (Fig. E.2) nous obtenons la valeur correspondante $y = \phi(z)$ (Fig. E.2.b). La fonction d'anamorphose $\phi(\cdot)$ est donc définie comme $\phi(z) = G^{-1}[F(z)]$. Si $G(\cdot)$ est la fonction de répartition de la loi normale $\mathcal{N}(0,1)$, en appliquant la transformation $\phi(\cdot)$ aux 1096

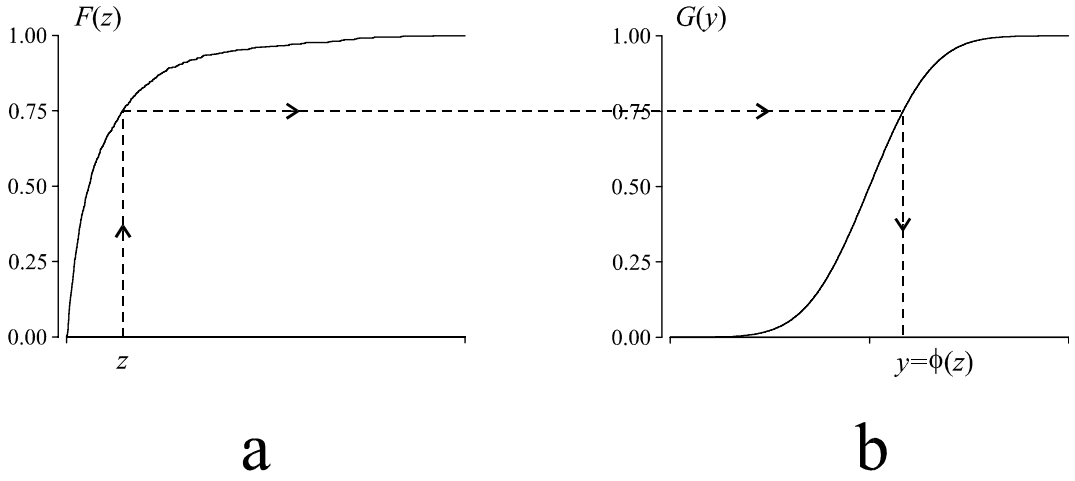
[1] Données communiquées par le Dr. Marie-Agnès Moravie.

diamètres de notre jeu de données nous obtenons leurs *normal scores*, qui s'avèrent effectivement distribués approximativement selon la loi normale $\mathcal{N}(0,1)$ (Fig. E.1.b). La transformation réciproque $z = \phi^{-1}(y)$ s'effectue de la même manière, en partant de la fonction de répartition de Y pour aller vers celle de Z, autrement dit, en définissant $\phi^{-1}(\cdot)$ comme $\phi^{-1}(y) = F^{-1}[G(y)]$.

Cette procédure peut être implémentée en considérant que $F(z)$ et $G(y)$ sont deux fonctions de répartition empiriques, ou en considérant que $G(y)$ est la fonction de répartition théorique de Y. La première implémentation est évidemment plus générale que la seconde puisqu'elle permet de transformer les valeurs de Z vers n'importe quel type de distribution, en particulier la distribution d'un autre jeu de données.



Figure˜E.1: Anamorphose gaussienne appliquée au diamètre de 1096 arbres d'une forêt tropicale. (a) Histogramme des données (diamètre en cm). (b) Histogramme des *normal scores* correspondant aux données.



Figure˜E.2: Principe de la définition de la fonction d'anamorphose $y = \phi(z)$. (a) Fonction de répartition empirique $F(z)$ des données à transformer. (b) Fonction de répartition $G(y)$ de la loi normale $\mathcal{N}(0,1)$, discrétisée en 10^3 sous-intervalles sur l'intervalle $[-4,+4]$.

E.2 Anamorphose gaussienne

Considérons le cas de l'anamorphose gaussienne. Soit $\varphi(\cdot)$ la fonction densité de probabilité de la loi normale $\mathcal{N}(0,1)$:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}} \tag{E.1}$$

La fonction de répartition de la loi normale $\mathcal{N}(0,1)$ s'écrit :

$$\Phi(y) = \int_{-\infty}^{y} \varphi(x)\, dx \tag{E.2}$$

La symétrie de $\varphi(x)$ par rapport à $x = 0$ entraîne la propriété :

$$\Phi(y) + \Phi(-y) = 1 \tag{E.3}$$

La fonction réciproque $\psi(\cdot) = \Phi^{-1}(\cdot)$ est telle que $\Phi(\psi(\alpha)) = \alpha$, avec $\alpha \in]0,1[$. La symétrie de $\varphi(x)$ par rapport à $x = 0$ entraîne la propriété :

$$\psi(\alpha) + \psi(1-\alpha) = 0 \tag{E.4}$$

En considérant la fonction de répartition empirique de Z notée $F(z)$, la fonction de répartition gaussienne $\Phi(y)$ et sa fonction réciproque $\psi(\alpha)$, la transformation de z consiste à déterminer $\alpha = F(z)$, puis à calculer $y = \psi(\alpha)$. La transformation réciproque s'effectue simplement en calculant $\alpha = \Phi(y)$ puis en déterminant z tel que $F(z) = \alpha$. En pratique, il est nécessaire de disposer d'approximations numériques pour les fonctions $\psi(\alpha)$ et $\Phi(y)$.

E.2.1 Approximation de $\psi(\alpha)$

En vertu de la propriété (E.4), il suffit de savoir calculer les valeurs de $\psi(\alpha)$ pour $\alpha \in [0.5, 1[$. Une approximation numérique[2] de $\psi(\alpha)$ relativement précise se calcule selon (Aïvazian *et al.* 1986, p. 364, Abramowitz & Stegun 1972, p. 933) :

$$\psi(\alpha) = t - \left[\sum_{i=0}^{2} c_i t^i\right] \left[1 + \sum_{i=1}^{3} d_i t^i\right]^{-1} + \varepsilon(\alpha) \tag{E.5}$$

avec

$$\begin{aligned} t &= [-2\ln(1-\alpha)]^{1/2} \\ c_0 &= 2.515517,\ c_1 = 0.802853,\ c_2 = 0.010328 \\ d_1 &= 1.432788,\ d_2 = 0.189269,\ d_3 = 0.001308 \end{aligned}$$

et

$$|\varepsilon(\alpha)| < 4.5 \times 10^{-4}$$

Nous avons implémenté cette approximation en Delphi (avatar du Pascal) dans la fonction `F_Gauss_Reciproque`. Le type `EXTENDED` est un type de flottant codé sur 10 octets, comportant 19 ou 20 chiffres significatifs, et couvrant une portée s'étendant de 3.4×10^{-4932} à 1.1×10^{4932} (Borland 1996).

[2] Pour d'autres approximations, voir Abramowitz & Stegun (1972, p. 933), Beasley & Springer (1977) et Aïvazian *et al.* (1986, p. 364).

E.2.2 Approximation de $\Phi(y)$

En vertu de la propriété (E.3), il suffit de savoir calculer les valeurs de $\Phi(y)$ pour $y \in [0, \infty[$. Une approximation numérique[3] de $\Phi(y)$ très précise se calcule selon (Aïvazian *et al.* 1986, p. 363, Abramowitz & Stegun 1972, p. 932) :

$$\Phi(y) = 1 - \varphi(y) \sum_{i=1}^{5} b_i t^i + \varepsilon(y) \tag{E.6}$$

avec

$$\begin{aligned} t &= (1+py)^{-1},\ p = 0.2316419 \\ b_1 &= 0.319381530,\ b_2 = -0.356563782 \\ b_3 &= 1.781477937,\ b_4 = -1.821255978 \\ b_5 &= 1.330274429 \end{aligned}$$

et

$$|\varepsilon(y)| < 7.5 \times 10^{-8}$$

Nous avons implémenté cette approximation en Delphi dans la fonction `F_Gauss`. La valeur de π est approximée par la fonction `PI` de Delphi qui renvoie 3.1415926535897932385 (Borland 1996).

E.2.3 Pratique de l'anamorphose gaussienne

En théorie, pour une valeur $z \in \mathbb{R}$ quelconque nous devrions avoir $\phi^{-1}[\phi(z)] = z$, avec $\phi(z) = \psi[F(z)]$ et $\phi^{-1}(\cdot) = F^{-1}[\Phi(\cdot)]$. En pratique, comme $\psi(\cdot)$ et $\Phi(\cdot)$ sont définies par des approximations numériques, nous obtenons nécessairement $\phi^{-1}[\phi(z)] \neq z$, indépendamment de la définition opératoire de $F(\cdot)$ et de $F^{-1}(\cdot)$. Ce problème peut être évité en renonçant à l'utilisation directe des approximations numériques de $\psi(\cdot)$ et $\Phi(\cdot)$, et en se plaçant dans le cadre général de l'anamorphose d'une fonction de répartition empirique $F(\cdot)$ vers une autre fonction de répartition empirique $G(\cdot)$. Pour ce faire, il suffit de construire $G(\cdot)$ en discrétisant $\Phi(\cdot)$ très finement (*e.g.*, en 10^3 sous-intervalles) sur un intervalle donné (*e.g.*, $[-4, +4]$). En utilisant cette procédure, nous assurons l'égalité numérique de $\phi^{-1}[\phi(z)]$ et de z.

[3] Pour d'autres approximations, voir Abramowitz & Stegun (1972, pp. 932-933), I.D. Hill (1973) et Aïvazian *et al.* (1986, p. 363).

```
{------------------------------------------------------------------------}
{ Approximation de la réciproque de la fonction de répartition de      }
{ la loi normale                                                        }
{------------------------------------------------------------------------}

Function F_Gauss_Reciproque(proba : EXTENDED) : EXTENDED;

    type V_Coeff_T = ARRAY[1..3] of EXTENDED;

    const Coeff_c : V_Coeff_T = (2.515517, 0.802853, 0.010328);
          Coeff_d : V_Coeff_T = (1.432788, 0.189269, 0.001308);

    var num,deno,t,t2,t3 : EXTENDED;
                     Sec : BOOLEAN;

begin
    if (proba<0.5) then begin
        Sec:=TRUE;
        proba:=1.0-proba;
    end else
        Sec:=FALSE;

    t:=sqrt(-2.0*ln(1.0-proba));
    t2:=t*t;
    t3:=t2*t;
    num:=Coeff_c[1]+Coeff_c[2]*t+Coeff_c[3]*t2;
    deno:=1+Coeff_d[1]*t+Coeff_d[2]*t2+Coeff_d[3]*t3;

    if Sec then
        result:=-t+num/deno
    else
        result:=t-num/deno;
end;
```

```
{------------------------------------------------------------------------}
{ Approximation de la fonction de répartition de la loi normale          }
{------------------------------------------------------------------------}

Function F_Gauss(U : EXTENDED) : EXTENDED;

    type V_Coeff_T = ARRAY[1..5] of EXTENDED;

    const   Coeff_P : EXTENDED = 0.2316419;
            Coeff_B : V_Coeff_T = (0.319381530, -0.356563782, 1.781477937,
                                   -1.821255978, 1.330274429);

    var auxPI, Sum, t, t2, t3, t4, t5 : EXTENDED;
                              Sec : BOOLEAN;

begin
    if (U<0.0) then begin
        Sec:=TRUE;
        U:=-U;
    end else
        Sec:=FALSE;

    auxPI:=1.0/sqrt(2.0*PI);
    t:=1.0/(1.0+Coeff_P*U);
    t2:=t*t;
    t3:=t2*t;
    t4:=t3*t;
    t5:=t4*t;
    Sum=Coeff_B[1]*t+Coeff_B[2]*t2+Coeff_B[3]*t3+Coeff_B[4]*t4+Coeff_B[5]*t5;

    if Sec then
        result:=auxPI*exp(-U*U/2.0)*Sum
    else
        result:=1.0-auxPI*exp(-U*U/2.0)*Sum;
end;
```

Annexe F

Modèles de variogrammes

La structure d'autocorrélation spatiale des FAST-2 peut être décrite par au moins six modèles de variogrammes à seuil : périodique, gaussien, cubique, pentasphérique, sphérique et exponentiel (Jian *et al.* 1996). En ce qui concerne les modèles dont le comportement à l'origine est linéaire, les modèles les plus largement employés sont les modèles sphérique et exponentiel, le modèle pentasphérique étant rarement utilisé (*e.g.*, von Steiger *et al.* 1996). Parmi les modèles dont le comportement à l'origine est parabolique, le modèle gaussien est plus souvent cité que le modèle cubique, le modèle périodique n'étant pratiquement jamais utilisé. Il existe d'autres modèles, notamment le modèle circulaire (Dalenius *et al.* 1960, McBratney & Webster 1986) rarement utilisé (*e.g.*, Atkinson 1996), le modèle de Whittle (Whittle 1954, McBratney & Webster 1986) rarement utilisé (*e.g.*, Oliver *et al.* 1998), ainsi que le modèle quadratique rationnel (Cressie 1991, p. 61), également rarement utilisé (*e.g.*, Chadœuf *et al.* 1998).

Les modèles s'écrivent comme des fonctions $\gamma(h, \theta)$ avec $\theta = (c_0, c, a)^{\mathrm{T}}$. Pour alléger la présentation, nous ne précisons plus par la suite que $\gamma(0) = 0$ quel que soit le modèle. Nous donnons ici les expressions des modèles $\gamma(h)$, de leurs intégrales $\gamma_i(h)$, ainsi que de leurs dérivées partielles par rapport aux paramètres[1].

F.1 Modèle exponentiel

Le modèle exponentiel et ses dérivées partielles s'écrivent :

$$\begin{aligned} \gamma(h) &= c_0 + c\left(1 - e^{-3\frac{h}{a}}\right) \\ \frac{\partial\gamma}{\partial c_0} &= 1 \\ \frac{\partial\gamma}{\partial c} &= 1 - e^{-3\frac{h}{a}} \\ \frac{\partial\gamma}{\partial a} &= -\frac{3ch}{a^2}e^{-3\frac{h}{a}} \end{aligned} \tag{F.1}$$

[1] Les dérivées partielles sont utilisées pour l'ajustement automatique par la méthode de Levenberg-Marquardt (Press *et al.* 1989).

L'intégrale du modèle exponentiel s'écrit :

$$\gamma_i(h) = c_0 h + c\left(h + \frac{a}{3}e^{-3\frac{h}{a}}\right) + k \tag{F.2}$$

avec $k = -\frac{ca}{3}$ une constante d'intégration assurant $\gamma_i(0) = 0$. Ses dérivées partielles s'écrivent :

$$\begin{aligned}
\frac{\partial \gamma_i}{\partial c_0} &= h \\
\frac{\partial \gamma_i}{\partial c} &= h + \frac{a}{3}\left(e^{-3\frac{h}{a}} - 1\right) \\
\frac{\partial \gamma_i}{\partial a} &= c\left[\left(\frac{1}{3} + \frac{h}{a}\right)e^{-3\frac{h}{a}} - \frac{1}{3}\right]
\end{aligned}$$

F.2 Modèle pentasphérique

Le modèle pentasphérique s'écrit :

$$\gamma(h) = \begin{cases} c_0 + c\left[\frac{15}{8}\frac{h}{a} - \frac{5}{4}\left(\frac{h}{a}\right)^3 + \frac{3}{8}\left(\frac{h}{a}\right)^5\right] & \text{si } h < a \\ c_0 + c & \text{si } h \geq a \end{cases} \tag{F.3}$$

Pour $h < a$, ses dérivées partielles s'écrivent :

$$\begin{aligned}
\frac{\partial \gamma}{\partial c_0} &= 1 \\
\frac{\partial \gamma}{\partial c} &= \frac{15}{8}\frac{h}{a} - \frac{5}{4}\left(\frac{h}{a}\right)^3 + \frac{3}{8}\left(\frac{h}{a}\right)^5 \\
\frac{\partial \gamma}{\partial a} &= \frac{c}{a}\left[-\frac{15}{8}\frac{h}{a} + \frac{15}{4}\left(\frac{h}{a}\right)^3 - \frac{15}{8}\left(\frac{h}{a}\right)^5\right]
\end{aligned}$$

et pour $h \geq a$, $\frac{\partial \gamma}{\partial c_0} = \frac{\partial \gamma}{\partial c} = 1$ et $\frac{\partial \gamma}{\partial a} = 0$. L'intégrale du modèle pentasphérique s'écrit :

$$\gamma_i(h) = \begin{cases} c_0 h + ch\left[\frac{15}{16}\frac{h}{a} - \frac{5}{16}\left(\frac{h}{a}\right)^3 + \frac{1}{16}\left(\frac{h}{a}\right)^5\right] & \text{si } h < a \\ c_0 h + ch + k & \text{si } h \geq a \end{cases} \tag{F.4}$$

avec $k = -\frac{5}{16}ca$ une constante d'intégration assurant la continuité en $h = a$. Pour $h < a$, ses dérivées partielles s'écrivent :

$$\begin{aligned}
\frac{\partial \gamma_i}{\partial c_0} &= h \\
\frac{\partial \gamma_i}{\partial c} &= h\left[\frac{15}{16}\frac{h}{a} - \frac{5}{16}\left(\frac{h}{a}\right)^3 + \frac{1}{16}\left(\frac{h}{a}\right)^5\right] \\
\frac{\partial \gamma_i}{\partial a} &= c\left[-\frac{15}{16}\left(\frac{h}{a}\right)^2 + \frac{15}{16}\left(\frac{h}{a}\right)^4 - \frac{5}{16}\left(\frac{h}{a}\right)^6\right]
\end{aligned}$$

et pour $h \geq a$, $\frac{\partial \gamma_i}{\partial c_0} = h$, $\frac{\partial \gamma_i}{\partial c} = h - \frac{5}{16}a$ et $\frac{\partial \gamma_i}{\partial a} = -\frac{5}{16}c$.

F.3 Modèle sphérique

Le modèle sphérique s'écrit :

$$\gamma(h) = \begin{cases} c_0 + c\left[\frac{3}{2}\frac{h}{a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right] & \text{si } h < a \\ c_0 + c & \text{si } h \geq a \end{cases} \tag{F.5}$$

Pour $h < a$, ses dérivées partielles s'écrivent :

$$\begin{aligned} \frac{\partial \gamma}{\partial c_0} &= 1 \\ \frac{\partial \gamma}{\partial c} &= \frac{3}{2}\frac{h}{a} - \frac{1}{2}\left(\frac{h}{a}\right)^3 \\ \frac{\partial \gamma}{\partial a} &= \frac{c}{a}\left[-\frac{3}{2}\frac{h}{a} + \frac{3}{2}\left(\frac{h}{a}\right)^3\right] \end{aligned}$$

et pour $h \geq a$, $\frac{\partial \gamma}{\partial c_0} = \frac{\partial \gamma}{\partial c} = 1$ et $\frac{\partial \gamma}{\partial a} = 0$. L'intégrale du modèle sphérique s'écrit :

$$\gamma_i(h) = \begin{cases} c_0 h + ch\left[\frac{3}{4}\frac{h}{a} - \frac{1}{8}\left(\frac{h}{a}\right)^3\right] & \text{si } h < a \\ c_0 h + ch + k & \text{si } h \geq a \end{cases} \tag{F.6}$$

avec $k = -\frac{3}{8}ac$ une constante d'intégration assurant la continuité en $h = a$. Pour $h < a$, ses dérivées partielles s'écrivent :

$$\begin{aligned} \frac{\partial \gamma_i}{\partial c_0} &= h \\ \frac{\partial \gamma_i}{\partial c} &= h\left[\frac{3}{4}\frac{h}{a} - \frac{1}{8}\left(\frac{h}{a}\right)^3\right] \\ \frac{\partial \gamma_i}{\partial a} &= c\left[-\frac{3}{4}\left(\frac{h}{a}\right)^2 + \frac{3}{8}\left(\frac{h}{a}\right)^4\right] \end{aligned}$$

et pour $h \geq a$, $\frac{\partial \gamma_i}{\partial c_0} = h$, $\frac{\partial \gamma_i}{\partial c} = h - \frac{3}{8}a$ et $\frac{\partial \gamma_i}{\partial a} = -\frac{3}{8}c$.

F.4 Modèle cubique

Le modèle cubique s'écrit :

$$\gamma(h) = \begin{cases} c_0 + c\left[7\left(\frac{h}{a}\right)^2 - \frac{35}{4}\left(\frac{h}{a}\right)^3 + \frac{7}{2}\left(\frac{h}{a}\right)^5 - \frac{3}{4}\left(\frac{h}{a}\right)^7\right] & \text{si } h < a \\ c_0 + c & \text{si } h \geq a \end{cases} \tag{F.7}$$

Pour $h < a$, ses dérivées partielles s'écrivent :

$$\begin{aligned} \frac{\partial \gamma}{\partial c_0} &= 1 \\ \frac{\partial \gamma}{\partial c} &= 7\left(\frac{h}{a}\right)^2 - \frac{35}{4}\left(\frac{h}{a}\right)^3 + \frac{7}{2}\left(\frac{h}{a}\right)^5 - \frac{3}{4}\left(\frac{h}{a}\right) \\ \frac{\partial \gamma}{\partial a} &= \frac{c}{a}\left[-14\left(\frac{h}{a}\right)^2 + \frac{105}{4}\left(\frac{h}{a}\right)^3 - \frac{35}{2}\left(\frac{h}{a}\right)^5 + \frac{21}{4}\left(\frac{h}{a}\right)^7\right] \end{aligned}$$

et pour $h \geq a$, $\frac{\partial \gamma}{\partial c_0} = \frac{\partial \gamma}{\partial c} = 1$ et $\frac{\partial \gamma}{\partial a} = 0$. L'intégrale du modèle cubique s'écrit :

$$\gamma_i(h) = \begin{cases} c_0 h + ch\left[\frac{7}{3}\left(\frac{h}{a}\right)^2 - \frac{35}{16}\left(\frac{h}{a}\right)^3 + \frac{7}{12}\left(\frac{h}{a}\right)^5 - \frac{3}{32}\left(\frac{h}{a}\right)^7\right] & \text{si } h < a \\ c_0 h + ch + k & \text{si } h \geq a \end{cases} \tag{F.8}$$

avec $k = -\frac{35}{96}ca$ une constante d'intégration assurant la continuité en $h = a$. Pour $h < a$, ses dérivées partielles s'écrivent :

$$\begin{aligned} \frac{\partial \gamma_i}{\partial c_0} &= h \\ \frac{\partial \gamma_i}{\partial c} &= h\left[\frac{7}{3}\left(\frac{h}{a}\right)^2 - \frac{35}{16}\left(\frac{h}{a}\right)^3 + \frac{7}{12}\left(\frac{h}{a}\right)^5 - \frac{3}{32}\left(\frac{h}{a}\right)^7\right] \\ \frac{\partial \gamma_i}{\partial a} &= c\left[-\frac{14}{3}\left(\frac{h}{a}\right)^3 + \frac{105}{16}\left(\frac{h}{a}\right)^4 - \frac{35}{12}\left(\frac{h}{a}\right)^6 + \frac{21}{32}\left(\frac{h}{a}\right)^8\right] \end{aligned}$$

et pour $h \geq a$, $\frac{\partial \gamma_i}{\partial c_0} = h$, $\frac{\partial \gamma_i}{\partial c} = h - \frac{35}{96}a$ et $\frac{\partial \gamma_i}{\partial a} = -\frac{35}{96}c$.

F.5 Modèle gaussien

Le modèle gaussien et ses dérivées partielles s'écrivent :

$$\begin{aligned} \gamma(h) &= c_0 + c\left[1 - e^{-3\left(\frac{h}{a}\right)^2}\right] \\ \frac{\partial \gamma}{\partial c_0} &= 1 \\ \frac{\partial \gamma}{\partial c} &= 1 - e^{-3\left(\frac{h}{a}\right)^2} \\ \frac{\partial \gamma}{\partial a} &= -\frac{6c}{a}\left(\frac{h}{a}\right)^2 e^{-3\left(\frac{h}{a}\right)^2} \end{aligned} \tag{F.9}$$

L'intégrale du modèle gaussien et ses dérivées partielles s'écrivent :

$$\begin{aligned} \gamma_i(h) &= c_0 h + c\left[h - \frac{a}{6}\sqrt{\pi}\sqrt{3}\,\mathrm{erf}\left(\frac{\sqrt{3}}{a}h\right)\right] \\ \frac{\partial \gamma_i}{\partial c_0} &= h \\ \frac{\partial \gamma_i}{\partial c} &= h - \frac{a}{6}\sqrt{\pi}\sqrt{3}\,\mathrm{erf}\left(\frac{\sqrt{3}}{a}h\right) \\ \frac{\partial \gamma_i}{\partial a} &= c\left[\frac{h}{a}e^{-3\left(\frac{h}{a}\right)^2} - \frac{1}{6}\sqrt{\pi}\sqrt{3}\,\mathrm{erf}\left(\frac{\sqrt{3}}{a}h\right)\right] \end{aligned} \tag{F.10}$$

avec la fonction d'erreur :

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt \tag{F.11}$$

F.6 Modèle périodique

Le modèle périodique et ses dérivées partielles s'écrivent :

$$\begin{aligned}
\gamma(h) &= c_0 + c\left[1 - \frac{a}{\beta h}\sin\left(\beta\frac{h}{a}\right)\right] \\
\frac{\partial\gamma}{\partial c_0} &= 1 \\
\frac{\partial\gamma}{\partial c} &= 1 - \frac{a}{\beta h}\sin\left(\beta\frac{h}{a}\right) \\
\frac{\partial\gamma}{\partial a} &= c\left[\frac{1}{a}\cos\left(\beta\frac{h}{a}\right) - \frac{1}{\beta h}\sin\left(\beta\frac{h}{a}\right)\right]
\end{aligned} \tag{F.12}$$

avec $\beta = 4.4934$. L'intégrale du modèle périodique et ses dérivées partielles s'écrivent :

$$\begin{aligned}
\gamma_i(h) &= c_0 h + c\left[h - \frac{a}{\beta}\,\mathrm{Si}\left(\beta\frac{h}{a}\right)\right] \\
\frac{\partial\gamma_i}{\partial c_0} &= h \\
\frac{\partial\gamma_i}{\partial c} &= h - \frac{a}{\beta}\,\mathrm{Si}\left(\beta\frac{h}{a}\right) \\
\frac{\partial\gamma_i}{\partial a} &= \frac{c}{\beta}\left[\sin\left(\beta\frac{h}{a}\right) - \mathrm{Si}\left(\beta\frac{h}{a}\right)\right]
\end{aligned} \tag{F.13}$$

avec le sinus intégral :

$$\mathrm{Si}(x) = \int_0^x \frac{\sin t}{t}dt \tag{F.14}$$

Annexe G

Exemples d'applications

G.1 Analyse de l'autocorrélation spatiale

G.1.1 Choix de la fonction

Dans la Section (3.3) nous avons mentionné plusieurs types de fonctions d'autocorrélation, essentiellement le c de Geary ou le variogramme, le I de Moran ou la fonction de covariance, et les fonctions de covariance ou de corrélation dites *non ergodiques*.

Toutes ces fonctions ont été utilisées en pratique, le I de Moran étant d'utilisation courante en biologie évolutive (*e.g.*, Sokal & Oden 1978a, 1978b, Sokal & Wartenberg 1983, Sokal *et al.* 1989a, 1989b, Epperson 1990, 1995, Epperson & Li 1996, Hossaert-McKey *et al.* 1996, Doligez & Joly 1997, Mahy & Nève 1997, Caujapé-Castells & Pedrola-Monfort 1997, Petit *et al.* 1997, Sokal *et al.* 1997, Sokal & Thomson 1998) et les fonctions géostatistiques (variogramme, covariance non ergodique) plus particulièrement utilisées en écologie des populations. Par exemple, en utilisant le variogramme :

- Pont (1986) tente de caractériser la distribution spatiale de différents stades du crustacé *Acanthocyclops robustus* (Cyclopidae) dans une rizière de Camargue,
- Schotzko & O'Keeffe (1989) décrivent la distribution spatiale de la punaise *Lygus hesperus* (Miridae) dans des champs de lentilles du nord de l'Idaho et de l'Est de l'état de Washington (USA),
- Sarnelle *et al.* (1993) démontrent l'effet des gastéropodes herbivores *Physella* sp. sur la structure spatiale de la biomasse algale benthique,
- Larkin *et al.* (1995) décrivent la structure spatiale d'une épidémie causée par *Phytophthora capsici* dans les champs d'une pipéracée,
- Cardina *et al.* (1996) montrent que la banque de graines et les populations de semis des "mauvaises herbes" sont spatialement autocorrélées,
- Delaville *et al.* (1996) et Rossi *et al.* (1996) étudient l'autocorrélation spatiale des populations des nématodes présents dans le sol de champs de canne à sucre de Martinique,
- Cannavacciuolo *et al.* (1998) caractérisent la structure spatiale de la biomasse des juvéniles et des adultes de deux espèces de lombrics dans une prairie de Bretagne,

- Kuuluvainen *et al.* (1998) calculent le variogramme du DBH (*Diameter at Breast Height*) et de la hauteur des arbres d'une forêt naturelle dominée par *Pinus sylvestris*, dans l'est de la Finlande.

A l'aide de la covariance non ergodique :

- Weisz *et al.* (1995a) caractérisent la structure spatiale des différents stades du doryphore (*Leptinotarsa decemlineata*),
- Dandurand *et al.* (1995) étudient la répartion spatiale des zoospores de *Pythium ultimum* sur des racines de petits pois,
- Dandurand *et al.* (1997) étudient la colonisation de la surface des racines de petits pois par la bactérie *Pseudomonas fluorescens*.

L'utilisation privilégiée du I de Moran en biologie évolutive et du variogramme ou de la covariance non ergodique en écologie des populations est pure contingence : toutes les fonctions peuvent être utilisées, indépendamment de la discipline.

G.1.2 Analyse omnidirectionnelle vs. directionnelle

Le calcul de fonctions directionnelles requiert des données suffisamment abondantes. Ainsi, il est recommandé de disposer d'au moins $n = 300$ données afin de calculer des variogrammes directionnels (Oliver *et al.* 1989b).

Néanmoins, avec seulement $n = 100$ données il est possible de calculer une fonction pour quelques directions principales, en utilisant des classes d'angles suffisamment larges. Typiquement, les variogrammes directionnels sont calculés pour quatre directions, à 0°, 45°, 90° et 135° (*e.g.*, Di *et al.* 1989, Höck *et al.* 1993, Cardina *et al.* 1995, Sharov *et al.* 1995, Brooker *et al.* 1995, Larkin *et al.* 1995, Goovaerts 1997, p. 99, Brannan & Hamlett 1998, Cannavacciuolo *et al.* 1998). En revanche, un jeu de données de taille $n = 44$ (Chang *et al.* 1998) ou $n = 48$ (Pont 1986) peut déjà être considéré comme insuffisant pour une analyse omnidirectionnelle, et *a fortiori* pour une analyse directionnelle.

Dans certaines applications, la situation est très favorable :

- Larkin *et al.* (1995) calculent des variogrammes directionnels dans quatre directions à partir de $n = 400$ quadrats, afin d'étudier la structure spatiale d'une épidémie chez une pipéracée,
- Delcourt *et al.* (1996) disposent également de $n = 400$ quadrats, dans le cadre de l'étude de la variabilité spatiale de la concentration en nutriments dans le sol,
- Guertal & Elkins (1996) disposent de $n = 529$ points pour analyser la variation spatiale des radiations photosynthétiquement actives dans une serre.

G.1.3 Test de signification globale

La pratique du test global d'une fonction d'autocorrélation au moyen de la correction de Bonferroni (Section 3.3.4.3) a été diffusée en écologie par Sokal & Thomson (1987), puis Legendre & Troussellier (1988), Legendre & Fortin (1989), et en biologie évolutique par Sokal *et al.* (1987), puis Sokal & Jacquez (1991).

Cette correction est utilisée notamment par :

- Liang *et al.* (1997) dans la quantification de l'autocorrélation spatiale de la densité du lépidoptère *Lambdina fiscellaria lugubrosa* (Geometridae) le long des routes de Colombie britannique,
- Legendre *et al.* (1997) dans la description de la structure spatiale des populations des bivalves *Macomona liliana* et *Austrovenus stutchburyi* en Nouvelle-Zélande,
- par Rossi & Quenehervé (1998) dans l'analyse des relations entre les nématodes et la texture du sol, ou le contenu du sol en carbone et en argile.

La correction de Šidák est utilisée notamment par Doligez & Joly (1997) dans l'étude de la structure spatiale génétique d'une forêt tropicale de *Carapa procera* (Meliaceae) dans le site de Paracou, en Guyane française.

G.2 Krigeage

Le krigeage est recommandé dans des domaines aussi divers que la géographie physique (Oliver *et al.* 1989a, 1989b), la pédologie (Webster & Burgess 1980a, 1980b, Burgess & Webster 1980, Yost *et al.* 1982, Morkoc *et al.* 1987, Webster & Oliver 1989, Webster 1991), l'écologie des sols (Robertson 1987), la biologie des sols (Wallace & Hawkins 1994, Rossi *et al.* 1995), la foresterie (Köhl & Gertner 1997, Gunnarsson *et al.* 1998), l'halieutique (Simard *et al.* 1993), l'entomologie (Liebhold *et al.* 1993, Crist 1998), la biogéographie (Maurer 1994), la phytopathologie (Lecoustre *et al.* 1989) ou encore l'agriculture de précision (Donald 1994, Heisel *et al.* 1996, Delcourt *et al.* 1996).

Le krigeage est de plus en plus utilisé pour cartographier des phénomènes abiotiques tels que :

- les propriétés du sol (Robertson *et al.* 1988),
- la concentration en nutriments dans le sol (Robertson 1987, Jackson & Caldwell 1993b),
- l'eau disponible dans les sols (Dahiya *et al.* 1990, Brooker *et al.* 1995),
- la quantité de carbone stocké dans le sol (Liski & Westman 1997),
- la température de l'air (Hudson & Wackernagel 1994, Holdaway 1996),
- la température de la surface de la mer (Gohin 1987, 1990, Gohin & Langlois 1993),
- la température, la salinité et les concentrations en nutriments de l'eau de mer (Toompuu & Wulff 1996)[1],
- la salinité de l'eau d'une nappe phréatique (Söderström 1992),
- l'évapotranspiration (Ashraf *et al.* 1997),
- les infiltrations de pesticides (Persicani 1995),

[1] Toompuu & Wulff (1996) et Zhou (1998) utilisent en fait une méthode d'interpolation dans le cadre de l'*analyse objective* développée en URSS pour la météorologie par Gandin, équivalente au krigeage (Cressie 1989, 1990), l'analyse objective étant elle-même équivalente à la géostatistique (Chauvet 1994).

- la contamination des sols par les métaux lourds (Okx *et al.* 1993, Tao 1995, von Steiger *et al.* 1996, Söderström & Eriksson 1996, Juang & Lee 1998b, Juang *et al.* 1998),
- le risque de contamination des sols (Goovaerts *et al.* 1997),
- la probabilité de délimitation incorrecte des aires à risque dans un site contaminé (Juang & Lee 1998a),
- la concentration en ozone dans les villes (Sheshinski 1979, Liu & Rossini 1996),
- la probabilité de pollution par le dioxyde de soufre (Soares *et al.* 1993),
- les dépôts de composants acidifiants et/ou de cations basiques dus aux précipitations (Haas 1990, van Leeuwen *et al.* 1996).

Le krigeage est également largement utilisé pour cartographier des phénomènes biotiques tels que :

- les fréquences géniques (Piazza *et al.* 1981),
- la densité ou l'abondance des oiseaux (Maurer 1994, Villard & Maurer 1996), des orignaux (McKenney *et al.* 1998), des poissons (Petitgas 1993, Simard *et al.* 1993, Pelletier & Parma 1994, Maravelias *et al.* 1996), du plancton (Zhou 1998)[1] , des crevettes (Porter *et al.* 1997), des nématodes (Wallace & Hawkins 1994, Rossi *et al.* 1995, 1996), des lombrics (Cannavacciuolo *et al.* 1998), des tiques (Nicholson & Mather 1996), d'un coléoptère carabique (Franceschini *et al.* 1997), des pucerons et de leurs parasitoïdes (Longley *et al.* 1997), des oeufs d'une cochenille (Speight *et al.* 1998), des différents stades d'un coléoptère ravageur des cultures (Weisz *et al.* 1995b), des "mauvaise herbes" (Cardina *et al.* 1996, Johnson *et al.* 1996),
- la biomasse végétale en milieu semi-aride (Phinn *et al.* 1996),
- la biomasse des oeufs d'un lépidoptère ravageur des forêts (Liebhold *et al.* 1991, Gribko *et al.* 1995, Weseloh 1996),
- les limites de l'aire de répartition d'un lépidoptère ravageur des forêts (Sharov *et al.* 1995),
- les probabilités de défioliation (Hohn *et al.* 1993),
- la probabilité de présence de termites (Crist 1998),
- la richesse spécifique de lépidoptères Rhopalocères (Carroll & Pearson 1998a, 1998b),
- les épidémies (Lecoustre *et al.* 1989, Lannou & Savary 1991, Carrat & Valleron 1992, 1993, Nelson *et al.* 1999),
- les risques de cancer (Webster *et al.* 1994, Oliver *et al.* 1998),
- la prévalence d'un protozoaire pathogène (White *et al.* 1998),
- les pertes d'aiguilles et de feuilles des arbres (Köhl & Gertner 1997),
- la qualité des sols (Smith *et al.* 1993, Halvorson *et al.* 1996),
- la productivité des forêts plantées (Höck *et al.* 1993),
- le rendement des cultures (Birrell 1996, Wopereis *et al.* 1996),
- la contamination fécale de l'eau de mer dans le cadre du contrôle microbiologique de la mytiliculture (Beliaeff & Cochard 1995).

G.3 Test de Mantel partiel

Le test de Mantel partiel est largement utilisé en écologie afin de régler le problème de l'autocorrélation spatiale dans le test de la corrélation entre deux variables régionalisées (Section 9.1.7.1). Les exemples qui suivent permettent d'illustrer cette approche.

Afin d'étudier à l'échelle européenne le rôle relatif de l'isolement par la distance et de l'histoire postglaciaire des patterns de variation nucléaire chez *Quercus petraea,* Le Corre *et al.* (1997) testent la corrélation entre les distances génétiques nucléaires établies à partir de site RAPD (*Random Amplified Polymorphic DNA*) ou de sites d'allozymes, et deux variables explicatives constituées par les distances géographiques (G) et les distances calculées à partir du polymorphisme de l'ADN des chloroplastes (DML). En utilisant un test d'association entre matrices (test de Mantel, Section 3.1.4), ces auteurs montrent que les deux variables explicatives G et DML sont corrélées, ce qui peut conduire à une fausse corrélation avec les distances génétiques nucléaires. Le Corre *et al.* (1997) considèrent par la suite uniquement des corrélations partielles et utilisent des tests de Mantel partiels. En fait, la corrélation entre G et DML indique simplement que le polymorphisme de l'ADN des chloroplastes est spatialement autocorrélé. Le problème peut donc être reformulé en termes de corrélation en présence d'autocorrélation spatiale pour le polymorphisme de l'ADN des chloroplastes. Néanmoins, la manipulation de données sous la forme de matrices de distances interdit de recourir au test de corrélation de Pearson car les distances ne sont pas indépendantes (Section 3.1.4).

Despland & Houle (1997) étudient la variation interannuelle de la croissance et de l'effort de reproduction dans une population de *Pinus banksiana* à sa limite d'aire de répartition dans le nord du Québec. En utilisant des tests de Mantel partiels, ces auteurs évaluent en particulier la corrélation entre la croissance radiale et l'effort de reproduction, ainsi que la corrélation entre chacune des variables biologiques considérées et des variables climatiques. Despland & Houle (1997) affirment que le test de Mantel partiel est similaire au test de la corrélation de Pearson partielle mais qu'il tient compte de l'autocorrélation, sans toutefois en faire la démonstration. Dans cette étude, l'autocorrélation est considérée dans un plan formé par une dimension spatiale (hauteur de l'arbre) et par une dimension temporelle (année de formation d'un anneau ou année de pollinisation).

Vieira *et al.* (1998) étudient à l'échelle planétaire la corrélation entre la température minimale et le nombre de copies des rétrotransposons 412 et roo/B104 chez *Drosophila simulans*. Rien n'impose *a priori* de manipuler les données sous la forme de matrices de distances ou de similarités, mais ces auteurs proposent de tenir compte de l'autocorrélation en latitude pour la température minimale et pour le nombre de copies des rétrotransposons en utilisant des tests de Mantel partiels.

G.4 Fonctions croisées

Le variogramme croisé peut être utilisé pour analyser la covariation spatiale :

- d'une même espèce à deux dates différentes ou bien entre deux espèces, *e.g.* les carabes *Dyschirius globosus* et *Pterostichus coerulescens* (*cf.* Rossi *et al.* 1992),
- d'une espèce par rapport à un *pool* d'espèces (Delaville *et al.* 1996, Rossi *et al.*1996),

- de deux écophases d'une même espèce (Cardina *et al.* 1996),
- de deux nutriments du sol (Lavado *et al.* 1996),
- de deux mesures de diversité (Kuuluvainen *et al.* 1998),
- de deux indicatrices concernant deux aspects d'un même phénomène, afin d'approfondir l'analyse de sa structure spatiale (Petitgas & Levenez 1996, Barange & Hampton 1997).

Delaville *et al.* (1996) et Rossi *et al.* (1996) étudient l'interaction entre différentes espèces de nématodes dans un champ de canne à sucre en Martinique. Un variogramme croisé est calculé entre la densité de *Criconemella onoensis* et la somme des densités de trois autres espèces de nématodes, *Helicotylenchus erythrinae*, *Hemicriconemoides cocophillus* et *Pratylenchus zeae*. Dans un des sites d'étude, le variogramme croisé est négatif, ce qui révèle un phénomène de ségrégation spatiale entre *C. onoensis* et le *pool* des trois autres espèces de nématodes. En fait, cette association spatiale négative est parfaitement visible dans les représentations cartographiques des densités.

Cardina *et al.* (1996) étudient la continuité spatiale entre la banque de graines et les populations de semis de *Chenopodium album* et d'un *pool* d'herbes annuelles, dans des champs de soja (*Glycine max*) de l'Ohio (USA), pour les années 1990 à 1993. En utilisant des variogrammes croisés, ces auteurs mettent en évidence une relation spatiale significative pour *C. album* uniquement dans les champs non labourés, pour les années 1990 à 1992, et pour toutes les années sans labour dans le cas des herbes annuelles considérées. Les auteurs calculent également le ρ de Spearman entre la densité de graines et la densité de semis. Pour l'année 1993 et un champ non labouré, les auteurs obtiennent un coefficient $\rho_{obs} = 0.53$ dans le cas de *C. album,* alors que le variogramme croisé ne montre par de dépendance spatiale claire. Cet exemple illustre bien la différence qui existe entre l'*association a-spatiale* (ρ de Spearman) et l'*association spatiale* (variogramme croisé), les deux n'allant pas forcément de pair (Hubert *et al.* 1985). Cardina *et al.* (1996) complètent leur analyse en cartographiant la densité de graines et de semis par krigeage par bloc, puis en comparant visuellement les images obtenues.

Lavado *et al.* (1996) examinent l'impact du pâturage de la pampa d'Argentine sur la variabilité spatiale des nutriments du sol. A l'aide de variogrammes croisés, ces auteurs montrent que le carbone organique et l'azote total sont fortement spatialement corrélés, indépendamment du pâturage, mais que la structure de la corégionalisation est spatialement plus régulière dans le cas des aires pâturées.

Kuuluvainen *et al.* (1998) décrivent la structure spatiale d'une forêt naturelle dominée par le pin *Pinus sylvestris*, située dans l'est de la Finlande. Des variogrammes croisés calculés entre la diversité des hauteurs des arbres et la diversité spécifique montrent que l'hétérogénéité verticale augmente avec la diversité des essences. En outre, la diversité est également associée à l'aire couverte par la base des arbres dans chaque quadrat, ce qui suggère que la diversité spécifique augmente avec la biomasse des arbres.

Le variogramme croisé peut également être utilisé afin d'affiner l'étude de la structure spatiale d'une seule VR en la décomposant en plusieurs indicatrices, pour différentes valeurs seuil. Dans ce contexte particulier, Petitgas & Levenez (1996) étudient la localisation des hautes densités au sein du domaine occupé par des bancs de poissons. Ces auteurs définissent une indicatrice I_a pour les petits bancs de poissons (SSM) et/ou les bancs compacts (SCO), et une indicatrice I_b d'unités d'échantillonnage riches (densité supérieure à

100 t par mille nautique carré) qui contiennent des SSM et/ou des SCO. En calculant un variogramme croisé entre I_a et I_b , Petitgas & Levenez (1996) montrent que les fortes densités associées à la présence de bancs ne se situent pas, en moyenne, au centre de l'aire de présence des bancs de poissons. Avec le même type d'approche, Barange & Hampton (1997) définissent des indicatrices pour la densité d'anchois (*Engraulis capensis*) et de sardines (*Sardinops sagax*), et calculent des variogrammes croisés entre indicatrices d'une même variable afin de même en évidence un phénomène de transition entre les régions de fortes et de faibles densités.

Bien que le variogramme croisé semble constituer la fonction la plus utilisée en écologie, la covariance croisée est potentiellement plus riche sur le plan descriptif puisqu'elle permet de déceler un éventuel effet retard (Lecoustre & de Reffye 1986). La covariance croisée doit par conséquent être calculée à la fois pour les directions $\mathbf{h}$ et $-\mathbf{h}$ afin d'obtenir une image complète de la covariation spatiale (Rossi *et al.* 1992). En fait, de même que dans le cas de l'analyse de l'autocorrélation spatiale, l'analyse de la covariation spatiale devrait être menée à l'aide de différentes fonctions croisées, notamment le variogramme croisé et la covariance croisée non ergodique (Rossi *et al.* 1992). En pratique, ces recommandations ne sont pas toujours suivies. Par exemple, Marshall *et al.* (1998) quantifient la dépendance spatiale entre les trois espèces de nématodes *Longidorus elongatus*, *L. goodeyi* et *Rotylenchus goodeyi*, en calculant la covariance croisée non ergodique, standardisée sous la forme d'une corrélation croisée non ergodique (Section 3.3.3, p. 50), mais uniquement dans deux directions $\mathbf{h}$ orthogonales. Griffiths *et al.* (1996) étudient la corrélation croisée entre les ectomycorrhizes *Gautieria* et *Hysterangium*, mais sans préciser quelle est la fonction utilisée, les directions considérées, et sans qu'aucun résultat ne soit figuré !

G.5 Comparaison visuelle

Bien que d'un intérêt primordial dans l'étude de l'association, les méthodes statistiques de comparaison spatiale de deux cartes quantitatives sont extrêmement peu développées (Minns *et al.* 1996). L'approche la plus élémentaire — mais peut-être encore la plus satisfaisante à l'heure actuelle — est la comparaison visuelle des cartes. Ainsi, en écologie végétale, Gittins (1968) se contente d'apprécier visuellement les similarités entre plusieurs surfaces de tendance et pose le problème de leur comparaison objective. Trente ans plus tard, ce problème ne semble toujours pas avoir reçu de solution entièrement satisfaisante. En effet, Hill *et al.* (1998) exploitent la capacité de l'Homme à appréhender les relations spatiales visuellement lors de la validation de leur modèle d'invasion du nord-ouest de la Méditerranée par l'algue *Caulerpa taxifolia*, Cardina *et al.* (1996) apprécient visuellement la correspondance entre les images en niveaux de gris des densités de graines et de semis de *Chenopodium album*. Il en est de même pour Birrell *et al.* (1996) en ce qui concerne le rendement de cultures cartographié selon différentes techniques, pour Gerhards *et al.* (1997) étudiant la stabilité de la répartition spatiale de "mauvaises herbes" au cours de quatre années consécutives, et pour Cannavacciuolo *et al.* (1998) mettant en relation la biomasse d'adultes de lombrics et l'hydromorphie du sol.

Références bibliographiques

1. **Ababou R., A.C. Bagtzoglou & E.F. Wood (1994)**. On the condition number of covariance matrices in kriging, estimation, and simulation of random fields. *Mathematical Geology*, **26** : 99-133.
2. **Abramowitz M. & I.A. Stegun (1972)**. Handbook of mathematical functions. Dover Publications, New York, USA, 1046 p.
3. **Addicott J.F., J.M. Aho, M.F. Antolin, D.K. Padilla, J.S. Richardson & D.A. Soluk (1987)**. Ecological neighborhoods: scaling environmental patterns. *Oikos*, **49** : 340-346.
4. **Adjeroud M. (1997)**. Factors influencing spatial patterns on coral reefs around Moorea, French Polynesia. *Marine Ecology - Progress Series*, **159** : 105-119.
5. **Agresti A. (1990)**. Categorical data analysis. Wiley, New York, USA, 558 p.
6. **Ahn C.W., M.F. Baumgardner & L.L. Biehl (1999)**. Delineation of soil variability using geostatistics and fuzzy clustering analyses of hyperspectral data. *Soil Science Society of America Journal*, **63** : 142-150.
7. **Aho A. & J. Ullman (1993)**. Concepts fondamentaux de l'informatique. Dunod, Paris, France, 856 p.
8. **Ahuja R.K., K. Melhorn, J.B. Orlin & R.E. Tarjan (1990)**. Faster algorithms for the shortest path problem. *Journal of the Association for Computing Machinery*, **37** : 213-223.
9. **Aïvazian S., I. Enukov & L. Mechalkine (1986)**. Eléments de modélisation et traitement primaire des données. Mir, Moscou, URSS, 389 p.
10. **Akl S.G. & G.T. Toussaint (1978)**. A fast convex hull algorithm. *Information Processing Letters*, **7** : 219-222.
11. **Akl S.G. (1979)**. Two remarks on a convex hull algorithm. *Information Processing Letters*, **8** : 108-109.
12. **Alabert F. (1987)**. The practice of fast conditional simulations through the LU decomposition of the covariance matrix. *Mathematical Geology*, **19** : 369-386.
13. **Aldrich J. (1995)**. Correlation genuine and spurious in Pearson and Yule. *Statistical Science*, **10** : 364-376.
14. **Alekseev V.B. (1995)**. Graph, planar. *In* : Hazewinkel M. (Ed.), *Encyclopedia of mathematics*, vol. 2, pp. 878-879. Kluwer Academic Publishers, Dordrecht, The Netherlands.
15. **Almeida A.S. & A.G. Journel (1994)**. Joint simulation of multiple variables with a Markov-type coregionalization model. *Mathematical Geology*, **26** : 565-588.
16. **Alt F.B. (1982)**. Bonferroni inequalities and intervals. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 1, pp. 294-300. Wiley, New York, USA.
17. **Altman N.S. (1988)**. Bit-wise behavior of random generators. *SIAM Journal on Scientific and Statistical Computing*, **9** : 941-949.
18. **Anderson A.R.A., I.M. Young, B.D. Sleeman, B.S. Griffiths & W.M. Robertson (1997)**. Nematode movement along a chemical gradient in a structurally heterogeneous environment. 1. Experiment. *Fundamental and Applied Nematology*, **20** : 157-163.
19. **Anderson K.P. (1976)**. Simple algorithm for positioning a point close to a boundary. *Mathematical Geology*, **8** : 105-106.

20. **Anderson K.R. (1978)**. A reevaluation of an efficient algorithm for determining the convex hull of a finite planar set. *Information Processing Letters*, **7** : 53-57.
21. **Andrew A.M. (1979)**. Another efficient algorithm for convex hulls in two dimensions. *Information Processing Letters*, **9** : 216-219.
22. **Anh V., H. Duc & I. Shannon (1997)**. Spatial variability of Sydney air quality by cumulative semivariogram. *Atmospheric Environment*, **31** : 4073-4080.
23. **Anselin L. (1988)**. Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity. *Geographical Analysis*, **20** : 1-17.
24. **Anselin L. (1995)**. Local indicators of spatial association. *Geographical Analysis*, **27** : 93-115.
25. **Arbia G. (1993)**. The use of GIS in spatial statistical surveys. *International Statistical Review*, **61** : 339-359.
26. **Armstrong M. (1984a)**. Problems with universal kriging. *Mathematical Geology*, **16** : 101-108.
27. **Armstrong M. (1984b)**. Common problems seen in variograms. *Mathematical Geology*, **16** : 305-313.
28. **Armstrong M. (1992)**. Positive definiteness is not enough. *Mathematical Geology*, **24** : 135-143.
29. **Armstrong M. & P.H. Diamond (1984)**. Testing variograms for positive-definiteness. *Mathematical Geology*, **16** : 407-421.
30. **Armstrong M. & D. Jabin (1981)**. Variogram models must be positive-definite. *Mathematical Geology*, **13** : 455-459.
31. **Arnold A. & I. Guessarian (1993)**. Mathématiques pour l'informatique. Masson, Paris, France, 349 p.
32. **Ashraf M., J.C. Loftis & K.G. Hubbard (1997)**. Application of geostatistics to evaluate partial weather station networks. *Agricultural and Forest Meteorology*, **84** : 255-271.
33. **Asli M. & D. Marcotte (1995)**. Comparison of approaches to spatial estimation in a bivariate context. *Mathematical Geology*, **27** : 641-658.
34. **Aspie D. & R.J. Barnes (1990)**. Infill-sampling design and the cost of classification errors. *Mathematical Geology*, **22** : 915-932.
35. **Atkinson A.C. (1980)**. Test of pseudo-random numbers. *Applied Statistics*, **29** : 164-171.
36. **Atkinson A.C. (1982)**. Developments in the design of experiments. *International Statistical Review*, **50** : 161-177.
37. **Atkinson P.M. (1996)**. Optimal sampling strategies for raster-based geographical information systems. *Global Ecology and Biogeography Letters*, **5** : 271-280.
38. **Atlan H. (1972)**. L'organisation biologique et la théorie de l'information. Hermann, Paris, France, 300 p.
39. **Aubry P. (1996a)**. Géostatistique pour l'analyse et la modélisation des structures spatiales en écologie et en biologie des populations. DEA "Analyse et modélisation des systèmes biologiques", Université Claude Bernard - Lyon 1, 40 p.
40. **Aubry P. (1996b)**. Estimation géostatistique en écologie et en biologie des populations. DEA "Analyse et modélisation des systèmes biologiques", Université Claude Bernard - Lyon 1, 68 p.
41. **Aubry P. & D. Debouzie [1999a] (2000)**. Geostatistical estimation variance for the spatial mean in two-dimensional systematic sampling. *Ecology,* 81 : 543-553.

42. **Aubry P. & D. Debouzie [1999b]**. Estimation of the mean from a two-dimensional sample: the geostatistical model-based approach. *Ecology*, sous presse.

43. **Aubry P. & C. Egretaud (1994)**. Classification non dirigée optimale d'une image monocanal. *International Journal of Remote Sensing*, **15** : 3839-3843.
44. **Avnir D., O. Biham, D. Lidar & O. Malcai (1998)**. Is the geometry of nature fractal? *Science*, **279** : 39-40.
45. **Baczkowski A.J. & K.V. Mardia (1990)**. Prediction based upon maximizing squared correlation for stationary processes and simple kriging. *Journal of Applied Statistics*, **17** : 159-164.
46. **Baddeley A. (1982)**. Stochastic geometry: an introduction and reading-list. *International Statistical Review*, **50** : 179-193.
47. **Bailey R.A. (1983)**. Restricted randomization. *Biometrika*, **70** : 183-198.
48. **Bailey R.A. (1985)**. Restricted randomization versus blocking. *International Statistical Review*, **53** : 171-182.
49. **Bailey R.A. (1986)**. Randomization, constrained. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 7, pp. 524-530. Wiley, New York, USA.
50. **Baker F.B. (1997)**. A note on the proper use of the Numerical Recipes RAN1 random number generator. *Computational Statistics & Data Analysis*, **25** : 237-239.
51. **Barange M. & I. Hampton (1997)**. Spatial structure of co-occurring anchovy and sardine populations from acoustic data: implications for survey design. *Fisheries Oceanography*, **6** : 94-108.
52. **Barbujani G., N.L. Oden & R.R. Sokal (1989)**. Detecting regions of abrupt change in maps of biological variables. *Systematic Zoology*, **38** : 376-389.
53. **Bard Y. (1970)**. Comparison of gradient methods for the solution of nonlinear parameter estimation problems. *SIAM Journal of Numerical Analysis*, **7** : 157-186.
54. **Bárdossy A. (1988)**. Notes on the robustness of the kriging system. *Mathematical Geology*, **20** : 189-203.
55. **Barnes R.J. (1988)**. Bounding the required sample size for geologic site characterization. *Mathematical Geology*, **20** : 477-490.
56. **Barnes R.J. (1991)**. The variogram sill and the sample variance. *Mathematical Geology*, **23** : 673-678.
57. **Barnes R.J. & T.B. Johnson (1984)**. Positive kriging. *In* : Verly G., M. David, A.G. Journel & A. Marechal (Eds.) *NATO-ASI geostatistics for natural resources characterization*, pp. 231-244. D. Reidel Publishing Company, Dordrecht, The Netherlands.
58. **Barnes R.J. & A.G. Watson (1992)**. Efficient updating of kriging estimates and variances. *Mathematical Geology*, **24** : 129-133.
59. **Barnes R.J. & K. You (1992)**. Adding bounds to kriging. *Mathematical Geology*, **24** : 171-176.
60. **Barry R.P. & J.M. Ver Hoef (1996)**. Blackbox kriging: spatial prediction without specifying variogram models. *Journal of Agricultural, Biological, and Environmental Statistics*, **1** : 297-322.
61. **Bates D.M. & D.G. Watts (1988)**. Nonlinear regression analysis and its applications. Wiley, New York, USA, 365 p.
62. **Baudry J. & F. Baudry-Burel (1982)**. La mesure de la diversité spatiale. Relations avec la diversité spécifique. Utilisation dans les évaluations d'impact. *Acta Oecologica - Oecologia Generalis*, **3** : 177-190.

63. **Bazuhair A.S.A. & Z. Şen (1994)**. Cumulative semivariogram models of trace elements from springs in Saudi Arabia. *Nordic Hydrology*, **25** : 345-358.
64. **Beasley J.D. & S.G. Springer (1977)**. Algorithm AS 111. The percentage points of the normal distribution. *Applied Statistics*, **26** : 118-121.
65. **Beck M.W. (1997)**. Inference and generality in ecology: current problems and an experimental solution. *Oikos*, **78** : 265-273.
66. **Beckers F. & P. Bogaert (1998)**. Nonstationary of the mean and unbiased variogram estimation: extension of the weighted least-squares method. *Mathematical Geology*, **30** : 223-240.
67. **Belgrano A., P. Legendre, J.M. Dewarumez & S. Frontier (1995a)**. Spatial structure and ecological variation of meroplankton on the French-Belgian coast of the North sea. *Marine Ecology - Progress Series*, **128** : 43-50.
68. **Belgrano A., P. Legendre, J.M. Dewarumez & S. Frontier (1995b)**. Spatial structure and ecological variation of meroplankton on the Belgian-Dutch coast of the North sea. *Marine Ecology - Progress Series*, **128** : 51-59.
69. **Beliaeff B. & M.L. Cochard (1995)**. Applying geostatistics to identification of spatial patterns of fecal contamination in a mussel farming area (Havre de la Vanlée, France). *Water Research*, **29** : 1541-1548.
70. **Bellehumeur C. & P. Legendre (1997)**. Aggregation of sampling units: an analytical solution to predict variance. *Geographical Analysis*, **29** : 258-266.
71. **Bellehumeur C. & P. Legendre (1998)**. Multiscale sources of variation in ecological variables: modeling spatial dispersion, elaborating sampling designs. *Landscape Ecology*, **13** : 15-25.
72. **Bellehumeur C., P. Legendre & D. Marcotte (1997)**. Variance and spatial scales in a tropical rain forest: changing the size of sampling units. *Plant Ecology*, **130** : 89-98.
73. **Bellhouse D.R. (1977)**. Some optimal designs for sampling in two dimensions. *Biometrika*, **64** : 605-611.
74. **Bellhouse D.R. & B.C. Sutradhar (1988)**. Variance estimation for systematic sampling when autocorrelation is present. *Statistician*, **37** : 327-332.
75. **Benzaken V. & A. Doucet (1993)**. Bases de données orientées objet. Armand Colin, Paris, France, 126 p.
76. **Benzécri J.P. (1984)**. L'analyse des données. I. La taxinomie. Quatrième édition. Dunod, Paris, France, 635 p.
77. **Berge C. (1970)**. Graphes et hypergraphes. Dunod, Paris, France, 502 p.
78. **Berger J.O. & T. Sellke (1987)**. Testing a point null hypothesis: the irreconcilability of P values and evidence (with discussion, pp. 123-139). *Journal of the American Statistical Association*, **82** : 112-122.
79. **Berger M. (1977)**. Géométrie. I. Action de groupes, espaces affines et projectifs. Cedic & Fernand Nathan, Paris, France, 200 p.
80. **Bergeret F. & P. Besse (1997)**. Simulated annealing, weighted simulated annealing and genetic algorithm at work. *Computational Statistics Quarterly*, **12** : 447-465.
81. **Berntson G.M. & P. Stoll (1997)**. Correcting for finite spatial scales of self-similarity when calculating the fractal dimensions of real-world structures. *Proceedings of the Royal Society of London Series B*, **264** : 1531-1537.
82. **Berstel J., J.E. Pin & M. Pocchiola (1991)**. Mathématiques et informatique. Problèmes résolus. 2. Combinatoire et arithmétique. McGraw-Hill, Paris, France, 257 p.

83. **Bertin J. (1977)**. La graphique et le traitement graphique de l'information. Flammarion, Paris, France, 277 p.
84. **Bertsimas D. & J. Tsitsiklis (1993)**. Simulated annealing. *Statistical Science*, **8** : 10-15.
85. **Besag J. & P. Clifford (1989)**. Generalized Monte Carlo significance tests. *Biometrika*, **76** : 633-642.
86. **Besag J. & P. Clifford (1991)**. Sequential Monte Carlo p-values. *Biometrika*, **78** : 301-304.
87. **Biham O., O. Malcai, D.A. Lidar & D. Avnir (1998)**. Is nature fractal? - Response. *Science*, **279** : 785-786.
88. **Birks H.J.B. (1987)**. Recent methodological developments in quantitative descriptive biogeography. *Annales Zoologici Fennici*, **24** : 165-178.
89. **Birrell S.J., K.A. Sudduth & S.C. Borgelt (1996)**. Comparison of sensors and techniques for crop yield mapping. *Computers and Electronics in Agriculture*, **14** : 215-233.
90. **Bivand R. (1980)**. A Monte Carlo study of correlation estimation with spatially autocorrelated observations. *Quaestiones Geographicae*, **6** : 5-10.
91. **Blackburn G.A. & E.J. Milton (1996)**. Filling the gaps: remote sensing meets woodland ecology. *Global Ecology and Biogeography Letters*, **5** : 175-191.
92. **Blanc F., P. Chardy, A. Laurec & J.P. Reys (1976)**. Choix des métriques qualitatives en analyse d'inertie. Implications en écologie marine benthique. *Marine Biology*, **35** : 49-68.
93. **Blanché S., J. Casas, F. Bigler & K.A. Janssen-van Bergeijk (1996)**. An individual-based model of *Trichogramma* foraging behaviour: parameter estimation for single females. *Journal of Applied Ecology*, **33** : 425-434.
94. **Bliss C.I. (1971)**. The aggregation of species within spatial units (with discussion). *In* : Patil G.P., E.C. Pielou & W.E. Waters (Eds.), *Statistical Ecology. Volume 1. Spatial patterns and statistical distributions*, pp. 311-335. The Pennsylvania State University Press, University Park, Pennsylvania, USA.
95. **Blondel J. & J.D. Lebreton (1996)**. The biology of spatially structured populations: concluding remarks. *Acta Oecologica*, **17** : 687-693.
96. **Bocquet-Appel J.P. & J.N. Bacro (1994)**. Generalized wombling. *Systematic Biology*, **43** : 442-448.
97. **Bocquet-Appel J.P. & R.R. Sokal (1989)**. Spatial autocorrelation analysis of trend residuals in biological data. *Systematic Zoology*, **38** : 333-341.
98. **Boddy L., J.M. Wells, C. Culshaw & D.P. Donnelly (1999)**. Fractal analysis in studies of mycelium in soil. *Geoderma*, **88** : 301-328.
99. **Bogaert P. & D. Russo (1999)**. Optimal spatial sampling design for the estimation of the variogram based on a least squares approach. *Water Resources Research*, **35** : 1275-1289.
100. **Bonetto R.D. & J.L. Ladaga (1998)**. The variogram method for characterization of scanning electron microscopy images. *Scanning*, **20** : 457-463.
101. **Boots B.N. & C. Dufournaud (1994)**. A programming approach to minimizing and maximizing spatial autocorrelation statistics. *Geographical Analysis*, **26** : 54-66.
102. **Boots B.N. & G.F. Royle (1991)**. A conjecture on the maximum value of the principal eigenvalue of a planar graph. *Geographical Analysis*, **23** : 276-282.

103. **Borcard D. & P. Legendre (1994)**. Environmental control and spatial structure in ecological communities: an example using oribatid mites (Acari, Oribatei) (with discussion). *Environmental and Ecological Statistics*, **1** : 37-61.

104. **Borcard D., P. Legendre & P. Drapeau (1992)**. Partialling out the spatial component of ecological variation. *Ecology*, **73** : 1045-1055.

105. **Borel E. (1958)**. Les principes de la théorie des probabilités. Fascicule I. Principes et formules classiques du calcul des probabilités. Gauthier-Villars, Paris, France, 161 p.

106. **Borga M. & A. Vizzaccaro (1997)**. On the interpolation of hydrologic variables: formal equivalence of multiquadratic surface fitting and kriging. *Journal of Hydrology*, **195** : 160-171.

107. **Borgman L., M. Taheri & R. Hagan (1984)**. Three-dimensional frequency-domain simulations of geological variables. *In* : Verly G., M. David, A.G. Journel & A. Marechal (Eds.) *NATO-ASI geostatistics for natural resources characterization*, pp. 517-541. D. Reidel Publishing Company, Dordrecht, The Netherlands.

108. **Borland (1996)**. Guide de l'utilisateur. Borland Delphi pour Windows 95 et Windows NT. Version 2.0. Borland International, Inc., Scotts Valley, California, USA, 433 p.

109. **Boswell M.T. & G.P. Patil (1971)**. Chance mechanisms generating the logarithmic series distribution used in the analysis of number of species and individuals (with discussion). *In* : Patil G.P., E.C. Pielou & W.E. Waters (Eds.), *Statistical Ecology. Volume 1. Spatial patterns and statistical distributions*, pp. 99-130. The Pennsylvania State University Press, University Park, Pennsylvania, USA.

110. **Bouché M. (1994)**. La démarche objet. Concepts et outils. AFNOR, Paris-la-Défense, France, 313 p.

111. **Boucneau G., M. Vanmeirvenne, O. Thas & G. Hofman (1998)**. Integrating properties of soil map delineations into ordinary kriging. *European Journal of Soil Biology*, **49** : 213-229.

112. **Boufassa A. & M. Armstrong (1989)**. Comparison between different kriging estimators. *Mathematical Geology*, **21** : 331-345.

113. **Bouillé F. (1975)**. Structuration et saisie des données cartographiques. *In* : *Acquisition et structuration de l'information graphique*, pp. 1-32. Laboratoire de Tectonophysique, Université Pierre & Marie Curie - Paris 6, Paris, France.

114. **Bouillé F. (1977)**. Un modèle universel de banque de données, simultanément partageable, portable et répartie. Thèse de doctorat d'Etat, Université Pierre & Marie Curie - Paris 6, 447 p.

115. **Bourennane H., D. King, P. Chery & A. Bruand (1996)**. Improving the kriging of a soil variable using slope gradient as external drift. *European Journal of Soil Science*, **47** : 473-483.

116. **Bourgault G. (1992)**. Estimation et filtrage multidimensionnel de variables aléatoires régionalisées. PhD, Université de Montréal, Ecole Polytechnique, 289 p.

117. **Bourgault G. (1997)**. Using non-gaussian distributions in geostatistical simulations. *Mathematical Geology*, **29** : 315-334.

118. **Bourgault G. & D. Marcotte (1991)**. Multivariable variogram and its application to the linear model of coregionalization. *Mathematical Geology*, **23** : 899-928.

119. **Bourgault G., D. Marcotte & P. Legendre (1992)**. The multivariate co-variogram as a spatial weighting function in classification methods. *Mathematical Geology*, **24** : 463-478.

120. **Bouvier A. & M. George (1992)**. Dictionnaire des mathématiques. Troisième édition. Presses Universitaires de France, Paris, France, 834 p.

121. **Box G.E.P. (1954a)**. Some theorems on quadratic forms applied in the study of analysis of variance problems. I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, **25** : 290-302.

122. **Box G.E.P. (1954b)**. Some theorems on quadratic forms applied in the study of analysis of variance problems. II. Effects of inequality of variance and of correlation between errors in the two-way classification. *Annals of Mathematical Statistics*, **25** : 484-498.

123. **Bradley R. & J. Haslett (1992)**. Interactive graphics for the exploratory analysis of spatial data. The Interactive variogram cloud. *Sciences de la Terre, Série Informatique Géologique*, **31** : 373-386.

124. **Brannan K. & J.M. Hamlett (1998)**. Using geostatistics to select grid-cell layouts for the AGNPS model. *Transactions of the ASAE*, **41** : 1011-1018.

125. **Brassel K.E. & D. Reif (1979)**. A procedure to generate Thiessen polygons. *Geographical Analysis*, **11** : 289-303.

126. **Bregt A.K., A.B. McBratney & M.C.S. Wopereis (1991)**. Construction of isolinear maps of soil attributes with empirical confidence limites. *Soil Science Society of America Journal*, **55** : 14-19.

127. **Brillouin L. (1959)**. La science et la théorie de l'information (1988). Editions Jacques Gabay, Sceaux, France, 302 p.

128. **Brockman F.J. & C.J. Murray (1997)**. Subsurface microbiological heterogeneity: current knowledge, descriptive approaches and applications. *FEMS Microbiology Reviews*, **20** : 231-247.

129. **Brooker P.I. (1985)**. Two-dimensional simulation by turning bands. *Mathematical Geology*, **17** : 81-90.

130. **Brooker P.I. & M.A. Stewart (1994)**. A comparative study of simulation techniques for two dimensional data honouring specified exponential semivariograms. *Journal of the Australian Mathematical Society Series B*, **36** : 249-260.

131. **Brooker P.I., J.P. Winchester & A.C. Adams (1995)**. A geostatistical study of soil data from an irrigated vineyard near Waikerie, South Australia. *Environment International*, **21** : 699-704.

132. **Brunel C. (1986)**. Etude éco-entomologique des zones humides de la Chaussée-Tirancourt Vallée de la Somme. *Acta Oecologica - Oecologia Applicata*, **7** : 367-388.

133. **Brus D.J. & J.J. de Gruijter (1993)**. Design-based versus model-based estimates of spatial means: theory and application in environmental soil science. *Environmetrics*, **4** : 123-152.

134. **Brus D.J. & J.J. de Gruijter (1994)**. Estimation of non-ergodic variograms and their sampling variance by design-based sampling strategies. *Mathematical Geology*, **26** : 437-454.

135. **Brus D.J. & J.J. de Gruijter (1997)**. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma*, **80** : 1-59.

136. **Brus D.J., J.J. de Gruijter, B.A. Marsman, R. Visschers, A.K. Bregt, A. Brreuwsma & J. Bouma (1996)**. The performance of spatial interpolation methods and choropleth maps to estimate properties at points: a soil survey case study. *Environmetrics*, **7** : 1-16.

137. **Brus D.J., L.E.E.M. Spatjens & J.J. de Gruijter (1999)**. A sampling scheme for estimating the mean extractable phosphorus concentration of fields for environmental regulation. *Geoderma*, **89** : 129-148.
138. **Buckland S.T. (1984)**. Monte Carlo confidence intervals. *Biometrics*, **40** : 811-817.
139. **Buckland S.T. & D.A. Elston (1993)**. Empirical models for the spatial distribution of wildlife. *Journal of Applied Ecology*, **30** : 478-495.
140. **Burgess T.M. & R. Webster (1980a)**. Optimal interpolation and isarithmic mapping of soil properties. I. The semi-variogram and punctual kriging. *Journal of Soil Science*, **31** : 315-331.
141. **Burgess T.M. & R. Webster (1980b)**. Optimal interpolation and isarithmic mapping of soil properties. II. Block kriging. *Journal of Soil Science*, **31**: 333-341.
142. **Burgess T.M., R. Webster & A.B. McBratney (1981)**. Optimal interpolation and isarithmic mapping of soil properties. IV. Sampling strategy. *Journal of Soil Science*, **32** : 643-659.
143. **Burrough P.A. (1981)**. Fractal dimensions of landscapes and other environmental data. *Nature*, **294** : 240-242.
144. **Burrough P.A. (1983)**. Multiscale sources of spatial variation in soil. II. A non-brownian fractal model and its application in soil survey. *Journal of Soil Science*, **34** : 599-620.
145. **Burrough P.A. (1996)**. Principles of geographical information systems for land resources assessment. Clarendon Press - Oxford University Press, Oxford, UK, 194 p.
146. **Burrough P.A., P.F.M. van Gaans & R. Hootsmans (1997)**. Continuous classification in soil survey: spatial correlation, confusion and boundaries. *Geoderma*, **77** : 115-135.
147. **Bykat A. (1978)**. Convex hull of a finite set of points in two dimensions. *Information Processing Letters*, **7** : 296-298.
148. **Caloz R. & C. Collet (1997)**. Geographic information systems (GIS) and remote sensing in aquatic botany: methodological aspects. *Aquatic Botany*, **58** : 209-228.
149. **Campbell J.B. (1978)**. Locating boundaries between mapping units. *Mathematical Geology*, **10** : 289-299.
150. **Cannavacciuolo M., A. Bellido, D. Cluzeau, C. Gascuel & P. Trehen (1998)**. A geostatistical approach to the study of earthworm distribution in grassland. *Applied Soil Ecology*, **9** : 345-349.
151. **Cantwell M.D. & R.T.T. Forman (1993)**. Landscape graphs: ecological modeling with graph theory to detect configurations common to diverse landscapes. *Landscape Ecology*, **8** : 239-255.
152. **Cardina J., G.A. Johnson & D.H. Sparrow (1997)**. The nature and consequence of weed spatial distribution. *Weed Science*, **45** : 364-373.
153. **Cardina J., D.H. Sparrow & E.L. McCoy (1995)**. Analysis of spatial distribution of Common Lambsquarters (*Chenopodium album*) in no-till Soybean (*Glycine max*). *Weed Science*, **43**: 258-268.
154. **Cardina J., D.H. Sparrow & E.L. McCoy (1996)**. Spatial relationship between seedbank and seedling populations of Common Lambsquarters (*Chenopodium album*) and annual grasses. *Weed Science*, **44** : 298-308.
155. **Carle S.F. & G.E. Fogg (1996)**. Transition probability-based indicator geostatistics. *Mathematical Geology*, **28** : 453-476.
156. **Carr J.R., R.E. Bailey & E.D. Deng (1985)**. Use of indicator variograms for an enhanced spatial analysis. *Mathematical Geology*, **17** : 797-811.

157. **Carr J.R. & W.B. Benzer (1991)**. On the practice of estimating fractal dimension. *Mathematical Geology*, **23** : 945-958.

158. **Carr J.R. & N.H. Mao (1993)**. Comparison of disjunctive kriging to generalized probability kriging in application to the estimation of simulated and real data. *In* : Soares A. (Ed.), *Geostatistics Tróia '92*, pp. 13-24. Kluwer Academic Publishers, Dordrecht, The Netherlands.

159. **Carr J.R. & E. Myers (1990)**. Efficiency of different equation solvers in cokriging. *Computers & Geosciences*, **16** : 705-716.

160. **Carrat F. & A.J. Valleron (1992)**. Epidemiologic mapping using the "kriging" method: application to an influenza-like illness epidemic in France. *American Journal of Epidemiology*, **135** : 1293-1300.

161. **Carrat F. & A.J. Valleron (1993)**. Cartographie épidémiologique et prévision spatiale par krigeage. Application à l'étude des syndromes grippaux en France. *In* : Asselain B., M. Boniface, C. Duby, C. Lopez, J.P. Masson & J. Tranchefort (Eds.), *Biométrie et analyse de données spatio-temporelles*, pp. 96-110. Société Française de Biométrie, ENSA, Rennes, France.

162. **Carroll S.S. & D.L. Pearson (1998a)**. Spatial modeling of butterfly species richness using tiger beetles (Cicindelidae) as a bioindicator taxon. *Ecological Applications*, **8** : 531-543.

163. **Carroll S.S. & D.L. Pearson (1998b)**. The effects of scale and sample size on the accuracy of spatial predictions of tiger beetle (Cicindelidae) species richness. *Ecography*, **21** : 401-414.

164. **Cartan M. (1978)**. Inventaires et cartographies de répartitions d'espèces. Faune et flore. Editions du CNRS, Paris, France, 127 p.

165. **Casella G. & R.L. Berger (1987a)**. Reconciling bayesian and frequentist evidence in the one-sided testing problem (with discussion, pp. 123-139). *Journal of the American Statistical Association*, **82** : 106-111.

166. **Casella G. & R.L. Berger (1987b)**. Discussion of G. Casella and R.L. Berger "Reconciling bayesian and frequentist evidence in the one-sided testing problem": rejoinder. *Journal of the American Statistical Association*, **82** : 133-135.

167. **Cassel C.M., C.E. Särndal & J.H. Wretman (1977)**. Foundations of inference in survey sampling. Wiley, London, UK, 192 p.

168. **Cassie R.M. (1962)**. Frequency distribution models in the ecology of plankton and other organisms. *Journal of Animal Ecology*, **31** : 65-92.

169. **Castelier E. & P. Laurenge (1993)**. Krigeage de la dérive dans le cas d'un échantillonnage régulier. ENSMP, Fontainebleau, France, 20 p.

170. **Caujapé-Castells J. & J. Pedrola-Monfort (1997)**. Space-time patterns of genetic structure within a stand of *Androcymbium gramineum* (Cav.) McBride (Colchicaceae). *Heredity*, **79** : 341-349.

171. **Cayley A. (1859)**. On contour and slope lines. *Philosophical Magazine*, **18**: 264-268.

172. **Cerioli A. (1997)**. Modified tests of independence in 2 x 2 tables with spatial data. *Biometrics*, **53** : 619-628.

173. **Černý V. (1985)**. Thermodynamic approach to the traveling salesman problem: an efficient simulation algorithm. *Journal of Optimization Theory and Application*, **45** : 41-51.

174. **Cesaroni D., P. Matarazzo, G. Allegrucci & V. Sbordoni (1997)**. Comparing patterns of geographic variation in cave crickets by combining geostatistic methods and Mantel tests. *Journal of Biogeography*, **24** : 419-431.

175. **Chadœuf J. & P. Monestiez (1989)**. A new tessellation model derivated from the Voronoi model: properties, simulation, estimation. *Acta Stereologica*, **8** : 225-230.

176. **Chadœuf J., C. Moran & M. Goulard (1998)**. A note on extension variances in $\mathbb{R}^2$. *Mathematical Geology*, **30** : 575-587.

177. **Chami H. (1984)**. Contribution à l'optimisation d'un réseau de stations de mesure : utilisation de techniques d'interpolation spatiale pour la localisation des stations. Thèse de $3^{ème}$ cycle, Université des Sciences et Techniques du Languedoc - Montpellier 2, 158 p.

178. **Chan G. & A.T.A. Wood (1997)**. Algorithm AS 312. An algorithm for simulating stationary Gaussian random fields. *Applied Statistics*, **46** : 171-181.

179. **Chang Y.H., M.D. Scrimshaw, R.H.C. Emmerson & J.N. Lester (1998)**. Geostatistical analysis of sampling uncertainty at the Tollesbury Managed Retreat site in Blackwater Estuary, Essex, UK: kriging and cokriging approach to minimise sampling density. *Science of the Total Environment*, **221** : 43-57.

180. **Charmet G. & F. Balfourier (1995)**. The use of geostatistics for sampling a core collection of perennial ryegrass populations. *Genetic Resources and Crop Evolution*, **42** : 303-309.

181. **Charmet G., F. Balfourier & P. Monestiez (1994)**. Hierarchical clustering of perennial ryegrass populations with geographic contiguity constraint. *Theoretical and Applied Genetics*, **88** : 42-48.

182. **Chassery J.M. & A. Montanvert (1991)**. Géométrie discrète en analyse d'images. Hermès, Paris, France, 358 p.

183. **Chauvet P. (1982)**. The variogram cloud. *In* : *17th APCOM Symposium*, pp. 757-764. Colorado School of Mines, Golden, Colorado, USA.

184. **Chauvet P. (1985)**. Réflexions sur le variogramme et la géostatistique linéaire. ENSMP, Fontainebleau, France, 8 p.

185. **Chauvet P. (1986)**. La géostatistique non stationnaire : KU ou FAI-k ? ENSMP, Fontainebleau, France, 25 p.

186. **Chauvet P. (1988)**. Réflexions sur les pondérateurs négatifs du krigeage. *Sciences de la Terre, Série Informatique Géologique*, **28** : 65-113.

187. **Chauvet P. (1993)**. Processing data with a spatial support: geostatistics and its methods. ENSMP, Fontainebleau, France, 41 p.

188. **Chauvet P. (1994)**. Aide-mémoire de géostatistique linéaire. ENSMP, Fontainebleau, France, 210 p.

189. **Cheetham A.H. & J.E. Hazel (1969)**. Binary presence-absence similarity coefficients. *Journal of Paleontology*, **43** : 1130-1136.

190. **Chellemi D.O., K.G. Rohrbach, R.S. Yost & R.M. Sonoda (1988)**. Analysis of the spatial pattern of plant pathogens and diseased plants using geostatistics. *Phytopathology*, **78** : 221-226.

191. **Chen J., J.W. Hopmans & G.E. Fogg (1995)**. Sampling design for soil moisture measurements in large field trials. *Soil Science*, **159** : 155-161.

192. **Cheriton D. & R.G. Tarjan (1976)**. Finding minimum spanning trees. *SIAM Journal on Computing*, **5** : 724-742.

193. **Chessel D. (1978)**. Description non paramétrique de la dispersion spatiale des individus d'une espèce. *In* : Legay J.M. & R. Tomassone (Eds.), *Biométrie et écologie*, pp. 45-135. Société Française de Biométrie, Jouy-en-Josas, France.
194. **Chessel D. (1981)**. The spatial autocorrelation matrix. *Vegetatio*, **46** : 177-180.
195. **Chessel D. (1992)**. Echanges interdisciplinaires en analyse des données écologiques. Habilitation à diriger des recherches, Université Claude Bernard - Lyon 1, 108 p.
196. **Chessel D. & J.P. Croze (1978)**. Un indice de dispersion pour les mesures de présence-absence : application à la répartition des animaux et des plantes. *Bulletin d'Ecologie*, **9** : 19-28.
197. **Chessel D. & P. Mercier (1993)**. Couplage de triplets statistiques et liaisons espèces-environnement. *In* : Lebreton J.D. & B. Asselain (Eds.), *Biométrie et environnement*, pp. 15-43. Masson, Paris, France.
198. **Chessel D. & R. Sabatier (1993)**. Couplage de triplets statistiques et graphes de voisinage. *In* : Asselain B., M. Boniface, C. Duby, C. Lopez, J.P. Masson & J. Tranchefort (Eds.), *Biométrie et analyse de données spatio-temporelles*, pp. 28-37. Société Française de Biométrie, ENSA, Rennes, France.
199. **Chessel D., J. Thioulouse & S. Champely (1997)**. Autocorrélation et composantes cartographiables. *In* : Chessel D., J. Thioulouse, S. Dolédec & J.M. Olivier (Eds.), *Documentation thématique ADE-4. VII. Cartographie*, pp. 1-26. CNRS, Université Claude Bernard - Lyon 1, Lyon, France.
200. **Chevenet F. (1994)**. Un environnement coopératif de résolution de problèmes pour l'analyse statistique en écologie. Thèse de doctorat, Université Claude Bernard - Lyon 1, 189 p.
201. **Chiarello E. (1994)**. Pyramide stochastique et écologie du paysage : modélisation des structures spatiales par images de synthèse. Thèse de doctorat, Université Claude Bernard - Lyon 1, 179 p.
202. **Chiarello E., J.M. Jolion & C. Amoros (1993)**. Model of spatial organization of vegetation in river floodplains. *Acta Stereologica*, **12** : 255-261.
203. **Chiarello E., J.M. Jolion & C. Amoros (1996)**. Regions growing with the stochastic pyramid: application in landscape ecology. *Pattern Recognition*, **29** : 61-75.
204. **Choquet G. (1992)**. Cours de topologie. Masson, Paris, France, 317 p.
205. **Chou Y.H. & S. Soret (1996)**. Neighborhood effects in bird distributions, Navarre, Spain. *Environmental Management*, **20** : 675-687.
206. **Christakos G. (1992)**. Random field models in earth sciences. Academic Press, San Diego, USA, 474 p.
207. **Christakos G. & B.R. Killam (1993)**. Sampling design for classifying contaminant level using annealing search algorithms. *Water Resources Research*, **29**: 4063-4076.
208. **Christakos G. & R.A. Olea (1992)**. Sampling desing for spatially distributed hydrogeologic and environmental processes. *Advances in Water Resources*, **15** : 219-237.
209. **Christensen R. (1990)**. The equivalence of predictions from universal kriging and intrinsic random-function kriging. *Mathematical Geology*, **22** : 655-664.
210. **Chu J. (1996)**. Fast sequential indicator simulation: beyong reproduction of indicator variograms. *Mathematical Geology*, **28** : 923-936.
211. **Chun Z.X. (1990)**. Méthodologie de conception d'un système expert pour la généralisation cartographique. Thèse de doctorat, Ecole Nationale des Ponts et Chaussées, 234 p.
212. **Clark I. (1976)**. Some auxiliary functions for the spherical model of geostatistics. *Computers & Geosciences*, **1** : 255-263.

213. **Cliff A.D. (1970)**. Computing the spatial correspondence between geographical patterns. *Transactions of the Institute British Geographers*, **50** : 143-154.

214. **Cliff A.D. & J.K. Ord (1969)**. The problem of spatial autocorrelation. *In* : Scott A.J. (Ed.), *London papers in regional science. 1. Studies in regional science*, pp. 25-55. Pion, London, UK.

215. **Cliff A.D. & J.K. Ord (1971)**. Evaluating the percentage points of a spatial autocorrelation coefficient. *Geographical Analysis*, **3** : 51-61.

216. **Cliff A.D. & J.K. Ord (1973)**. Spatial autocorrelation. London, UK, 178 p.

217. **Cliff A.D. & J.K. Ord (1975)**. The comparison of means when samples consist of spatially autocorrelated observations. *Environment and Planning A*, **7** : 725-734.

218. **Cliff A.D. & J.K. Ord (1977)**. "Large sample-size distribution of statistics used in testing for spatial correlation" : a comment. *Geographical Analysis*, **9** : 297-299.

219. **Cliff A.D. & J.K. Ord (1981)**. Spatial processes. Models & applications. Pion, London, UK, 266 p.

220. **Clifford P. (1998)**. Discussion of Diggle *et al.*"Model-based geostatistics". *Applied Statistics*, **47** : 331-332.

221. **Clifford P. & S. Richardson (1985)**. Testing the association between two spatial processes. *Statistics & Decisions*, **2** : 155-160.

222. **Clifford P., S. Richardson & D. Hemon (1989)**. Assessing the significance of the correlation between two spatial processes. *Biometrics*, **45** : 123-134.

223. **Cline A.K., C.B. Moler, G.W. Stewart & J.H. Wilkinsons (1979)**. An estimate for the condition number of a matrix. *SIAM Journal of Numerical Analysis*, **16** : 368-375.

224. **Cochran W.G. (1946)**. Relative accuracy of systematic and stratified random samples for a certain class of populations. *Annals of Mathematical Statistics*, **17** : 164-177.

225. **Cochran W.G. (1977)**. Sampling techniques. Third edition. Wiley, New York, USA, 428 p.

226. **Cochran W.G., F. Mosteller & J.W. Tukey (1954)**. Principles of sampling. *Journal of the American Statistical Association*, **49** : 13-35.

227. **Collet C. (1992)**. Systèmes d'information géographique en mode image. Presses Polytechniques et Universitaires Romandes, Lausanne, Suisse, 186 p.

228. **Cook S.E. & N.A. Coles (1997)**. A comparison of soil survey methods in relation to catchment hydrology. *Australian Journal of Soil Research*, **35** : 1379-1395.

229. **Cooper S.D., L. Barmuta, O. Sarnelle, K. Kratz & S. Diehl (1997)**. Quantifying spatial heterogeneity in streams. *Journal of the North American Benthological Society*, **16** : 174-188.

230. **Coquillard P. & D.R.C. Hill (1997)**. Modélisation et simulation d'écosystèmes. Des modèles déterministes aux simulations à événements discrets. Masson, Paris, France, 273 p.

231. **Cornelius J.M. & J.F. Reynolds (1991)**. On determining the statistical significance of discontinuities within ordered ecological data. *Ecology*, **72** : 2057-2070.

232. **Corsten L.C.A. (1985)**. Current statistical issues in agricultural research. *Statistica Neerlandica*, **39** : 159-168.

233. **Corsten L.C.A. (1989)**. Interpolation and optimal linear prediction. *Statistica Neerlandica*, **43** : 69-84.

234. **Corsten L.C.A. & A. Stein (1994)**. Nested sampling for estimating spatial semivariograms compared to other designs. *Applied Statistics*, **10** : 103-122.

235. **Costa R., A.A. Peixoto, G. Barbujani & C.P. Kyriacou (1992)**. A latidudinal cline in a Drosophila clock gene. *Proceedings of the Royal Society of London Series B*, **250** : 43-49.
236. **Cousin R. (1988)**. Apport de la théorie des catégories à la représentation des connaissances. Thèse de doctorat d'Etat, Université Pierre & Marie Curie - Paris 6, 314 p.
237. **Cousins S.H. (1991)**. Species diversity measurement: choosing the right index. *Trends in Ecology & Evolution*, **6** : 190-192.
238. **Cox D.D., L.H. Cox & K.B. Ensor (1997)**. Spatial sampling and the environment: some issues and directions. *Environmental and Ecological Statistics*, **4** : 219-233.
239. **Cressie N. (1985)**. Fitting variogram models by weighted least squares. *Mathematical Geology*, **17** : 563-586.
240. **Cressie N. (1988a)**. Spatial prediction and ordinary kriging. *Mathematical Geology*, **20** : 405- 421.
241. **Cressie N. (1988b)**. Variogram. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 9, pp. 489-491. Wiley, New York, USA.
242. **Cressie N. (1989)**. Geostatistics. *American Statistician*, **43** : 197-202.
243. **Cressie N. (1990)**. The origins of kriging. *Mathematical Geology*, **22** : 239-252.
244. **Cressie N. (1991)**. Statistics for spatial data. Wiley, New York, USA, 900 p.
245. **Cressie N. (1993)**. Aggregation in geostatistical problems. *In* : Soares A. (Ed.), *Geostatistics Tróia '92*, pp. 25-36. Kluwer Academic Publishers, Dordrecht, The Netherlands.
246. **Cressie N. & D.M. Hawkins (1980)**. Robust estimation of the variogram. *Mathematical Geology*, **12** : 115-125.
247. **Crist T.O. (1998)**. The spatial distribution of termites in shortgrass steppe: a geostatistical approach. *Oecologia*, **114** : 410-416.
248. **Critten D.L. (1997)**. Fractal dimension relationships and values associated with certain plant canopies. *Journal of Agricultural Engineering Research*, **67** : 61-72.
249. **Croft F. & B. Kessler (1996)**. Remote sensing, image processing, and GIS. Trends and Forecasts. *Journal of Forestry*, **94** : 31-35.
250. **Croner C.M., J. Sperling & F.R. Broome (1996)**. Geographic information systems (GIS): new perspectives in understanding human health and environmental relationships. *Statistics in Medicine*, **15** : 1961-1977.
251. **Crovello T.J. (1981)**. Quantitative biogeography: an overview. *Taxon*, **30** : 563-575.
252. **Crowley P.H. (1992)**. Resampling methods for computation-intensive data analysis in ecology and evolution. *Annual Review of Ecology and Systematics*, **23** : 405-447.
253. **Crozel D. & M. David (1985)**. Global estimation variance: formulas and calculation. *Mathematical Geology*, **17** : 785-796.
254. **Csillag F. & S. Kabos (1996)**. Hierarchical decomposition of variance with applications in environmental mapping based on satellite images. *Mathematical Geology*, **28** : 385-405.
255. **Cullinan V.I. & J.M. Thomas (1992)**. A comparison of quantitative methods for examining landscape pattern and scale. *Landscape Ecology*, **7**: 211-227.
256. **Czárán T. & S. Bartha (1992)**. Spatiotemporal dynamic models of plant populations and communities. *Trends in Ecology & Evolution*, **7** : 38-42.
257. **Daget J. (1979)**. Les modèles mathématiques en écologie. Masson, Paris, France, 172 p.
258. **Daget P.H. (1980)**. Le nombre de diversité de Hill : un concept unificateur dans la théorie de la diversité écologique. *Acta Oecologica - Oecologia Generalis*, **1** : 51-70.

259. **Dagnelie P. (1986)**. Théorie et méthodes statistiques. Vol. 2. Les Presses Agronomiques de Gembloux, Gembloux, Belgique, 463 p.

260. **Dahiya I.S., D.J. Dahiya, A.V. Shanwal, R.D. Laura & R.P. Agrawal (1990)**. Geostatistical analysis of temporal variation in water content of sand dune soils. *International Journal of Tropical Agriculture*, **8** : 54-65.

261. **Dalenius T., J. Hajek & S. Zubrzycki (1960)**. On plane sampling and related geometrical problems. *In* : *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 125-150. University of California Press, Berkeley, California, USA.

262. **Dandurand L.M., G.R. Knudsen & D.J. Schotzko (1995)**. Quantification of *Pythium ultimum* var. *soprangiiferum* zoospore encystment patterns using geostatistics. *Phytopathology*, **85** : 186-190.

263. **Dandurand L.M., D.J. Schotzko & G.R. Knudsen (1997)**. Spatial patterns of rhizoplane populations of *Pseudomonas fluorescens*. *Applied and Environmental Microbiology*, **63** : 3211-3217.

264. **Dang M.V. (1998)**. Classification de données spatiales : modèles probabilistes et critères de partitionnement. Thèse de doctorat, Université de Technologie de Compiègne, 250 p.

265. **David B. (1991)**. Modélisation, représentation et gestion d'information géographique. Une approche en relationnel étendu. Thèse de doctorat, Université Pierre & Marie Curie - Paris 6, 215 p.

266. **David M. (1977)**. Geostatistical ore reserve estimation. Elsevier, New York, USA, 364 p.

267. **Davis B.M. (1987)**. Uses and abuses of cross-validation in geostatistics. *Mathematical Geology*, **19** : 241-248.

268. **Davis F.W. & J. Dozier (1990)**. Information analysis of a spatial database for ecological land classification. *Photogrammetric Engineering and Remote Sensing*, **56** : 605-613.

269. **Davis G.J. & M.D. Morris (1997)**. Six factors which affect the condition number of matrices associated with kriging. *Mathematical Geology*, **29** : 669-683.

270. **Davis J.C. (1986)**. Statistics and data analysis in geology. Wiley, New York, USA, 646 p.

271. **Davis M.W. (1987a)**. Production of conditional simulations via the LU triangular decomposition of the covariance matrix. *Mathematical Geology*, **19** : 91-98.

272. **Davis M.W. (1987b)**. Generating large stochastic simulations. The matrix polynomial approximation method. *Mathematical Geology*, **19** : 99-107.

273. **Davis M.W.D. & M. David (1980)**. An algorithm for finding the position of a point relative to a fixed boundary. *Mathematical Geology*, **12** : 61-68.

274. **Dawid A.P. (1983)**. Inference, statistical: I. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 4, pp. 84-105. Wiley, New York, USA.

275. **De Cesare L. & D. Posa (1995)**. A simulation technique of a non-Gaussian spatial process. *Computational Statistics & Data Analysis*, **20** : 543-555.

276. **De Finetti B. (1979)**. Probability and exchangeability from a subjective point of view. *International Statistical Review*, **47** : 129-135.

277. **De Fouquet C. (1993)**. Géostatistique orientée vers le traitement des données territoriales: notions-clés et exemples. *Biométrie Praximétrie*, **33** : 113-146.

278. **De Gruijter J.J. & C.J.F. ter Braak (1990)**. Model-free estimation from spatial samples: a reappraisal of classical sampling theory. *Mathematical Geology*, **22** : 407-415.

279. **De Gruijter J.J. & C.J.F. ter Braak (1992)**. Design-based versus model-based sampling strategies: comment on R.J. Barnes "Bounding the required sample size for geologic site characterization". *Mathematical Geology*, **24** : 859-864.

280. **De Gruijter J.J., D.J.J. Walvoort & P.F.M. van Gaans (1997)**. Continuous soil maps. A fuzzy set approach to bridge the gap between aggregation levels of process and distribution models. *Geoderma*, **77** : 169-195.

281. **De Kwaadsteniet J.W. (1990)**. On some fundamental weak spots of kriging technique and their consequences. *Journal of Hydrology*, **114** : 277-284.

282. **Debinski D.M. & P.S. Humphrey (1997)**. An integrated approach to biological diversity assessment. *Natural Areas Journal*, **17** : 355-365.

283. **Debouzie D., A. Bendjedid, T. Bensid & N. Gautier (1996)**. *Stippa tenacissima* aerial biomass estimated at regional scale in an Algerian steppe, using geostatistical tools. *Vegetatio*, **124** : 173-181.

284. **Deichman U. & L. Anselin (1994)**. Exploratory spatial data analysis of categorical variables: an application to african farming systems data. *In* : Harts J.J., H.F.L. Ottens, H.J. Scholten & J. van Arragon (Eds.) *EGIS/MARI '94. Fifth european conference and exhibition on geographical information systems*, pp. 2107-2116. EGIS Foundation, Utrecht, The Netherlands.

285. **Delaville L., J.P. Rossi & P. Quenehervé (1996)**. Plant row and soil factors influencing the microspatial patterns of plant-parasitic nematodes on sugarcane in Martinique. *Fundamental and Applied Nematology*, **19** : 321-328.

286. **Delay F. & G.H. de Marsily (1994)**. The integral of the semivariogram: a powerful method for adjusting the semivariogram in geostatistics. *Mathematical Geology*, **26** : 301-321.

287. **Delcourt H., P.L. Darius & J. de Baerdemaeker (1996)**. The spatial variability of some aspects of topsoil fertility in two Belgian fields. *Computers and Electronics in Agriculture*, **14** : 179-196.

288. **Delfiner P. & G. Matheron (1980)**. Les fonctions aléatoires intrinsèques d'ordre k. ENSMP, Fontainebleau, France, 36 p.

289. **Delobel C., C. Lécluse & P. Richard (1991)**. Bases de données : des systèmes relationnels aux systèmes à objets. InterEditions, Paris, France, 460 p.

290. **Demazure M. (1988)**. Introduction à l'algorithmique numérique et à la programmation en Pascal. Cours et exercices corrigés. McGraw-Hill, Paris, France, 621 p.

291. **Despland E. & G. Houle (1997)**. Climate influences on growth and reproduction of *Pinus banksiana* (Pinaceae) at the limit of the species distribution in eastern North America. *American Journal of Botany*, **84** : 928-937.

292. **Dessaint F. & J.P. Caussanel (1994)**. Trend surface analysis: a simple tool for modelling spatial patterns of weeds. *Crop Protection*, **13** : 433-438.

293. **Deutsch C.V. (1996)**. Correcting for negative weights in ordinary kriging. *Computers & Geosciences*, **22** : 765-773.

294. **Deutsch C.V. & P.W. Cockerham (1994)**. Practical considerations in the application of simulated annealing to stochastic simulation. *Mathematical Geology*, **26** : 67-82.

295. **Deutsch C.V. & A.G. Journel (1992)**. GSLIB. Geostatistical Software Library and user's guide. Oxford University Press, New York, USA, 340 p.

296. **Deverly F. (1984a)**. Application de la géostatistique aux problèmes d'échantillonnage. *Sciences de la Terre, Série Informatique Géologique*, **18** : 27-45.

297. **Deverly F. (1984b)**. Echantillonnage et géostatistique. Thèse de Docteur-Ingénieur, ENSMP, Fontainebleau, 128 p.

298. **Devillers O., S. Meiser & M. Teillaud (1992)**. Fully dynamic Delaunay triangulation in logarithmic expected time per operation. *Computational Geometry - Theory and Applications*, **2** : 55-80.
299. **Devroye L. & G.T. Toussaint (1981)**. A note on linear expected time algorithms for finding convex hulls. *Computing*, **26** : 361-366.
300. **Di H.J., B.B. Trangmar & R.A. Kemp (1989)**. Use of geostatistics in designing sampling strategies for soil survey. *Soil Science Society of America Journal*, **53** : 1163-1167.
301. **Diaconis P. & B. Efron (1989)**. Méthodes de calculs statistiques intensifs sur ordinateurs. *In* : *Le calcul intensif*, pp. 123-141. Belin, Paris, France.
302. **Diamond P. & M. Armstrong (1984)**. Robustness of variograms and conditioning of kriging matrices. *Mathematical Geology*, **16** : 809-822.
303. **Diaz G., A. Zucca, M.D. Setzu & C. Cappai (1997)**. Chromatin pattern by variogram analysis. *Microscopy Research and Technique*, **39** : 305-311.
304. **Diday E. (1971)**. Une nouvelle méthode en classification automatique et reconnaissance des formes, la méthode des nuées dynamiques. *Revue de Statistique Appliquée*, **19** : 19-33.
305. **Diday E., J. Lemaire, J. Pouget & F. Testu (1982)**. Eléments d'analyse de données. Dunod, Paris, France, 464 p.
306. **Dietrich C.R. (1990)**. Sensitivity of kriging and spline interpolation to data perturbation. *Mathematics and Computers in Simulation*, **32** : 191-196.
307. **Dietrich C.R. (1991)**. Modality of the restricted likelihood for spatial Gaussian random fields. *Biometrika*, **78** : 833-839.
308. **Dietrich C.R. (1993a)**. Computationally efficient Cholesky factorization of a covariance matrix with block Toeplitz structure. *Journal of Statistical Computation and Simulation*, **45** : 203-218.
309. **Dietrich C.R. (1993b)**. Computationally efficient generation of Gaussian conditional simulations over regular sample grids. *Mathematical Geology*, **25** : 439-451.
310. **Dietrich C.R. (1995)**. A simple and efficient space domain implementation of the turning bands method. *Water Resources Research*, **31** : 147-156.
311. **Dietrich C.R. & G.N. Newsam (1993)**. A fast and exact method for multidimensional Gaussian stochastic simulations. *Water Resources Research*, **29**: 2861-2869.
312. **Dietrich C.R. & G.N. Newsam (1994)**. Generating correlated Gaussian random fields by orthogonal polynomial approximation to the square root of the covariance matrix. *Journal of Statistical Computation and Simulation*, **50** : 91-109.
313. **Dietrich C.R. & G.N. Newsam (1995)**. Efficient generation of conditional simulations by Chebyshev matrix polynomial approximations to the symmetric square root of the covariance matrix. *Mathematical Geology*, **27** : 207-228.
314. **Dietrich C.R. & G.N. Newsam (1996)**. A fast exact method for multidimensional Gaussian stochastic simulations: extension to realizations conditioned on direct and indirect measurements. *Water Resources Research*, **32** : 1643-1652.
315. **Dietrich C.R. & G.N. Newsam (1997)**. Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix. *SIAM Journal on Scientific Computing*, **18** : 1088-1107.
316. **Dietz E.J. (1983)**. Permutation tests for association between two distance matrices. *Systematic Zoology*, **32** : 21-26.
317. **Diggle P.J. (1983)**. Statistical analysis of spatial point patterns. Academic Press, London, UK, 148 p.

318. **Dijkstra E.W. (1959)**. A note on two problems in connexion with graphs. *Numberische Mathematik*, **1** : 269-271.
319. **Dimitrakopoulos R. (1990)**. Conditional simulation of intrinsic random function of order k. *Mathematical Geology*, **22** : 361-380.
320. **Dimitrakopoulos R. (1993)**. Artificially Intelligent geostatistics: a framework accomodating qualitative knowledge-information. *Mathematical Geology*, **25** : 261-279.
321. **Ding Y. & A. Fotheringham (1992)**. The integration of spatial analysis and GIS. *Computers Environment and Urban Systems*, **16** : 3-19.
322. **Diniz-Filho J.A.F. & L.M. Bini (1996)**. Assessing the relationship between multivariate community structure and environmental variables. *Marine Ecology - Progress Series*, **143** : 303-306.
323. **Dixmier J. (1981)**. Topologie générale. Presses Universitaires de France, Paris, France, 164 p.
324. **Dodge Y. (1993)**. Statistique. Dictionnaire encyclopédique. Dunod, Paris, France, 409 p.
325. **Dodge Y. (1996)**. A natural random number generator. *International Statistical Review*, **64** : 329-344.
326. **Dodson R. & D. Marks (1997)**. Daily air temperature interpolated at high spatial resolution over a large mountainous region. *Climate Research*, **8**: 1-20.
327. **Doligez A. & H.I. Joly (1997)**. Genetic diversity and spatial structure within a natural stand of a tropical forest tree species, *Carapa procera* (Meliaceae), in French Guiana. *Heredity*, **79** : 72-82.
328. **Domburg P., J.J. de Gruijter & D.J. Brus (1994)**. A structured approach to designing soil survey schemes with prediction of sampling error from variograms. *Geoderma*, **62** : 151-164.
329. **Domburg P., J.J. de Gruijter & P. van Beek (1997)**. Designing efficient soil survey schemes with a knowledge-based system using dynamic programming. *Geoderma*, **75** : 183-201.
330. **Donald W.W. (1994)**. Geostatistics for mapping weeds, with a Canada Thistle (*Cirsium arvense*) patch as a case study. *Weed Science*, **42** : 648-657.
331. **Donnay J.P. (1994)**. Construction d'un indice de conformité locale entre deux surfaces transcrites en mode image. *In* : Harts J.J., H.F.L. Ottens, H.J. Scholten & J. van Arragon (Eds.) *EGIS/MARI '94. Fifth european conference and exhibition on geographical information systems*, pp. 1131-1139. EGIS Foundation, Utrecht, The Netherlands.
332. **Douglas M.E. & J.A. Endler (1982)**. Quantitative matrix comparisons in ecological and evolutionary investigations. *Journal of Theoretical Biology*, **99** : 777-795.
333. **Dow M.M. & J.M. Cheverud (1985)**. Comparison of distance matrices in studies of population structure and genetic microdifferenciation: quadratic assigment. *American Journal of Physical Anthropology*, **68** : 367-373.
334. **Dowd P.A. (1992)**. A review of recent developments in geostatistics. *Computers & Geosciences*, **17** : 1481-1500.
335. **Dubrule O. (1983a)**. Two methods with different objectives: splines and kriging. *Mathematical Geology*, **15** : 245-257.
336. **Dubrule O. (1983b)**. Cross validation of kriging in a unique neighborhood. *Mathematical Geology*, **15** : 687-699.
337. **Dubrule O. (1984)**. Comparing splines and kriging. *Computers & Geosciences*, **10** : 327-338.

338. **Dubrule O. & C. Kostov (1986)**. An interpolation method taking into account inequality constraints. I. Methodology. *Mathematical Geology*, **18** : 33-51.
339. **Dunn R. & A.R. Harrison (1993)**. Two-dimensional systematic sampling of land use. *Applied Statistics*, **42** : 585-601.
340. **Durbin J. & G.S. Watson (1950)**. Testing for serial correlation in least squares regression. I. *Biometrika*, **37** : 409-428.
341. **Dutilleul P. (1993)**. Modifying the t test for assessing the correlation between two spatial processes. *Biometrics*, **49** : 305-314.
342. **Dutilleul P., C. Bellehumeur & P. Legendre (1993)**. L'intervalle de confiance de la moyenne spatiale d'un processus spatial autocorrélé. Ministère de l'Environnement du Québec & Environnement Canada. Montréal, Québec, 37 p.
343. **Dutreix C. (1986)**. Essai de biogéographie écologique numérique à l'échelle régionale. Etude en Bourgogne de la super-famille Papilionoidea (Lepidoptera). Thèse d'Université, Université Aix-Marseille 1, 175 p.
344. **Easley D.H., L.E. Borgman & D. Weber (1991)**. Monitoring well placement using conditional simulation of hydraulic head. *Mathematical Geology*, **23** : 1059-1080.
345. **Eddy W.F. (1977)**. A new convex hull algorithm for planar sets. *ACM Transactions on Mathematical Software*, **3** : 398-403.
346. **Edelsbrunner H., T.S. Tan & R. Waupotitsch (1992)**. An $O(n^2 \log n)$ time algorithm for the minmax angle triangulation. *SIAM Journal on Scientific and Statistical Computing*, **13** : 994-1008.
347. **Edgington E.S. (1986)**. Randomization tests. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 7, pp. 530-538. Wiley, New York, USA.
348. **Edgington E.S. (1987)**. Randomization tests. Second edition. Marcel Dekker, New York, USA, 341 p.
349. **Efron B. & G. Gong (1983)**. A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician*, **37** : 36-48.
350. **Efron B. & R.J. Tibshirani (1993)**. An introduction to the bootstrap. Chapman & Hall, New York, USA, 436 p.
351. **Eghball B., G.W. Hergert, G.W. Lesoing & R.B. Ferguson (1999)**. Fractal analysis of spatial and temporal variability. *Geoderma*, **88** : 349-362.
352. **El Bahi M. (1981)**. Représentations transitives des variables régionalisées : examen des fondements théoriques et essai d'utilisation pour l'analyse spatio-temporelle d'une population d'Insectes (*Aedes cataphylla* L.). Thèse de $3^{ème}$ cycle, Université Claude Bernard - Lyon 1, 131 p.
353. **Elbaz M. & J.C. Spehner (1990)**. Construction du diagramme de Voronoï dans le plan en utilisant la structure de carte. *Bigre*, **67** : 171-181.
354. **Englund E.J. (1990)**. A variance of geostatisticians. *Mathematical Geology*, **22** : 417-455.
355. **Epperson B.K. (1990)**. Spatial autocorrelation of genotypes under directional selection. *Genetics*, **124** : 757-771.
356. **Epperson B.K. (1995)**. Spatial distribution of genotypes under isolation by distance. *Genetics*, **140** : 1431-1440.
357. **Epperson B.K. & E.R. Alvarez-Buylla (1997)**. Limited seed dispersal and genetic structure in life stages of *Cecropia obtusifolia*. *Evolution*, **51** : 275-282.
358. **Epperson B.K., Z. Huang & T.Q. Li (1999)**. Measures of spatial structure in samples of genotypes for multiallelic loci. *Genetical Research - Cambridge*, **73** : 251-261.

359. **Epperson B.K. & T.Q. Li (1996)**. Measurement of genetic structure within populations using Moran's spatial autocorrelation statistics. *Proceedings of the National Academy of Sciences of the USA*, **93** : 10528-10532.

360. **Epperson B.K. & T.Q. Li (1997)**. Gene dispersal and spatial genetic structure. *Evolution*, **51** : 672-681.

361. **Eshel A. (1998)**. On the fractal dimensions of a root system. *Plant Cell and Environment*, **21** : 247-251.

362. **Evans J.W., F. Harary & M.S. Lynn (1967)**. On the computer enumeration of finite topologies. *Communications of the ACM*, **10** : 295-297.

363. **Eynon B.P. & P. Switzer (1982)**. The variability of acide rainfall. Technical report N° 58. Department of statistics, Stanford University. Stanford, California, USA.

364. **Fahrig L. (1988)**. A general model of populations in patchy habitats. *Applied Mathematics and Computation*, **27** : 53-66.

365. **Fahrig L. & G. Merriam (1985)**. Habitat patch connectivity and population survival. *Ecology*, **66** : 1762-1768.

366. **Fahrig L. & G. Merriam (1994)**. Conservation of fragmented populations. *Conservation Biology*, **8** : 50-59.

367. **Fairfield-Smith H. (1938)**. An empirical law describing heterogeneity in the yields of agricultural crops. *Journal of Agricultural Science*, **28**: 1-23.

368. **Fang J.H. & P.P. Wang (1997)**. Random field generation using simulated annealing vs. fractal-based stochastic interpolation. *Mathematical Geology*, **29** : 849-858.

369. **Fang K.T., Y. Wang & P.M. Bentler (1994)**. Some applications of number-theoretic methods in statistics. *Statistical Science*, **9** : 416-428.

370. **Fariña A.C., J. Freire & E. Gonzales-Gurriaran (1997)**. Demersal fish assemblages in the Galician continental shelf and upper slope (NW Spain): spatial structure and long term changes. *Estuarine Coastal and Shelf Science*, **44** : 435-454.

371. **Faust K. & A.K. Romney (1985)**. The effect of skewed distributions on matrix permutation tests. *British Journal of Mathematical and Statistical Psychology*, **38** : 152-160.

372. **Feo T.A. & M.G.C. Resende (1995)**. Greedy randomized adaptative search procedure. *Journal of Global Optimization*, **6** : 109-133.

373. **Ferri M. & M. Piccioni (1992)**. Optimal selection of statistical units. *Computational Statistics & Data Analysis*, **13** : 47-61.

374. **Fielding A.H. & J.F. Bell (1997)**. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24** : 38-49.

375. **Fingleton B. (1983)**. Independence, stationary, categorical spatial data and the chi-squared test. *Environment and Planning A*, **15** : 483-499.

376. **Fisher D.R. (1968)**. A study of faunal resemblance using numerical taxonomy and factor analysis. *Systematic Zoology*, **17** : 48-63.

377. **Fisher W.D. (1958)**. On grouping for maximum homogeneity. *Journal of the American Statistical Association*, **53** : 789-798.

378. **Fons J., T. Sauras, J. Romanya & V.R. Vallejo (1997)**. Sampling strategies in forest soils. *Annales des Sciences Forestieres*, **54** : 493-499.

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

380. **Fortin M.J. (1997)**. Effects of data types on vegetation boundary delineation. *Canadian Journal of Forest Research*, **27** : 1851-1858.

381. **Fortin M.J. (1999)**. Effects of quadrat size and data measurement on the detection of boundaries. *Journal of Vegetation Science*, **10** : 43-50.

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

383. **Fortin M.J., P. Drapeau & G.M. Jacquez (1996)**. Quantification of the spatial co-occurrences of ecological boundaries. *Oikos*, **77** : 51-60.

384. **Fortin M.J., P. Drapeau & P. Legendre (1989)**. Spatial autocorrelation and sampling design in plant ecology. *Vegetatio*, **83** : 209-222.

385. **Fournier A. (1979)**. Comments on convex hull of a finite set of points in two dimensions. *Information Processing Letters*, **8** : 173-175.

386. **Fox B.L. (1993)**. Integrating and accelerating tabu search, simulated annealing, and genetic algorithms. *Annals of Operation Research*, **41** : 47-67.

387. **Franceschini G., M. Cannavacciuolo & F. Burel (1997)**. A geostatistical analysis of the spatial distribution of *Abax parallelepipedus* (Coleoptera, Carabidae) in a woodlot. *European Journal of Soil Biology*, **33** : 117-122.

388. **Franquet E. & D. Chessel (1994)**. Approche statistique des composantes spatiales et temporelles de la relation faune-milieu. *Comptes Rendus de l'Académie des Sciences série III - Sciences de la Vie*, **317** : 202-206.

389. **Freire J., E. Gonzalez-Gurriaran & I. Olaso (1992)**. Spatial distribution of *Munida intermedia*and *M. sarsi (*Crustacea: Anomura) on the Galician continental shelf NW Spain: application of geostatistical analysis. *Estuarine Coastal and Shelf Science*, **35** : 637-648.

390. **Froidevaux C., M.C. Gaudel & M. Soria (1993)**. Types de données et algorithmes. Ediscience international, Paris, France, 577 p.

391. **Frontier S. (1983a)**. Choix et contraintes de l'échantillonnage écologique. *In* : Frontier S. (Ed.), *Stratégies d'échantillonnage en écologie*, pp. 15-62. Masson, Paris, France.

392. **Frontier S. (1983b)**. L'échantillonnage de la diversité biologique. *In* : Frontier S. (Ed.), *Stratégies d'échantillonnage en écologie*, pp. 416-436. Masson, Paris, France.

393. **Frontier S. & D. Pichod-Viale (1991)**. Ecosystèmes : structure, fonctionnement, évolution. Masson, Paris, France, 392 p.

394. **Gabow H.N., Z. Galil, T. Spencer & R.E. Tarjan (1986)**. Efficient algorithms for finding minimum spanning trees in undirected and directed graph. *Combinatorica*, **6** : 109-122.

395. **Gabriel K.R. & R.R. Sokal (1969)**. A new statistical approach to geographic variation analysis. *Systematic Zoology*, **18** : 259-278.

396. **Gaines S.D. & M.W. Denny (1993)**. The largest, smallest, highest, lowest, longest and shortest: extremes in ecology. *Ecology*, **74** : 1677-1692.

397. **Galambos J. (1982)**. Exchangeability. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 2, pp. 573-577. Wiley, New York, USA.

398. **Gallerani-Lawson E.J. & A.R. Rodgers (1997)**. Differences in home-range size computed in commonly used software programs. *Wildlife Society Bulletin*, **25** : 721-729.

399. **Galli A. & H. Wackernagel (1987)**. Multivariate geostatistical methods for spatial data analysis. ENSMP, Fontainebleau, France, 8 p.

400. **Gallichand J., D. Marcotte, S.O. Prasher & R.S. Broughton (1992)**. Optimal sampling density of hydraulic conductivity for subsurface drainage in the Nole delta. *Agricultural Water Management*, **20** : 299-312.

401. **Gao H., J. Wang & P. Zhao (1996)**. The updated kriging variance and optimal sample design. *Mathematical Geology*, **28** : 295-313.
402. **Gardner R.H. (1998)**. Pattern, process, and the analysis of spatial scales. *In* : Peterson D.L. & V.T. Parker (Eds.), *Ecological Scale*, pp. 17-34. Columbia University Press, New York, USA.
403. **Gascuel-Odoux C. & P. Boivin (1994)**. Variability of variograms and spatial estimates due to soil sampling: a case study. *Geoderma*, **62** : 165-182.
404. **Gascuel-Odoux C., C. Walter & M. Voltz (1993)**. Intérêt du couplage des méthodes géostatistiques et de cartographie des sols pour l'estimation spatiale. *Sciences du Sol*, **31** : 193-213.
405. **Gatrell A.C. (1977)**. Complexity and redundancy in binary maps. *Geographical Analysis*, **9** : 29-41.
406. **Gatrell A.C. (1979)**. Autocorrelation in spaces. *Environment and Planning A*, **11** : 507-516.
407. **Gautschi W. (1957)**. Some remarks on systematics sampling. *Annals of Mathematical Statistics*, **28** : 385-394.
408. **Geary R.C. (1954)**. The contiguity ratio and statistical mapping (with discussion). *Incorporated Statistician*, **5** : 115-145.
409. **Genton M.G. (1998a)**. Highly robust variogram estimation. *Mathematical Geology*, **30** : 213-221.
410. **Genton M.G. (1998b)**. Variogram fitting by generalized least squares using an explicit formula for the covariance structure. *Mathematical Geology*, **30** : 323-345.
411. **Geoghegan J., L.A. Wainger & N.E. Bockstael (1997)**. Spatial landscape indices in a hedonic framework: an ecological economics analysis using GIS. *Ecological Economics*, **23** : 251-264.
412. **Gerhards R., D.Y. Wyse-Pester, D. Mortensen & G.A. Johnson (1997)**. Characterizing spatial stability of weed populations using interpolated maps. *Weed Science*, **45** : 108-119.
413. **Getis A. (1989)**. A spatial association model approach to the identification of spatial dependence. *Geographical Analysis*, **21** : 251-259.
414. **Ghosh B. (1951)**. Random distances within a rectangle and between two rectangles. *Bulletin of the Calcutta Mathematical Society*, **43** : 17-24.
415. **Gibbons J.D. (1986)**. P values. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 7, pp. 366-368. Wiley, New York, USA.
416. **Gilbert M. (1997)**. A user-friendly PC-based GIS for forest entomology: an attempt to combine existing software. *In* : Gregoire J.C., A.M. Liebhold, F.M. Stephen, K.R. Day & S.M. Salom (Eds.), *Integrating Cultural Tactics into the Management of Bark Beetle and Reforestation Pests*, pp. 54-61. US Dept. Agr., 5 Radnor Corporate Ctr.
417. **Gittins R. (1968)**. Trend-surface analysis of ecological data. *Journal of Ecology*, **56** : 845-869.
418. **Glover F. (1989)**. Tabu search. Part I. *ORSA Journal on Computing*, **1**: 190-206.
419. **Glover F. (1990a)**. Tabu search. Part II. *ORSA Journal on Computing*, **2**: 4-32.
420. **Glover F. (1990b)**. Tabu search: a tutorial. *Interfaces*, **20** : 74-94.
421. **Glover F. (1994)**. Tabu search for nonlinear and parametric optimization (with links to genetic algorithms). *Discrete Applied Mathematics*, **49** : 231-255.
422. **Glover F., J.P. Kelly & M. Laguna (1995)**. Genetic algorithms and tabu search; hybrids for optimization. *Computers and Operations Research*, **22** : 111-134.

423. **Glover F., E. Taillard & D. Werra (1993)**. A user's guide to tabu search. *Annals of Operation Research*, **41** : 3-28.

424. **Gneiting T. (1998)**. Closed form solutions of the two-dimensional turning bands equation. *Mathematical Geology*, **30** : 379-390.

425. **Gohin F. (1987)**. Analyse géostatistique des champs thermiques de surface de la mer. Thèse de Docteur-Ingénieur, ENSMP, Fontainebleau, 103 p.

426. **Gohin F. (1990)**. Analyse géostatistique et cartographie des champs thermiques à partir d'observations obtenues par bateaux et par télédétection spatiale. *In* : Frontier S. (Ed.), *Biométrie et océanographie*, pp. 65-78. IFREMER, Plouzané, France.

427. **Gohin F. & G. Langlois (1993)**. Using geostatistics to merge in situ measurements and remotely-sensed observations of sea surface temperature. *International Journal of Remote Sensing*, **14** : 9-19.

428. **Goldberg D.E. (1994)**. Algorithmes génétiques. Exploration, optimisation et apprentissage automatique. Addison-Wesley, Paris, France, 417 p.

429. **Golub G.H. & C.F. van Loan (1983)**. Matrix computations. Johns Hopkins University Press, Baltimore, Maryland, USA, 476 p.

430. **Gómez-Hernández J.J. & R.M. Srivastava (1990)**. ISIM3D: an ANSI-C three-dimensional multiple indicator conditional simulation program. *Computers & Geosciences*, **16** : 395-440.

431. **Gonzales R.C. & P. Wintz (1987)**. Digital image processing. Second edition. Addison-Wesley, Reading, Massachusetts, USA, 503 p.

432. **Gonzalez-Gurriaran E., J. Freire & L. Fernandez (1993)**. Geostatistical analysis of spatial distribution of Liocarcinus depurator, *Macropipus tuberculatus* and *Polybius henslowii* (Crustacea: Brachyura) over the Galician continental shelf NW Spain. *Marine Biology*, **115** : 453-461.

433. **Good P. (1994)**. Permutation tests. A practical guide to resampling methods for testing hypotheses. Springer-Verlag, New York, USA, 228 p.

434. **Goodall D.W. (1974)**. A new method for the analysis of spatial pattern by random pairing of quadrats. *Vegetatio*, **29** : 135-146.

435. **Goodchild M.F. (1980)**. Fractals and the accuracy of geographical measures. *Mathematical Geology*, **12** : 85-98.

436. **Goovaerts P. (1992)**. Factorial kriging analysis: a useful tool for exploring the structure of multivariate spatial soil information. *Journal of Soil Science*, **43** : 597-619.

437. **Goovaerts P. (1993)**. Spatial orthogonality of the principal components computed from coregionalized variables. *Mathematical Geology*, **25** : 281-302.

438. **Goovaerts P. (1994a)**. On a controversial method for modeling a coregionalization. *Mathematical Geology*, **26** : 197-204.

439. **Goovaerts P. (1994b)**. Comparative performance of indicator algorithms for modeling conditional probability distribution functions. *Mathematical Geology*, **26** : 389-411.

440. **Goovaerts P. (1996)**. Stochastic simulation of categorical variables using a classification algorithm and simulated annealing. *Mathematical Geology*, **28** : 909-921.

441. **Goovaerts P. (1997)**. Geostatistics for natural resources evaluation. Oxford University Press, New York, USA, 483 p.

442. **Goovaerts P. (1998a)**. Ordinary cokriging revisited. *Mathematical Geology*, **30** : 21-42.

443. **Goovaerts P. (1998b)**. Accounting for estimation optimality criteria in simulated annealing. *Mathematical Geology*, **30** : 511-534.

444. **Goovaerts P. (1999)**. Geostatistics in soil science: state-of-the-art and perspectives. *Geoderma*, **89** : 1-45.
445. **Goovaerts P. & P. Sonnet (1993)**. Study of spatial and temporal variations of hydrogeochemical variables using factorial kriging analysis. *In* : Soares A. (Ed.), *Geostatistics Tróia '92*, pp. 745-756. Kluwer Academic Publishers, Dordrecht, The Netherlands.
446. **Goovaerts P., R. Webster & J.P. Dubois (1997)**. Assessing the risk of soil contamination in the Swiss Jura using indicator geostatistics. *Environmental and Ecological Statistics*, **4** : 31-48.
447. **Gordon A.D. & C.R. Finden (1985)**. Classification of spatially-located data. *Computational Statistics Quarterly*, **2** : 315-328.
448. **Gotway C.A. (1991)**. Fitting semivariogram models by weighted least squares. *Computers & Geosciences*, **17** : 171-172.
449. **Gotway C.A., R.B. Ferguson, G.W. Hergert & T.A. Peterson (1996)**. Comparison of kriging and inverse-distance methods for mapping soil parameters. *Soil Science Society of America Journal*, **60** : 1237-1247.
450. **Goulard M. (1988)**. Champs spatiaux et statistique multidimensionnelle. Thèse de doctorat, Université des Sciences et Techniques du Languedoc - Montpellier 2, 177 p.
451. **Gower J.C. (1971)**. A general coefficient of similarity and some of its properties. *Biometrics*, **27** : 857-874.
452. **Gower J.C. & P. Legendre (1986)**. Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, **3** : 5-48.
453. **Graham R.L. (1972)**. An efficient algorithm for determining the convex hull of a finite planar set. *Information Processing Letters*, **1** : 132-133.
454. **Green C.D. (1973)**. A path entropy function for rooted trees. *Journal of the Association for Computing Machinery*, **20** : 378-384.
455. **Green P.J. & R. Sibson (1977)**. Computing Dirichlet tessellations in the plane. *Computer Journal*, **21** : 168-173.
456. **Greig-Smith P. (1952)**. The use of random and contiguous quadrats in the study of the structure of plant communities. *Annals of Botany*, **16** : 293-316.
457. **Gribko L.S., A.M. Liebhold & M.E. Hohn (1995)**. Model to predict Gypsy moth (Lepidoptera: Lymantriidae) defoliation using kriging and logistic regression. *Environmental Entomology*, **24** : 529-537.
458. **Griffith D.A., R. Haining & G. Arbia (1994)**. Heterogeneity of attribute sampling error in spatial data sets. *Geographical Analysis*, **26** : 300-320.
459. **Griffiths R.P., G.A. Bradshaw, B. Marks & G.W. Lienkaemper (1996)**. Spatial distribution of ectomycorrhizal mats in coniferous forests of the Pacific Northwest, USA. *Plant and Soil*, **180** : 147-158.
460. **Groleau P. & D. Marcotte (1997)**. The border effect of simulated annealing. *Mathematical Geology*, **29** : 585-592.
461. **Grötschel M. & L. Lovász (1995)**. Combinatorial optimization. *In* : Graham R.L., M. Grötschel & L. Lovász (Eds.), *Handbook of combinatorics*, vol. 2, pp. 1541-1597. Elsevier, Amsterdam, The Netherlands.
462. **Grundy P.M. & M.J.R. Healy (1950)**. Restricted randomization and quasi-Latin squares. *Journal of the Royal Statistical Society Series B*, **12**: 286-291.
463. **Guertal E.A. & C.B. Elkins (1996)**. Spatial variability of photosynthetically active radiation in a greenhouse. *Journal of the American Society for Horticultural Science*, **121** : 321-325.

464. **Guiblin P. (1997)**. Analyse géostatistique de campagnes (acoustique et chalutage) sur le hareng écossais. Thèse de doctorat, ENSMP, Fontainebleau, 81 p.

465. **Guillard J., D. Gerdeaux & G. Brun (1992)**. The use of geostatistics to analyse data from an echo-integration survey of fish stock Lake Sainte-Croix. *Fisheries Research*, **13** : 395.

466. **Guillobez S. & M. Arnaud (1998)**. Regionalized soil roughness indices. *Soil & Tillage Research*, **45** : 419-432.

467. **Gunnarsson F., S. Holm, P. Holmgren & T. Thuresson (1998)**. On the potential of kriging for forest management planning. *Scandinavian Journal of Forest Research*, **13** : 237-245.

468. **Gurland J. & P. Hinz (1971)**. Estimating parameters, testing fit, and analyzing untransformed data pertaining to the negative binomial and other distributions. *In* : Patil G.P., E.C. Pielou & W.E. Waters (Eds.), *Statistical Ecology. Volume 1. Spatial patterns and statistical distributions*, pp. 143-178. The Pennsylvania State University Press, University Park, Pennsylvania, USA.

469. **Gustafson E.J. (1998)**. Quantifying landscape spatial pattern: What is the state of the art? *Ecosystems*, **1** : 143-156.

470. **Guttorp P. (1994)**. Discussion of A.G. Journel "Resampling from stochastic simulations". *Environmental and Ecological Statistics*, **1** : 85-86.

471. **Haas T.C. (1990)**. Kriging and automated variogram modeling within a moving window. *Atmospheric Environment*, **24** : 1759-1769.

472. **Habib R., D. Tisne-Agostini, M.P. Vanniere & P. Monestiez (1991)**. Geostatistical method for independent sampling in kiwifruit vine to estimate yield components. *New Zealand Journal of Crop and Horticultural Science*, **19** : 329-335.

473. **Hacker R. (1962)**. Certification of algorithm 112. Position of point relative to polygon. *Communications of the ACM*, **5** : 606.

474. **Haining R. (1988)**. Estimating spatial means with an application to remotely sensed data. *Communications in Statistics - Theory and Methods*, **17** : 573-597.

475. **Haining R. (1991)**. Bivariate correlation with spatial data. *Geographical Analysis*, **23** : 210-227.

476. **Haining R. & G. Arbia (1993)**. Error propagation through map operations. *Technometrics*, **35** : 293-305.

477. **Haining R., J. Ma & S. Wise (1996)**. Design of a software system for interactive spatial statistical analysis linked to a GIS. *Computational Statistics Quarterly*, **11** : 449-466.

478. **Hall P. (1988)**. On confidence intervals for spatial parameters estimated from nonreplicated data. *Biometrics*, **44** : 271-277.

479. **Halvorson J.J., J.L. Smith, H. Bolton & R.E. Rossi (1995)**. Evaluating shrub-associated spatial patterns of soil properties in a shrub-steppe ecosystem using multiple-variable geostatistics. *Soil Science Society of America Journal*, **59** : 1476-1487.

480. **Halvorson J.J., J.L. Smith & R.I. Papendick (1996)**. Integration of multiple soil parameters to evaluate soil quality: a field example. *Biology and Fertility of Soils*, **21** : 207-214.

481. **Hamlett J.M., R. Horton & N.A.C. Cressie (1986)**. Resistant and exploratory techniques for use in semivariogram analyses. *Soil Science Society of America Journal*, **50** : 868-875.

482. **Hansen H.S. (1994)**. Spatial autocorrelation in vector-topological geographical information systems. *In* : Harts J.J., H.F.L. Ottens, H.J. Scholten & J. van Arragon (Eds.) *EGIS/MARI '94. Fifth european conference and exhibition on geographical information systems*, pp. 1252-1261. EGIS Foundation, Utrecht, The Netherlands.

483. **Hansen M.H. (1987)**. Some history and reminiscences on survey sampling. *Statistical Science*, **2** : 180-190.

484. **Hansen M.H. & W.N. Hurwitz (1943)**. On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, **14** : 333-362.

485. **Hansen M.H., W.G. Madow & B.J. Tepping (1983)**. An evaluation of model-dependent and probability-sampling inferences in samples surveys (with discussion). *Journal of the American Statistical Association*, **78** : 776-807.

486. **Hanski I. (1994a)**. Spatial scale, patchiness and population dynamics on land. *Philosophical Transactions of the Royal Society of London Series B*, **343** : 19-25.

487. **Hanski I. (1994b)**. Patch-occupancy dynamics in fragmented landscapes. *Trends in Ecology & Evolution*, **9** : 131-135.

488. **Hanski I., M. Kuussaari & M. Nieminen (1994)**. Metapopulation structure and migration in the butterfly *Melitaea cinxia*. *Ecology*, **75** : 747-762.

489. **Hanski I., T. Pakkala, M. Kuussaari & G. Lei (1995)**. Metapopulation persistence of an endangered butterfly in a fragmented landscape. *Oikos*, **72** : 21-28.

490. **Hartigan J.A. & M.A. Wong (1979)**. Algorithm AS 136. A k-means clustering algorithm. *Applied Statistics*, **28** : 100-108.

491. **Harvey P.K. (1981)**. A simple algorithm for the unique characterization of convex polygons. *Computers & Geosciences*, **7** : 387-392.

492. **Haslett J., R. Bradley, P. Craig, A. Unwin & G. Wills (1991)**. Dynamic graphics for exploring spatial data with application to locating global and local anomalies. *American Statistician*, **45** : 234-242.

493. **Haslett J.R. (1991)**. Species inventories as ecological databases in a Geographical Information System. *In* : *8th International colloquium of the European Invertebrate Survey*, pp. 41-44. Bruxelles, Belgique.

494. **Haslett J.R. (1994)**. The landscape approach in ecology. *Trends in Ecology & Evolution*, **9** : 486-487.

495. **Hastings A. (1994)**. Conservation and spatial structure: theoretical approaches. *In* : Levin S.A. (Ed.), *Frontiers in mathematical biology*, pp. 496-503. Springer-Verlag, Berlin, Germany.

496. **Hastings H.M. & G. Sugihara (1993)**. Fractals. A user's guide for the natural sciences. Oxford University Press, New York, USA, 235 p.

497. **Haton J.P., N. Bouzid, F. Charpillet, M.C. Haton, B. Lâasri, H. Lâasri, P. Marquis, T. Mondot & A. Napoli (1991)**. Le raisonnement en intelligence artificielle. Modèles, techniques et architectures pour les systèmes à bases de connaissances. InterEditions, Paris, France, 480 p.

498. **He F., P. Legendre & C. Bellehumeur (1994)**. Diversity pattern and spatial scale: a study of a tropical rain forest of Malaysia. *Environmental and Ecological Statistics*, **1** : 265-286.

499. **Hedayat A.S. & B.K. Sinha (1991)**. Design and inference in finite population sampling. Wiley, New York, USA, 377 p.

500. **Heil G.W. & W.P.A. van Deursen (1996)**. Searching for patterns and processes: modelling of vegetation dynamics with geographical information systems and remote sensing. *Acta Botanica Neerlandica*, **45** : 543-556.

501. **Heilbron D.C. (1978)**. Comparison of estimators of the variance of systematic sampling. *Biometrika*, **65** : 429-433.

502. **Heisel T., C. Andreasen & A.K. Ersboll (1996)**. Annual weed distributions can be mapped with kriging. *Weed Research*, **36** : 325-337.

503. **Hendricks-Franssen H.J.W.M.H., A.C. van Eijnsbergen & A. Stein (1997)**. Use of spatial prediction techniques and fuzzy classification for mapping soil pollutants. *Geoderma*, **77** : 243-262.

504. **Henley S. (1976)**. Autocorrelation coefficients from irregularly spaced areal data. *Computers & Geosciences*, **2** : 437-438.

505. **Herman J. (1986)**. Analyse de données qualitatives. 1. Traitement d'enquêtes, échantillons, répartitions, associations. Masson, Paris, France, 181 p.

506. **Hess G.R. & J.M. Bay (1997)**. Generating confidence intervals for composition-based landscape indexes. *Landscape Ecology*, **12** : 309-320.

507. **Heuvelink G.M. (1997)**. Discussion of D.J. Brus and J.J. de Gruijter "Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil". *Geoderma*, **80** : 49-51.

508. **Heuvelink G.M. & E.J. Pebesma (1999)**. Spatial aggregation and soil process modelling. *Geoderma*, **89** : 47-65.

509. **Hewett T.A. (1993)**. Modelling reservoir heterogeneity with fractals. *In* : Soares A. (Ed.), *Geostatistics Tróia '92*, pp. 455-466. Kluwer Academic Publishers, Dordrecht, The Netherlands.

510. **Hill D., P. Coquillard, J. Devaugelas & A. Meinesz (1998)**. An algorithmic model for invasive species: Application to *Caulerpa taxifolia* (Vahl) C. Agardh development in the North-Western Mediterranean Sea. *Ecological Modelling*, **109** : 251-265.

511. **Hill I.D. (1973)**. Algorithm AS 66. The normal integral. *Applied Statistics*, **22** : 424-427.

512. **Hill M.F. & H. Caswell (1999)**. Habitat fragmentation and extinction thresholds on fractal landscapes. *Ecology Letters*, **2** : 121-127.

513. **Hill M.O. (1973)**. The intensity of spatial pattern in plant communities. *Journal of Ecology*, **61** : 225-235.

514. **Hill M.O. (1991)**. Patterns of species distribution in Britain elucidated by canonical correspondence analysis. *Journal of Biogeography*, **18**: 247-255.

515. **Hinkley D. (1983)**. Jackknife methods. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 4, pp. 280-287. Wiley, New York, USA.

516. **Hinkley D.V. (1987)**. Discussion of G. Casella and R.L. Berger "Reconciling bayesian and frequentist evidence in the one-sided testing problem" and of J.O. Berger and T. Sellke "Testing a point null hypothesis: the irreconcilability of P values and evidence". *Journal of the American Statistical Association*, **82** : 128-129.

517. **Höck B.K., T.W. Payn & J.W. Shirley (1993)**. Using a geographic information system and geostatistics to estimate site index of *Pinus radiata* for Kaingaroa forest, New Zealand. *New Zealand Journal of Forestry Science*, **23** : 264-277.

518. **Hoeffding W. (1948)**. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, **19** : 293-325.

519. **Hoel P.G. (1943)**. The accuracy of sampling methods in ecology. *Annals of Mathematical Statistics*, **14** : 289-300.

520. **Hohn M.E., A.M. Liebhold & L.S. Gribko (1993)**. Geostatistical model for forecasting spatial dynamics of defoliation caused by the Gypsy Moth Lepidoptera: Lymantriidae. *Environmental Entomology*, **22** : 1066-1075.

521. **Holdaway M.R. (1996)**. Spatial modeling and interpolation of monthly temperature using kriging. *Climate Research*, **6** : 215-225.

522. **Holland J.H. (1962)**. Outline for a logical theory of adaptative systems. *Journal of the Association for Computing Machinery*, **9** : 297-314.

523. **Holland J.H. (1992)**. Les algorithmes génétiques. *Pour la Science*, **179** : 44-51.

524. **Holt R.D. (1997)**. From metapopulation dynamics to community structure. Some consequences of spatial heterogeneity. *In* : Hanski I. & M. Gilpin (Eds.), *Metapopulation biology*, pp. 149-164. Academic Press, London, UK.

525. **Hoosbeek M.R., A. Stein, H. Vanreuler & B.H. Janssen (1998)**. Interpolation of agronomic data from plot to field scale: using a clustered versus a spatially randomized block design. *Geoderma*, **81** : 265-280.

526. **Hopcroft J. & R. Tarjan (1974)**. Efficient planarity testing. *Journal of the Association for Computing Machinery*, **21** : 549-568.

527. **Horvitz D.G. & D.J. Thompson (1952)**. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47** : 663-685.

528. **Hosking J.R.M. (1986)**. Bonferroni inequalities and intervals. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 7, pp. 119-121. Wiley, New York, USA.

529. **Hossaert-McKey M., M. Valero, D. Magda, M. Jarry, J. Cuguen & P. Vernet (1996)**. The evolving genetic history of a population of *Lathyrus sylvestris*: evidence from temporal and spatial genetic structure. *Evolution*, **50** : 1808-1821.

530. **Hosseini E., J. Gallichand & D. Marcotte (1994)**. Theoretical and experimental performance of spatial interpolation methods for soil salinity analysis. *Transactions of the ASAE*, **37** : 1799-1807.

531. **Houle G. (1996)**. Environment filters and seedling recruitment on a coastal dune in subartic Quebec (Canada). *Canadian Journal of Botany*, **74** : 1507-1513.

532. **Houllier F. (1986)**. Echantillonnage et modélisation de la dynamique des peuplements forestiers. Application au cas de l'inventaire forestier national. Thèse de doctorat, Université Claude Bernard - Lyon 1, 267 p.

533. **Houllier F. (1992)**. Analyse et modélisation de la dynamique des peuplements forestiers. Applications à la gestion des ressources forestières. Habilitation à diriger des recherches, Université Claude Bernard - Lyon 1, 67 p.

534. **Howarth R.J. & S.A.M. Earle (1979)**. Application of a generalized power transformation to geochemical data. *Mathematical Geology*, **11** : 45-62.

535. **Hoyle M.H. (1973)**. Transformation. An introduction and a bibliography. *International Statistical Review*, **41** : 203-223.

536. **Huang C.W. & T.Y. Shih (1997)**. On the complexity of point-in-polygon algorithms. *Computers & Geosciences*, **23** : 109-118.

537. **Hubálek Z. (1982)**. Coefficients of association and similarity based on binary presence-absence data: an evaluation. *Biological Reviews*, **57** : 669-689.

538. **Hubert L.J. (1985)**. Combinatorial data analysis: association and partial association. *Psychometrika*, **50** : 449-467.

539. **Hubert L.J. & P. Arabie (1991)**. The assessment of spatial autocorrelation through constrained multiple regression. *Geographical Analysis*, **23** : 95-111.

540. **Hubert L.J. & R.G. Golledge (1982)**. Measuring association between spatially defined variables: Tjostheim's index and some extensions. *Geographical Analysis*, **14** : 273-278.

541. **Hubert L.J., R.G. Golledge & C.M. Costanzo (1981)**. Generalized procedures for evaluating spatial autocorrelation. *Geographical Analysis*, **13** : 224-233.

542. **Hubert L.J., R.G. Golledge, C.M. Costanzo & N. Gale (1985)**. Measuring association between spatially defined variables: an alternative procedure. *Geographical Analysis*, **17** : 36-46.

543. **Hudson G. (1993)**. Kriging temperature in Scotland using the external drift method. *In* : Soares A. (Ed.), *Geostatistics Tróia '92*, pp. 577-588. Kluwer Academic Publishers, Dordrecht, The Netherlands.

544. **Hudson G. & H. Wackernagel (1994)**. Mapping temperature using kriging with external drift: theory and an example from Scotland. *International Journal of Climatology*, **14** : 77-91.

545. **Hughes J.P. & D.P. Lettenmaier (1981)**. Data requirements for kriging: estimation and network design. *Water Resources Research*, **17** : 1641-1650.

546. **Hung M.S. & J.J. Divoky (1988)**. A computational study of efficient shortest path algorithms. *Computers and Operations Research*, **15** : 567-576.

547. **Hurlbert S.H. (1971)**. The nonconcept of species diversity: a critique and alternative parameters. *Ecology*, **52** : 577-586.

548. **Hurst R.L. & R.E. Knop (1972)**. Generation of random correlated normal variables. *Communications of the ACM*, **15** : 355-357.

549. **Hutchinson G.E. (1953)**. The concept of pattern in ecology. *Proceedings of the Academy of Natural Sciences of Philadephia*, **105** : 1-12.

550. **Iachan R. (1983)**. Asymptotic theory of systematic sampling. *Annals of Statistics*, **11** : 959-969.

551. **Iachan R. (1984)**. Sampling strategies, robustness and efficiency: the state of the art. *International Statistical Review*, **52** : 209-218.

552. **Iachan R. (1985)**. Plane sampling. *Statistics and Probability Letters*, **3** : 151-159.

553. **Ingber L. (1993)**. Simulated annealing: practice versus theory. *Mathematical and Computer Modelling*, **18** : 29-57.

554. **Isaaks E.H. & R.M. Srivastava (1988)**. Spatial continuity for probabilistic and deterministic geostatistics. *Mathematical Geology*, **20** : 313-341.

555. **Isaaks E.H. & R.M. Srivastava (1989)**. An introduction to applied geostatistics. Oxford University Press, New York, USA, 561 p.

556. **Ivanov A.B. (1995)**. Planigon. *In* : Hazewinkel M. (Ed.), *Encyclopedia of mathematics*, vol. 4, pp. 421-422. Kluwer Academic Publishers, Dordrecht, The Netherlands.

557. **Jackson D.A. & K.M. Somers (1989)**. Are probability estimates from the permutation model of Mantel's test stable? *Canadian Journal of Zoology*, **67** : 766-769.

558. **Jackson R.B. & M.M. Caldwell (1993a)**. The scale of nutrient heterogeneity around individual plants and its quantification with geostatistics. *Ecology*, **74** : 612-614.

559. **Jackson R.B. & M.M. Caldwell (1993b)**. Geostatistical patterns of soil heterogeneity around individual perennial plants. *Journal of Ecology*, **81** : 683-692.

560. **Jacquez G.M. (1995)**. The map comparison problem: tests for the overlap of geographic boundaries. *Statistics in Medicine*, **14** : 2343-2361.

561. **Jaggi S., D.A. Quattrochi & N.S.G. Lam (1993)**. Implementation and operation of three fractal measurement algorithms for analysis of remote-sensing data. *Computers & Geosciences*, **19** : 745-767.

562. **Janson S. & J. Vegelius (1981)**. Measures of ecological association. *Oecologia*, **49** : 371-376.

563. **Jarvis R.A. (1973)**. On the identification of the convex hull of a finite set of points in the plane. *Information Processing Letters*, **2** : 18-21.

564. **Jelinek H.F. & E. Fernandez (1998)**. Neurons and fractals: how reliable and useful are calculations of fractal dimensions? *Journal of Neuroscience Methods*, **81** : 9-18.

565. **Jian X., R.A. Olea & Y.S. Yu (1996)**. Semivariogram modeling by weighted least squares. *Computers & Geosciences*, **22** : 387-397.

566. **Joe B. (1993)**. Construction of k-dimensional Delaunay triangulation using local transformations. *SIAM Journal on Scientific Computing*, **14** : 1415-1436.

567. **Joffre R., S. Rambal & F. Romane (1996)**. Local variations of ecosystem functions in Mediterranean evergreen oak woodland. *Annales des Sciences Forestières*, **53** : 561-570.

568. **Johnson D.S., C.R. Aragon, L.A. McGeoch & C. Schevon (1989)**. Optimization by simulated annealing: an experimental evaluation. Part I. Graph partitioning. *Operation Research*, **37** : 865-892.

569. **Johnson G.A., D.A. Mortensen & C.A. Gotway (1996)**. Spatial and temporal analysis of weed seedling populations using geostatistics. *Weed Science*, **44** : 704-710.

570. **Johnson L.B. (1990)**. Analyzing spatial and temporal phenomena using geographical information systems. A review of ecological applications. *Landscape Ecology*, **4** : 31-43.

571. **Jonsson B.G. & J. Moen (1998)**. Patterns in species associations in plant communities: the importance of scale. *Journal of Vegetation Science*, **9** : 327-332.

572. **Jorge L.A.B. & G.J. Garcia (1997)**. A study of habitat fragmentation in Southeastern Brazil using remote sensing and geographic information systems (GIS). *Forest Ecology and Management*, **98** : 35-47.

573. **Joseph L. & B.K. Bhaumik (1997)**. Improved estimation of the Box-Cox transform parameter and its application to hydrogeochemical data. *Mathematical Geology*, **29** : 963-976.

574. **Journel A.G. (1983a)**. Nonparametric estimation of spatial distributions. *Mathematical Geology*, **15** : 445-468.

575. **Journel A.G. (1983b)**. Geostatistics. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 3, pp. 424-431. Wiley, New York, USA.

576. **Journel A.G. (1985)**. The deterministic side of geostatistics. *Mathematical Geology*, **17** : 1-15.

577. **Journel A.G. (1986a)**. Geostatistics: models and tools for the Earth sciences. *Mathematical Geology*, **18** : 119-140.

578. **Journel A.G. (1986b)**. Constrained interpolation and qualitative information. The soft kriging approach. *Mathematical Geology*, **18** : 269-286.

579. **Journel A.G. (1988)**. New distance measures: the route toward truly non-Gaussian geostatistics. *Mathematical Geology*, **20** : 459-475.

580. **Journel A.G. (1989)**. Fundamentals of geostatistics in five lessons. Short course in geology: volume 8. American Geophysical Union, Washington, USA, 38 p.

581. **Journel A.G. (1992)**. Comment on M. Armstrong "Positive definiteness is not enough". *Mathematical Geology*, **24** : 145-147.

582. **Journel A.G. (1994a)**. Discussion of C.A. Gotway "The use of conditional simulation in nuclear-waste-site performance assessment". *Technometrics*, **36** : 149-150.

583. **Journel A.G. (1994b)**. Resampling from stochastic simulations. *Environmental and Ecological Statistics*, **1** : 63-91.

584. **Journel A.G. (1996)**. Modelling uncertainty and spatial dependence: stochastic imaging. *International Journal of Geographical Information Systems*, **10** : 517-522.

585. **Journel A.G. & C.J. Huijbregts (1978)**. Mining geostatistics. Academic Press, London, UK, 600 p.

586. **Journel A.G. & M.E. Rossi (1989)**. When do we need a trend model in kriging? *Mathematical Geology*, **21** : 715-739.

587. **Juang K.W. & D.Y. Lee (1998a)**. Simple indicator kriging for estimating the probability of incorrectly delineating hazardous areas in a contaminated site. *Environmental Science & Technology*, **32** : 2487-2493.

588. **Juang K.W. & D.Y. Lee (1998b)**. A comparison of three kriging methods using auxiliary variables in heavy-metal contaminated soils. *Journal of Environmental Quality*, **27** : 355-363.

589. **Juang K.W., D.Y. Lee & C.K. Hsiao (1998)**. Kriging with cumulative distribution function of order statistics for delineation of heavy-metal contaminated soils. *Soil Science*, **163** : 797-804.

590. **Jumars P.A., D. Thistle & M.L. Jones (1977)**. Detecting two-dimensional spatial structure in biological data. *Oecologia*, **28** : 109-123.

591. **Kacewicz M. (1991)**. Solving the kriging problem by using the Gram-Schmidt orthogonalization. *Mathematical Geology*, **23** : 111-118.

592. **Kadmon R. & A. Danin (1997)**. Floristic variation in Israel: a GIS analysis. *Flora*, **192** : 341-345.

593. **Kalikhman I. & I. Ostrovsky (1997)**. Patchy distribution fields: survey design and adequacy of reconstruction. *ICES Journal of Marine Science*, **54**: 809-818.

594. **Kampichler C. (1999)**. Fractal concepts in studies of soil fauna. *Geoderma*, **88** : 283-300.

595. **Karakostas K.X. (1990)**. Exact optimum sampling designs for autocorrelated finite population. *Journal of Statistical Planning and Inference*, **24** : 353-361.

596. **Karakostas K.X. & H.P. Wynn (1989)**. Systematic sampling for autocorrelated superpopulations. *Journal of Statistical Planning and Inference*, **22** : 181-195.

597. **Karmanov V.G. (1995)**. Quadratic programming. *In* : Hazewinkel M. (Ed.), *Encyclopedia of mathematics*, vol. 4, pp. 638-638. Kluwer Academic Publishers, Dordrecht, The Netherlands.

598. **Kaufmann A. (1968)**. Introduction à la combinatorique en vue des applications. Dunod, Paris, France, 609 p.

599. **Keitt T.H. (1997)**. Stability and complexity on a lattice: coexistence of species in an individual-based food web model. *Ecological Modelling*, **102** : 243-258.

600. **Kemp C.D. (1971)**. Properties of some discrete ecological distributions. *In* : Patil G.P., E.C. Pielou & W.E. Waters (Eds.), *Statistical Ecology. Volume 1. Spatial patterns and statistical distributions*, pp. 1-22. The Pennsylvania State University Press, University Park, Pennsylvania, USA.

601. **Kempton R.A. & R.W.M. Wedderburn (1978)**. A comparison of three measures of species diversity. *Biometrics*, **34** : 25-37.

602. **Kenkel N.C., M.L. Hendrie & I.E. Bella (1997)**. A long-term study of *Pinus banksiana* population dynamics. *Journal of Vegetation Science*, **8** : 241-254.
603. **Kershaw K.A. (1961)**. Association and co-variance analysis of plant communities. *Journal of Ecology*, **49** : 643-654.
604. **Kido T., K. Takagi & M. Nakanishi (1994)**. Analysis and comparisons of genetic algorithm, simulated annealing, tabu search, and evolutionary combination algorithm. *Informatica*, **18** : 399-410.
605. **Kirkpatrick S., C.D. Gelatt & M.P. Vecchi (1983)**. Optimization by simulated annealing. *Science*, **220** : 671-680.
606. **Kitagawa T. (1974)**. Brainware concept in intelligent and integrated information system. Research Institute of Fundamental Information Science. Kyushu University, Japan, 10 p.
607. **Kitanidis P.K. (1983)**. Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water Resources Research*, **19**: 909-921.
608. **Kitanidis P.K. (1985)**. Minimum-variance unbiased quadratic estimation of covariances of regionalized variables. *Mathematical Geology*, **17** : 195-208.
609. **Kitanidis P.K. (1991)**. Orthonormal residuals in geostatistics: model criticism and parameter estimation. *Mathematical Geology*, **23** : 741-758.
610. **Kitron U., L.H. Otieno, L.L. Hungerford, A. Odulaja, W.U. Brigham, O.O. Okello, M. Joselyn, M.M. Mohamed-Ahmed & E. Cook (1996)**. Spatial analysis of the distribution of tsetse flies in the Lambwe Valley, Kenya, using Landsat TM satellite imagery and GIS. *Journal of Animal Ecology*, **65** : 371-380.
611. **Knotters M., D.J. Brus & J.H. Oude Voshaar (1995)**. A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. *Geoderma*, **67**: 227-246.
612. **Koch G.G. & D.B. Gillings (1983)**. Inference, design based vs. model based. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 4, pp. 84-88. Wiley, New York, USA.
613. **Kocherlakota S. & K. Kocherlakota (1983)**. Generalized variance. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 3, pp. 354-357. Wiley, New York, USA.
614. **Kodratoff Y. & E. Diday (1991)**. Induction symbolique et numérique à partir des données. Cépaduès-Editions, Toulouse, France, 460 p.
615. **Koenig W.D. (1998)**. Spatial autocorrelation in California land birds. *Conservation Biology*, **12** : 612-620.
616. **Koenig W.D. (1999)**. Spatial autocorrelation of ecological phenomena. *Trends in Ecology & Evolution*, **14** : 22-26.
617. **Koenig W.D. & J.M.H. Knops (1998)**. Testing for spatial autocorrelation in ecological studies. *Ecography*, **21** : 423-429.
618. **Köhl M. & G. Gertner (1997)**. Geostatistics in evaluating forest damage surveys: considerations on methods for describing spatial distributions. *Forest Ecology and Management*, **95** : 131-140.
619. **Kooijman S.A.L.M. (1976)**. Some remarks on the statistical analysis of grids especially with respect to ecology. *Annals of Systems Research*, **5** : 113-132.
620. **Koop J.C. (1971)**. On splitting a systematic sample for variance estimation. *Annals of Mathematical Statistics*, **42** : 1084-1087.
621. **Koop J.C. (1990)**. Systematic sampling of two-dimensional surfaces and related problems. *Communications in Statistics - Theory and Methods*, **19** : 1701-1750.

622. **Kostov C. & O. Dubrule (1986)**. An interpolation method taking into account inequality constraints. II. Practical approach. *Mathematical Geology*, **18** : 53-73.

623. **Kotz S. & N.L. Johnson (1982)**. Bias. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 1, pp. 230-231. Wiley, New York, USA.

624. **Kowalski C.J. (1972)**. On the effects of non-normality on the distribution of the sample product-moment correlation coefficient. *Applied Statistics*, **21** : 1-12.

625. **Kruskal J.B. (1956)**. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, **7** : 48-50.

626. **Kruskal W. & F. Mosteller (1979a)**. Representative sampling. I. Non-scientific literature. *International Statistical Review*, **47** : 13-24.

627. **Kruskal W. & F. Mosteller (1979b)**. Representative sampling. II. Scientific literature, excluding statistics. *International Statistical Review*, **47** : 111-127.

628. **Kruskal W. & F. Mosteller (1979c)**. Representative sampling. III. The current statistical literature. *International Statistical Review*, **47** : 245-265.

629. **Kruskal W. & F. Mosteller (1980)**. Representative sampling, IV: the history of the concept in statistics, 1895-1939. *International Statistical Review*, **48** : 169-195.

630. **Kruskal W. & F. Mosteller (1988)**. Representative sampling. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 8, pp. 77-81. Wiley, New York, USA.

631. **Kunin W.E. (1997)**. Sample shape, spatial scale and species counts: implications for reserve design. *Biological Conservation*, **82** : 369-377.

632. **Kuuluvainen T., E. Jarvinen, T.J. Hokkanen, S. Rouvinen & K. Heikkinen (1998)**. Structural heterogeneity and spatial autocorrelation in a natural mature *Pinus sylvestris* dominated forest. *Ecography*, **21** : 159-174.

633. **Lagacherie P. (1992)**. Formalisation des lois de distribution des sols pour automatiser la cartographie pédologique à partir d'un secteur pris comme référence. Cas de la petite région naturelle Moyenne Vallée de l'Hérault. Thèse de doctorat, Université des Sciences et Techniques du Languedoc - Montpellier 2, 175 p.

634. **Lagro J. (1991)**. Assessing patch shape in landscape mosaics. *Photogrammetric Engineering and Remote Sensing*, **57** : 285-293.

635. **Lajaunie C. (1990)**. Comparing some approximate methods for building local confidence intervals for predicting regionalized variables. *Mathematical Geology*, **22** : 123-144.

636. **Lam N.S.N. (1983)**. Spatial interpolation methods: a review. *American Cartographer*, **10** : 129-149.

637. **Lamorey G. & E. Jacobson (1995)**. Estimation of semivariogram parameters and evaluation of the effects of data sparsity. *Mathematical Geology*, **27** : 327-358.

638. **Lannou C. & S. Savary (1991)**. The spatial structure of spontaneous epidemics of different diseases in a groundnut plot. *Netherlands Journal of Plant Pathology*, **97** : 355-368.

639. **Lark R.M. & H.C. Bolam (1997)**. Uncertainty in prediction and interpretation of spatially variable data on soils. *Geoderma*, **77** : 263-282.

640. **Larkin B.J. (1991)**. An ANSI C program to determine in expected linear time the vertices of the convex hull of a set of planar points. *Computers & Geosciences*, **17** : 431-443.

641. **Larkin R.P., M.L. Gumpertz & J.B. Ristaino (1995)**. Geostatistical analysis of *Phytophthora* epidemic development in commercial bell pepper fields. *Phytopathology*, **85** : 191-203.

642. **Larvet P. (1994)**. Analyse des systèmes : de l'approche fonctionnelle à l'approche objet. InterEditions, Paris, France, 320 p.

643. **Laslett G.M. (1994)**. Kriging and splines: an empirical comparison of their predictive performance in some applications (with discussion). *Journal of the American Statistical Association*, **89** : 391-409.

644. **Laslett G.M. (1997)**. Discussion of D.J. Brus and J.J. de Gruijter "Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil". *Geoderma*, **80** : 45-49.

645. **Laslett G.M. & A.B. McBratney (1990)**. Further comparison of spatial methods for predicting soil pH. *Soil Science Society of America Journal*, **54** : 1553-1558.

646. **Laurini R. & F. Milleret-Raffort (1989)**. L'ingénierie des connaissances spatiales. Hermès, Paris, France, 63 p.

647. **Laurini R. & F. Milleret-Raffort (1993)**. Les bases de données en géomatique. Hermès, Paris, France, 340 p.

648. **Laurini R. & D. Thompson (1992)**. Fundamentals of spatial information systems. Academic Press, London, UK, 680 p.

649. **Lavado R.S., J.O. Sierra & P.N. Hashimoto (1996)**. Impact of grazing on soil nutrients in a Pampean grassland. *Journal of Range Management*, **49** : 452-457.

650. **Le Boulengé E., P. Legendre, C. de Le Court, P. Le Boulenge-Nguyen & M. Languy (1996)**. Microgeographic morphological differentiation in Muskrats. *Journal of Mammalogy*, **77** : 684-701.

651. **Le Corre V., S. Dumolin-Lapegue & A. Kremer (1997)**. Genetic variation at allozyme and RAPD loci in sessile oak *Quercus petraea* (Matt) Liebl: the role of history and geography. *Molecular Ecology*, **6** : 519-529.

652. **Le Corre V., G. Roussel, A. Zanetto & A. Kremer (1998)**. Geographical structure of gene diversity in *Quercus petraea*(Matt.) Liebl. III. Patterns of variation identified by geostatistical analyses. *Heredity*, **80** : 464-473.

653. **Le Moigne J.L. (1984)**. L'informatique est-elle une science? *AFCET/Interfaces*, **17** : 3-7.

654. **Lebreton J.D., D. Chessel, R. Prodon & N. Yoccoz (1988a)**. L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. I. Variables de milieu quantitatives. *Acta Oecologica - Oecologia Generalis*, **9** : 53-67.

655. **Lebreton J.D., D. Chessel, M. Richardot-Coulet & N. Yoccoz (1988b)**. L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. II. Variables de milieu qualitatives. *Acta Oecologica - Oecologia Generalis*, **9** : 137-151.

656. **Lecoustre R. & P. de Reffye (1986)**. La théorie des variables régionalisées : ses applications possibles dans le domaine épidémiologique aux recherches agronomiques en particulier sur le palmier à l'huile et le cocotier. *Oléagineux*, **41** : 541-548.

657. **Lecoustre R., D. Fargette, C. Fauquet & P. de Reffye (1989)**. Analysis and mapping of the spatial spread of african Cassava mosaic virus using geostatistics and the kriging technique. *Phytopathology*, **79** : 913-920.

658. **Ledreux C., R. Jeansoulin, D. King & P. Lagacherie (1994)**. Un raisonnement spatial symbolique-numérique pour la reconnaissance d'unités de sol dans SAPRISTI. *In*: Servigne S. & R. Laurini (Eds.) *Les journées de la recherche Cassini*, pp. 35-44. LISI, INSA, Villeurbanne, France.

659. **Leduc A., Y.T. Prairie & Y. Bergeron (1994)**. Fractal dimension estimates of a fragmented landscape: sources of variability. *Landscape Ecology*, **9** : 279-286.

660. **Lee D.T. & F.P. Preparata (1984)**. Computational geometry. A survey. *IEEE Transactions on Computers*, **33** : 1072-1101.

661. **Lee Y.M. & J.H. Ellis (1997)**. Estimation and simulation of lognormal random fields. *Computers & Geosciences*, **23** : 19-31.

662. **Lefèbvre J., H. Roussel, E. Walter, D. Lecointe & W. Tabbara (1996)**. Prediction from wrong models: the kriging approach. *IEEE Antennas and Propagation Magazine*, **38** : 35-45.

663. **Legay J.M. (1986a)**. Quelques réflexions à propos d'écologie : défense de l'indisciplinarité. *Acta Oecologica - Oecologia Generalis*, **7** : 391-398.

664. **Legay J.M. (1986b)**. Qu'est-ce que la biométrie. *Courrier du CNRS*, **64** : 56-61.

665. **Legay J.M. & D. Debouzie (1985)**. Introduction à une biologie des populations. Masson, Paris, France, 149 p.

666. **Legendre P. (1993)**. Spatial autocorrelation: trouble or new paradigm? *Ecology*, **74** : 1659-1673.

667. **Legendre P. & M.J. Fortin (1989)**. Spatial pattern and ecological analysis. *Vegetatio*, **80** : 107-138.

668. **Legendre L. & P. Legendre (1984a)**. Ecologie numérique. Tome 1. Le traitement multiple des données écologiques. Deuxième édition. Masson, Paris, France, 260 p.

669. **Legendre L. & P. Legendre (1984b)**. Ecologie numérique. Tome 2. La structure des données écologiques. Masson, Paris, France, 335 p.

670. **Legendre P. & B.H. McArdle (1997)**. Comparison of surfaces. *Oceanologica Acta*, **20** : 27-41.

671. **Legendre P. & M. Troussellier (1988)**. Aquatic heterotrophic bacteria: modeling in the presence of spatial autocorrelation. *Limnology and Oceanography*, **33** : 1055-1067.

672. **Legendre P., M. Troussellier, V. Jarry & M.J. Fortin (1989)**. Design for simultaneous sampling of ecological variables: from concepts to numerical solutions. *Oikos*, **55** : 30-42.

673. **Legendre P., N.L. Oden, R.R. Sokal, A. Vaudor & J. Kim (1990)**. Approximate analysis of variance of spatially autocorrelated regional data. *Journal of Classification*, **7** : 53-75.

674. **Legendre P., S.F. Thrush, V.J. Cummings, P.K. Dayton, J. Grant, J.E. Hewitt, A.H. Hines, B.H. McArdle, R.D. Pridmore, D.C. Schneider, S.J. Turner, R.B. Whitlatch & M.R. Wilkinson (1997)**. Spatial structure of bivalves in a sandflat: scale and generating processes. *Journal of Experimental Marine Biology and Ecology*, **216** : 99-128.

675. **Levin S.A. (1992)**. The problem of pattern and scale in ecology. *Ecology*, **73** : 1943-1967.

676. **Li B.L., Y.T. Chu & D.K. Loh (1993)**. Event probability correlation analysis for comparison of two-phase ecological maps. *Ecological Modelling*, **69** : 287-302.

677. **Li H. & J.F. Reynolds (1993)**. A new contagion index to quantify spatial patterns of landscapes. *Landscape Ecology*, **8** : 155-162.

678. **Li H. & J.F. Reynolds (1994)**. A simulation experiment to quantify spatial heterogeneity in categorical maps. *Ecology*, **75** : 2446-2455.

679. **Li K.J. (1992)**. Contributions aux systèmes d'hypermédia : modélisation et indexation des objets spatio-temporels. Thèse de doctorat, Institut National des Sciences Appliquées de Lyon, 191 p.

680. **Liang Q.W., I.S. Otvos & G.E. Bradfield (1997)**. Forest roadside sampling of larvae and adults of the western hemlock looper, *Lambdina fiscellaria lugubrosa*. *Forest Ecology and Management*, **93** : 45-53.

681. **Liebhold A.M., R.E. Rossi & W.P. Kemp (1993)**. Geostatistics and Geographic Information Systems in applied insect ecology. *Annual Review of Entomology*, **38** : 303-327.

682. **Liebhold A.M., X. Zhang, M.E. Hohn, J.S. Elkinton, M. Ticehurst, G.L. Benzon & R.W. Campbell (1991)**. Geostatistical analysis of Gypsy Moth (Lepidoptera: Lymantriidae) egg mass populations. *Environmental Entomology*, **20**: 1407-1417.

683. **Lieth H. & G.W. Moore (1971)**. Computerized clustering of species in phytosociological tables and its utilization for field work. *In* : Patil G.P., E.C. Pielou & W.E. Waters (Eds.), *Statistical Ecology. Volume 1. Spatial patterns and statistical distributions*, pp. 403-422. The Pennsylvania State University Press, University Park, Pennsylvania, USA.

684. **Limic N. & A. Mikelic (1984)**. Constrained kriging using quadratic programming. *Mathematical Geology*, **16** : 423-429.

685. **Lin F.T., C.Y. Kao & C.C. Hsu (1993)**. Applying the genetic approach to simulated annealing in solving some NP-hard problems. *IEEE Transactions on Systems, Man, and Cybernetics*, **23** : 1752-1767.

686. **Linden D.R. & D.M. van Doren (1986)**. Parameters for characterizing tillage-induced soil surface roughness. *Soil Science Society of America Journal*, **50** : 1560-1565.

687. **Lindgren G. & I. Rychlik (1995)**. How relaible are contour curves? Confidence sets for level contours. *Bernoulli*, **1** : 301-319.

688. **Lindsay B.G. (1985)**. Nuisance parameters. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 6, pp. 373-376. Wiley, New York, USA.

689. **Liski J. & C.J. Westman (1997)**. Carbon storage in forest soil in Finland. 2. Size and regional patterns. *Biogeochemistry*, **36** : 261-274.

690. **Little L.S., D. Edwards & D.E. Porter (1997)**. Kriging in estuaries: as the crow flies, or as the fish swims? *Journal of Experimental Marine Biology and Ecology*, **213** : 1-11.

691. **Liu L.J.S. & A.J. Rossini (1996)**. Use of kriging models to predict 12-hour mean ozone concentrations in metropolitan Toronto. A pilot study. *Environment International*, **22** : 677-692.

692. **Liu R. & L.P. Herrington (1996)**. The expected cost of uncertainty in geographic data. *Journal of Forestry*, **94** : 27-31.

693. **Loehle C. & B.L. Li (1996)**. Statistical properties of ecological and geological fractals. *Ecological Modelling*, **85** : 271-284.

694. **Long A.E. & D.E. Myers (1997)**. A new form of the cokriging equations. *Mathematical Geology*, **29** : 685-703.

695. **Longley M., P.C. Jepson, J. Izquierdo & N. Sotherton (1997)**. Temporal and spatial changes in aphid and parasitoid populations following applications of deltamethrin in winter wheat. *Entomologica Experimentalis et Applicata*, **83** : 41-52.

696. **Longley P.A. & M. Batty (1989a)**. Fractal measurement and line generalization. *Computers & Geosciences*, **15** : 167-183.

697. **Longley P.A. & M. Batty (1989b)**. On the fractal measurement of geographical boundaries. *Geographical Analysis*, **21** : 47-67.

698. **Lord E.A. & C.B. Wilson (1986)**. The mathematical description of shape and form. Ellis Horwood, Chichester, UK, 260 p.

699. **Lovejoy S. (1982)**. Area-perimeter relation for rain and cloud areas. *Science*, **216** : 185-187.

700. **Ludwig J.A. & D.W. Goodall (1978)**. A comparison of paired-with-blocked-quadrat variance methods for the analysis of spatial pattern. *Vegetatio*, **38** : 49-59.

701. **Luo J. & B.J. Fox (1996)**. A review of the Mantel test in dietary studies: effect of sample size and inequality of samples sizes. *Wildlife Research*, **23** : 267-288.

702. **Luque S.S., R.G. Lathrop & J.A. Bognar (1994)**. Temporal and spatial changes in an area of the New Jersey Pine Barrens landscape. *Landscape Ecology*, **9** : 287-300.

703. **MacDonald G.M. & N.M. Waters (1988)**. The use of most predictable surfaces for the classification and mapping of taxon assemblages. *Vegetatio*, **74**: 125-135.

704. **Mackas D.L. (1984)**. Spatial autocorrelation on plankton community composition in a continental shelf ecosystem. *Limnology and Oceanography*, **29** : 451-471.

705. **MacLaren N.M. (1992)**. A limit on the usable length of a pseudorandom sequence. *Journal of Statistical Computation and Simulation*, **42** : 47-54.

706. **Madow L.H. (1946)**. Systematic sampling and its relation to other sampling designs. *Journal of the American Statistical Association*, **41** : 204-217.

707. **Madow W.G. (1949)**. On the theory of systematic sampling. II. *Annals of Mathematical Statistics*, **20** : 333-354.

708. **Madow W.G. (1953)**. On the theory of systematic sampling. III. Comparison of centered and random start systematic sampling. *Annals of Mathematical Statistics*, **24** : 101-106.

709. **Madow W.G. & L.H. Madow (1944)**. On the theory of systematic sampling. I. *Annals of Mathematical Statistics*, **15** : 1-24.

710. **Mahy G. & G. Neve (1997)**. The application of spatial autocorrelation methods to the study of *Calluna vulgaris* population genetics. *Belgian Journal of Botany*, **129** : 131-139.

711. **Malinen T. & H. Peltonen (1996)**. Optimal sampling and traditional versus model-based data analysis in acoustic fish stock assessment in Lake Vesijarvi. *Fisheries Research*, **26** : 295-308.

712. **Maling D.H. (1989)**. Measurements from maps. Principles and methods of cartometry. Pergamon Press, Oxford, UK, 577 p.

713. **Mallet J.L. (1974)**. Présentation d'un ensemble de méthodes et techniques de la cartographie automatique numérique. *Sciences de la Terre, Série Informatique Géologique*, **4** : 1-213.

714. **Mallet J.L. (1980)**. Régression sous contraintes linéaires : application au codage des variables aleatoires. *Revue de Statistique Appliquée*, **28** : 57-68.

715. **Mandelbrot B.B. (1967)**. How long is the coast of Britain? Statistical self-similarity and fractional dimension. *Science*, **156** : 636-638.

716. **Mandelbrot B.B. (1998)**. Is nature fractal? *Science*, **279** : 783.

717. **Mandelbrot B.B., D.E. Passoja & A.J. Paullay (1984)**. Fractal character of fracture surfaces of metals. *Nature*, **308** : 721-722.

718. **Manly B.F.J. (1986)**. Randomization and regression methods for testing for association with geographical, environmental and biological distances between populations. *Researches on Population Ecology*, **28** : 201-218.

719. **Manly B.F.J. (1991)**. Randomization and Monte Carlo methods in biology. Chapman & Hall, London, UK, 281 p.

720. **Manly B.F.J. (1993)**. Randomization, bootstrap and Monte Carlo methods in biology. Second edition. Chapman & Hall, London, UK, 399 p.

721. **Manly B.F.J. (1995)**. A note on the analysis of species co-occurrences. *Ecology*, **76** : 1109-1115.

722. **Mantel N. (1967)**. The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27** : 209-220.

723. **Mantel N. & R.S. Valand (1970)**. A technique of nonparametric multivariate analysis. *Biometrics*, **26** : 547-558.

724. **Mantoglou A. (1987)**. Digital simulation of multivariate two- and three-dimensional stochastic processes with a spectral turning bands method. *Mathematical Geology*, **19** : 129-149.

725. **Mantoglou A. & J.L. Wilson (1982)**. The turning bands method for simulation of random fields using line generation by a spectral method. *Water Resources Research*, **18** : 1379-1394.

726. **Maravelias C.D., D.G. Reid, E.J. Simmonds & J. Haralabous (1996)**. Spatial analysis and mapping of acoustic survey data in the presence of high local variability: geostatistical application to North Sea herring (*Clupea harengus*). *Canadian Journal of Fisheries and Aquatic Sciences*, **53** : 1497-1505.

727. **Marcotte D. (1995)**. Generalized cross-validation for covariance model. *Mathematical Geology*, **27** : 659-672.

728. **Marcotte D. (1996)**. Fast variogram computation with FFT. *Computers & Geosciences*, **22** : 1175-1186.

729. **Marcotte D. & M. David (1988)**. Trend surface analysis as a special case of IRF-k Kriging. *Mathematical Geology*, **20** : 821-824.

730. **Marcus L.F. & J.H. Vandermeer (1966)**. Regional trends in geographic variation. *Systematic Zoology*, **15** : 1-13.

731. **Mardia K.V. (1980)**. Some statistical inference problems in kriging. *Sciences de la Terre, Série Informatique Géologique*, **15** : 113-131.

732. **Mardia K.V. & R.J. Marshall (1984)**. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, **71**: 135-146.

733. **Mardia K.V. & A.J. Watkins (1989)**. On multimodality of the likelihood in the spatial linear model. *Biometrika*, **76** : 289-295.

734. **Margalef R. (1979)**. The organization of space. *Oikos*, **33** : 152-159.

735. **Mark D.M. (1978)**. Topological properties of geographic surfaces: applications in computer cartography. *In* : *Harvard papers on Geographic Information Systems*, pp. 1-12. Cambridge, Massachusetts, USA.

736. **Marsaglia G. (1972)**. The structure of linear congruential sequences. *In* : Zaremba S.K. (Ed.), *Applications of number theory to numerical analysis*, pp. 249-285. Academic Press, New York, USA.

737. **Marsaglia G. (1985)**. A current view of random number generators. *In* : *Computer science and statistics: the interface*, pp. 3-10. Elsevier, New York, USA.

738. **Marshall B., B. Boag, J.W. McNicol & R. Neilson (1998)**. A comparison of the spatial distributions of three plant-parasitic nematode species at three different scales. *Nematologica*, **44** : 303-320.

739. **Martin R.J. (1979)**. A subclass of lattice processes applied to a problem in planar sampling. *Biometrika*, **66** : 209-217.

740. **Masini G., A. Napoli, D. Colnet, D. Leonard & K. Tombre (1990)**. Les langages à objets. Langages de classes, langages de frames, langages d'acteurs. InterEditions, Paris, France, 584 p.

741. **Mason D.C., M. O'Conaill & I. McKendrick (1994)**. Variable resolution block kriging using a hierarchical spatial data structure. *International Journal of Geographical Information Systems*, **8** : 429-449.

742. **Matérn B. (1960)**. Spatial variation. Meddelanden från Statens Skogsforskningsinstitut, 49. Second edition (1986). Springer-Verlag, Berlin, Germany, 151 p.

743. **Matheron G. (1963)**. Principles of geostatistics. *Economic Geology*, **58** : 1246-1266.

744. **Matheron G. (1965)**. Les variables régionalisées et leur estimation. Une application de la théorie des fonctions aléatoires aux sciences de la nature. Masson, Paris, France, 305 p.

745. **Matheron G. (1969)**. Le krigeage universel. ENSMP, Fontainebleau, France, 83 p.

746. **Matheron G. (1970)**. Structures aléatoires et géologie mathématique. *Revue de l'Institut International de Statistique*, **38** : 1-11.

747. **Matheron G. (1973)**. The intrinsic random functions and their applications. *Advances in Applied Probability*, **5** : 439-468.

748. **Matheron G. (1978)**. Estimer et choisir. Essai sur la pratique des probabilités. ENSMP, Paris, France, 175 p.

749. **Matheron G. (1982)**. Pour une analyse krigeante des données régionalisées. ENSMP, Fontainebleau, France, 22 p.

750. **Matheron G. (1986)**. Philipian/Watsonian high (flying) philosophy. *Mathematical Geology*, **18** : 503-504.

751. **Matheron G. (1987)**. A simple answer to an elementary question. *Mathematical Geology*, **19** : 455-457.

752. **Mathieu M. (1990)**. Glossaire de cartographie. Comité Français de Cartographie, Paris, France, 171 p.

753. **Matloff N.S. (1980)**. Algorithm AS 148. The Jackknife. *Applied Statistics*, **29** : 115-117.

754. **Matula D.W. & R.R. Sokal (1980)**. Properties of Gabriel graphs relevant to geographic variation research and the clustering of points in the plane. *Geographical Analysis*, **12** : 205-222.

755. **Maurer B.A. (1994)**. Geographical population analysis: tools for the analysis of biodiversity. Blackwell, Oxford, UK, 130 p.

756. **Maurin J. (1975)**. Simulation déterministe du hasard. Masson, Paris, France, 122 p.

757. **Maus A. (1984)**. Delaunay triangulation and the convex hull of n points in expected linear time. *Bit*, **24** : 151-163.

758. **Maxwell J.C. (1870)**. On hills and dales. *Philosophical Magazine*, **40** : 421-427.

759. **Mayer D.G., J.A. Belward & K. Burrage (1998)**. Tabu search not an optimal choice for models of agricultural systems. *Agricultural Systems*, **58** : 243-251.

760. **McBratney A.B. & I.O.A. Odeh (1997)**. Application of fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions. *Geoderma*, **77** : 85-113.

761. **McBratney A.B. & R. Webster (1983a)**. How many observations are needed for regional estimation of soil properties? *Soil Science*, **135** : 177-183.

762. **McBratney A.B. & R. Webster (1983b)**. Optimal interpolation and isarithmic mapping of soil properties. V. Co-regionalization and multiple sampling strategy. *Journal of Soil Science*, **34** : 137-162.

763. **McBratney A.B. & R. Webster (1986)**. Choosing functions for semi-variograms of soil properties and fitting them to sampling estimates. *Journal of Soil Science*, **37** : 617-639.

764. **McBratney A.B., R. Webster & T.M. Burgess (1981)**. The design of optimal sampling schemes for local estimation and mapping of regionalized variables. I. Theory and method. *Computers & Geosciences*, **7** : 331-334.

765. **McCallum D. & D. Avis (1979)**. A linear algorithm for finding the convex hull of a simple polygon. *Information Processing Letters*, **9** : 201-206.

766. **McCarn D.W. & J.R. Carr (1992)**. Influence of numerical precision and equation solution algorithm on computation of kriging weights. *Computers & Geosciences*, **18** : 1127-1167.

767. **McCarty H.H. & N.E. Salisbury (1961)**. Visual comparison of isopleth maps as a means of determining correlations between spatially distributed phenomena. Department of Geography, State University of Iowa, Iowa City, USA, p.

768. **McFadden C.S. & K.Y. Aydin (1996)**. Spatial autocorrelation analysis of small-scale genetic structure in a clonal soft coral with limited larval dispersal. *Marine Biology*, **126** : 215-224.

769. **McKenney D.W., R.S. Rempel, L.A. Venier, Y.H. Wang & A.R. Bisset (1998)**. Development and application of a spatially explicit moose population model. *Canadian Journal of Zoology*, **76** : 1922-1931.

770. **Mead R. (1971)**. Models for interplant competition in irregularly distributed populations. *In* : Patil G.P., E.C. Pielou & W.E. Waters (Eds.), *Statistical Ecology. Volume 2. Sampling and modeling biological populations and population dynamics*, pp. 13-32. The Pennsylvania State University Press, University Park, Pennsylvania, USA.

771. **Mead R. (1974)**. A test for spatial pattern at several scales using data from a grid of contiguous quadrats. *Biometrics*, **30** : 295-307.

772. **Meija J. & I. Rodríguez-Iturbe (1974)**. On the synthesis of random fields from the spectrum: an application to the generation of hydrologic spatial processes. *Water Resources Research*, **10** : 705-711.

773. **Mendès-France M. (1984)**. Folding paper and thermodynamics. *Physics Reports - Review Section on Physics Letters*, **103** : 161-172.

774. **Mendès-France M. (1987)**. Dimension et entropie des courbes régulières. *In* : Cherbit G. (Ed.), *Fractals. Dimensions non entières et applications*, pp. 329-339. Masson, Paris, France.

775. **Méot A., P. Legendre & D. Borcard (1998)**. Partialling out the spatial component of ecological variation: questions and propositions in the linear modelling framework. *Environmental and Ecological Statistics*, **5**: 1-27.

776. **Mercer W.B. & A.D. Hall (1911)**. The experimental error of field trials. *Journal of Agricultural Science*, **4** : 107-132.

777. **Mercier F. (1997)**. Analyse et modélisation de la dynamique forestière guyanaise à l'aide de diagrammes de Voronoï. Thèse de doctorat, Université Claude Bernard - Lyon 1, 181 p.

778. **Merks J.W. (1992)**. Geostatistics or voodoo statistics? *Engineering and Mining Journal*, **193** : 45-49.

779. **Merriam D.F. & P.H.A. Sneath (1966)**. Quantitative comparison of contour maps. *Journal of Geophysical Research*, **71** : 1105-1115.

780. **Merriam G. (1988)**. Landscape dynamics in Farmland. *Trends in Ecology & Evolution*, **3** : 16-20.

781. **Metropolis N. & S. Ulam (1949)**. The Monte Carlo method. *Journal of the American Statistical Association*, **44** : 335-341.

782. **Metropolis N., A.W. Rosenbluth, M.N. Rosenbluth & A.H. Teller (1953)**. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, **21** : 1087-1092.

783. **Metzger J.P. & E. Muller (1996)**. Characterizing the complexity of landscape boundaries by remote sensing. *Landscape Ecology*, **11** : 65-77.

784. **Miclet L. (1984)**. Méthodes structurelles pour la reconnaissance des formes. Eyrolles, Paris, France, 184 p.

785. **Midgarden D.G., R.R. Youngman & S.J. Fleischer (1993)**. Spatial analysis of counts of Western Corn Rootworm (Coleoptera: Chrysomelidae) adults on yellow sticky traps in corn: geostatistics and dispersion indices. *Environmental Entomology*, **22** : 1124-1133.

786. **Mielke P.W. (1978)**. Clarification and appropriate inferences for Mantel and Valand's nonparametric multivariate analysis technique. *Biometrics*, **34** : 277-282.

787. **Miller R.G. (1974)**. The jackknife: a review. *Biometrika*, **61** : 1-15.

788. **Milne A. (1959)**. The centric systematic area-sample treated as a random sample. *Biometrics*, **15** : 270-297.

789. **Milne B.T. (1988)**. Measuring the fractal geometry of landscapes. *Applied Mathematics and Computation*, **27** : 67-79.

790. **Milne B.T. (1991)**. Lessons from applying fractal models to landscape patterns. *In* : Turner M.G. & R.H. Gardner (Eds.), *Quantitative methods in landscape ecology*, pp. 199-235. Springer-Verlag, New York, USA.

791. **Milne B.T. (1992)**. Spatial aggregation and neutral models in fractal landscapes. *American Naturalist*, **139** : 32-57.

792. **Minns C.K., C.N. Bakelaar, J.E. Moore, R.W. Dermott & R. Green (1996)**. Measuring differences between overlapping but unpaired spatial surveys using a geographic information system. *Environmental Monitoring and Assessment*, **43** : 237-253.

793. **Moilanen A. & I. Hanski (1998)**. Metapopulation dynamics: effects of habitat quality and landscape structure. *Ecology*, **79** : 2503-2515.

794. **Monmonier M.S. (1974)**. Measures of pattern complexity for choroplethic maps. *American Cartographer*, **1** : 159-169.

795. **Moran P.A.P. (1948)**. The interpretation of statistical maps. *Journal of the Royal Statistical Society Series B*, **10** : 243-251.

796. **Moran P.A.P. (1950)**. Notes on continuous stochastic phenomena. *Biometrika*, **37** : 17-23.

797. **Morkoc A., J.W. Biggar, D.R. Nielsen & D.E. Myers (1987)**. Kriging with generalized covariances. *Soil Science Society of America Journal*, **51** : 1126-1131.

798. **Morris M.D. (1991)**. On counting the number of data pairs for semivariogram estimation. *Mathematical Geology*, **23** : 929-943.

799. **Morse S.P. (1968)**. A mathematical model for the analysis of contour-line data. *Journal of the Association for Computing Machinery*, **15** : 205-220.

800. **Morse S.P. (1969)**. Concepts of use in contour map processing. *Communications of the ACM*, **12** : 147-152.

801. **Mowrer H.T. (1997)**. Propagating uncertainty through spatial estimation processes for old-growth subalpine forests using sequential Gaussian simulation in GIS. *Ecological Modelling*, **98** : 73-86.

802. **Mukherjee D. & M.V. Ratnaparkhi (1986)**. On the functional relationship between entropy and variance with related applications. *Communications in Statistics - Theory and Methods*, **15** : 291-311.

803. **Murray A.W.A. (1996)**. Comparison of geostatistical and random sample survey analyses of Antarctic krill acoustic data. *ICES Journal of Marine Science*, **53** : 415-421.

804. **Murray J.D. (1988)**. Spatial dispersal of species. *Trends in Ecology & Evolution*, **3** : 307-309.

805. **Myers D.E. (1982)**. Matrix formulation of co-kriging. *Mathematical Geology*, **14** : 249-257.

806. **Myers D.E. (1983)**. Estimation of linear combinations and co-kriging. *Mathematical Geology*, **15** : 633-637.

807. **Myers D.E. (1988)**. Interpolation with positive definite functions. *Sciences de la Terre, Série Informatique Géologique*, **28** : 251-265.

808. **Myers D.E. (1989)**. To be or not to be stationary? That is the question. *Mathematical Geology*, **21** : 347-362.

809. **Myers D.E. (1991a)**. Interpolation and estimation with spatially located data. *Chemometrics and Intelligent Laboratory Systems*, **11** : 209-228.

810. **Myers D.E. (1991b)**. Pseudo-cross variograms, positive-definiteness and cokriging. *Mathematical Geology*, **23** : 805-816.

811. **Myers D.E. (1994a)**. Spatial interpolation: an overview. *Geoderma*, **62** : 17-28.

812. **Myers D.E. (1994b)**. Discussion of D. Borcard and P. Legendre "Environmental control and spatial structure in ecological communities: an example using oribatid mites (Acari, Oribatei)". *Environmental and Ecological Statistics*, **1** : 53-55.

813. **Myers D.E. (1994c)**. Comment on Z. Şen "Cumulative semivariogram models of regionalized variables" and "Standard cumulative semivariograms of stationary stochastic processes and regional correlation". *Mathematical Geology*, **26** : 415-416.

814. **Myers J.C. (1997)**. Geostatistical error management. Quantifying uncertainty for environmental sampling and mapping. Van Nostrand Reinhold, New York, USA, 571 p.

815. **Myers D.E. & A. Journel (1990)**. Variograms with zonal anisotropies and noninvertible kriging systems. *Mathematical Geology*, **22** : 779-785.

816. **Myers W., G.P. Patil & K. Joly (1997)**. Echelon approach to areas of concern in synoptic regional monitoring. *Environmental and Ecological Statistics*, **4** : 131-152.

817. **Nelson M.R., R. Felix-Gastelum, T.V. Orum, L.J. Stowell & D.E. Myers (1994)**. Geographic information systems and geostatistics in the design and validation of regional plant virus management programs. *Phytopathology*, **84** : 898-905.

818. **Nelson M.R., T.V. Orum, R. Jaime-Garcia & A. Nadeem (1999)**. Applications of geographic information systems and geostatistics in plant disease epidemiology and management. *Plant Disease*, **83** : 308-319.

819. **Nguyen N.K. & A.J. Miller (1992)**. A review of some exchange algorithms for constructing discrete D-optimal designs. *Computational Statistics & Data Analysis*, **14** : 489-498.

820. **Nicholson M.C. & T.N. Mather (1996)**. Methods for evaluating lyme disease risks using geographic information systems and geospatial analysis. *Journal of Medical Entomology*, **33** : 711-720.

821. **Niederreiter H. (1978)**. Quasi-Monte Carlo methods and pseudo-random numbers. *Bulletin of the American Mathematical Society*, **84** : 957-1041.

822. **Noreen E.W. (1989)**. Computer-intensive methods for testing hypotheses: an introduction. Wiley, New York, USA, 229 p.

823. **Odeh I.O.A., A.B. McBratney & D.J. Chittleborough (1990)**. Design of optimal sample spacings for mapping soil using fuzzy-k-means and regionalized variable theory. *Geoderma*, **47** : 93-122.

824. **Odeh I.O.A., A.B. McBratney & D.J. Chittleborough (1995)**. Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. *Geoderma*, **67** : 215-226.

825. **Oden N.L. (1984)**. Assessing the significance of a spatial correlogram. *Geographical Analysis*, **16** : 1-16.

826. **Oden N.L. & R.R. Sokal (1986)**. Directional autocorrelation: an extension of spatial correlograms to two dimensions. *Systematic Zoology*, **35** : 608-617.

827. **Oden N.L. & R.R. Sokal (1992)**. An investigation of three-matrix permutation tests. *Journal of Classification*, **9** : 275-290.

828. **Oden N.L., R.R. Sokal, M.J. Fortin & H. Goebl (1993)**. Categorical wombling: detecting regions of significant change in spatially located categorical variables. *Geographical Analysis*, **25** : 315-336.

829. **O'Dowd R.F. (1991)**. Conditioning of coefficient matrices of ordinary kriging. *Mathematical Geology*, **23** : 721-739.

830. **Okabe A. (1976)**. A note on Geary's spatial contiguity ratio. *Geographical Analysis*, **8** : 315-318.

831. **Okx J.P., H. Leenaers & R.M. Krzanowski (1993)**. Probability kriging as a decision support tool for local soil pollution problems. *In* : Soares A. (Ed.), *Geostatistics Tróia '92*, pp. 673-683. Kluwer Academic Publishers, Dordrecht, The Netherlands.

832. **Olea R.A. (1984)**. Sampling design optimization for spatial functions. *Mathematical Geology*, **16** : 369-392.

833. **Olea R.A. (1992)**. Kriging. Understanding allays intimidation. *Geobyte*, **Oct. 92** : 12-17.

834. **Olea R.A. & V. Pawlowsky (1996)**. Compensating for estimation smoothing in kriging. *Mathematical Geology*, **28** : 407-417.

835. **Oleschko K., F. Brambila, F. Aceff & L.P. Mora (1998)**. From fractal analysis along a line to fractals on the plane. *Soil Tillage Research*, **45** : 389-406.

836. **Oliver D.S. (1995)**. Moving averages for Gaussian simulation in two and three dimensions. *Mathematical Geology*, **27** : 939-960.

837. **Oliver M.A. (1992)**. Some novel geostatistical applications in soil science. *In* : Bárdossy A. (Ed.) *Geostatistical methods: recent developments and applications in surface and subsurface hydrology*, pp. 142-153. UNESCO, Paris, France.

838. **Oliver M.A. & R. Webster (1986)**. Combining nested and linear sampling for determining the scale and form of spatial variation of regionalized variables. *Geographical Analysis*, **18** : 227-242.

839. **Oliver M.A. & R. Webster (1989)**. A geostatistical basis for spatial weighting in multivariate classification. *Mathematical Geology*, **21** : 15-35.

840. **Oliver M.A. & R. Webster (1991)**. How geostatistics can help you. *Soil Use and Management*, **7** : 206-217.

841. **Oliver M., R. Webster & J. Gerrard (1989a)**. Geostatistics in physical geography. Part I. Theory. *Transactions of the Institute British Geographers*, **14** : 259-269.

842. **Oliver M., R. Webster & J. Gerrard (1989b)**. Geostatistics in physical geography. Part II. Applications. *Transactions of the Institute British Geographers*, **14** : 270-286.

843. **Oliver M.A., R. Webster, C. Lajaunie, K.R. Muir, S.E. Parkes, A.H. Cameron, M.C.G. Stevens & J.R. Mann (1998)**. Binomial cokriging for estimating and mapping the risk of childhood cancer. *IMA Journal of Mathematics Applied in Medicine and Biology*, **15** : 279-297.

844. **Olsen A. (1994)**. Discussion of A.G. Journel "Resampling from stochastic simulations". *Environmental and Ecological Statistics*, **1** : 89.

845. **O'Neill R.V., J.R. Krummel, R.H. Gardner, G. Sugihara, B. Jackson, D.L. Deangelis, B.T. Milne, M.G. Turner, B. Zygmunt, S.W. Christensens, V.H. Dale & R.L. Graham (1988)**. Indices of landscape pattern. *Landscape Ecology*, **1** : 153-162.

846. **Ooi B.C. (1990)**. Efficient query processing in geographic information systems. Springer-Verlag, Berlin, Germany, 208 p.

847. **Ord J.K. (1985)**. Pearson system of distributions. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 6, pp. 655-659. Wiley, New York, USA.

848. **Ord J.K. (1988)**. Spatial processes. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 8, pp. 575-581. Wiely, New York, USA.

849. **Ord J.K. & A. Getis (1995)**. Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis*, **27** : 286-306.

850. **Ore O. (1970)**. Les graphes et leurs applications. Dunod, Paris, France, 144 p.

851. **O'Rourke J. (1982)**. Computing the relative neighborhood graph in the L_1 and L_∞ metrics. *Pattern Recognition*, **15** : 189-192.

852. **Orum T.V., D.M. Bigelow, M.R. Nelson, D.R. Howell & P.J. Cotty (1997)**. Spatial and temporal patterns of *Aspergillus flavus* strain composition and propagule density in Yuma County, Arizona, soils. *Plant Disease*, **81** : 911-916.

853. **Osborne J.G. (1942)**. Sampling errors of systematic and random surveys of cover-type areas. *Journal of the American Statistical Association*, **37**: 256-264.

854. **Overmars M.H. & J. van Leeuwen (1980)**. Further comments on Bykat's convex hull algorithm. *Information Processing Letters*, **10** : 209-212.

855. **Palmer M.W. (1988)**. Fractal geometry: a tool for describing spatial patterns of plant communities. *Vegetatio*, **75** : 91-102.

856. **Pan G. (1994)**. Restricted kriging: a link between sample value and sample configuration. *Mathematical Geology*, **26** : 135-155.

857. **Papageorgiou I. & K.X. Karakostas (1998)**. On optimal sampling designs for autocorrelated finite populations. *Biometrika*, **85** : 482-486.

858. **Papritz A., H.R. Kunsch & R. Webster (1993)**. On the pseudo cross-variogram. *Mathematical Geology*, **25** : 1015-1026.

859. **Papritz A. & R. Webster (1995a)**. Estimating temporal change in soil monitoring. I. Statistical theory. *European Journal of Soil Science*, **46** : 1-12.

860. **Papritz A. & R. Webster (1995b)**. Estimating temporal change in soil monitoring. II. Sampling from simulated fields. *European Journal of Soil Science*, **46** : 13-27.

861. **Pardo-Igúzquiza E. (1997a)**. MLREML: a computer program for the inference of spatial covariance parameters by maximum likelihood and restricted maximum likelihood. *Computers & Geosciences*, **23** : 153-162.

862. **Pardo-Igúzquiza E. (1997b)**. GCINFE: a computer program for inference of polynomial generalized covariance functions. *Computers & Geosciences*, **23** : 163-174.

863. **Pardo-Igúzquiza E. (1998a)**. Maximum likelihood estimation of spatial covariance parameters. *Mathematical Geology*, **30** : 95-108.

864. **Pardo-Igúzquiza E. (1998b)**. Optimal selection of number and location of rainfall gauges for areal rainfall estimation using geostatistics and simulated annealing. *Journal of Hydrology*, **210** : 206-220.

865. **Pardo-Igúzquiza E. & P.A. Dowd (1998)**. Maximum likelihood inference of spatial covariance parameters of soil properties. *Soil Science*, **163** : 212-219.

866. **Pariente D. (1994a)**. Définition et manipulation de champs continus contraints statistiquement et morphologiquement. *In* : Servigne S. & R. Laurini (Eds.) *Les journées de la recherche Cassini*, pp. 231-240. LISI, INSA, Villeurbanne, France.

867. **Pariente D. (1994b)**. Geographical interpolation and extrapolation by means of neural networks. *In* : Harts J.J., H.F.L. Ottens, H.J. Scholten & J. van Arragon (Eds.) *EGIS/MARI '94. Fifth european conference and exhibition on geographical information systems*, pp. 684-693. EGIS Foundation, Utrecht, The Netherlands.

868. **Pasquier B. (1987)**. Cartographie numérique des domaines. Structuration, modélisation, algorithmique. Thèse de doctorat d'Etat, Université Pierre & Marie Curie - Paris 6, 389 p.

869. **Paulsen J. (1984)**. Impact of random number generators in time series Monte Carlo simulation. *Journal of Statistical Computation and Simulation*, **19**: 23-33.

870. **Pearl J. (1990)**. Heuristique. Stratégies de recherche intelligente pour la résolution de problèmes par ordinateur. Cépaduès-Editions, Toulouse, France, 383 p.

871. **Pelletier D. & A.M. Parma (1994)**. Spatial distribution of pacific Halibut (*Hippoglossus stenolepis*): an application of geostatistics to longline survey data. *Canadian Journal of Fisheries and Aquatic Sciences*, **51** : 1506-1518.

872. **Pennycuick C.J. & N.C. Kline (1986)**. Units of measurment for fractal extent: applied to the coastal distribution of bald eagle nests in the Aleutian Islands (Alaska). *Oecologia*, **68** : 254-258.

873. **Perry J.N. (1995)**. Spatial analysis by distance indices. *Journal of Animal Ecology*, **64** : 303-314.

874. **Persicani D. (1995)**. Evaluation of soil classification and kriging for mapping herbicide leaching simulated by two models. *Soil Technology*, **8**: 17-30.

875. **Petersen G.W., J.C. Bell, K. McSweeney, G.A. Nielsen & P.C. Robert (1995)**. Geographic information systems in agronomy. *Advances in Agronomy*, **55** : 67-111.

876. **Petitgas P. (1991)**. Contributions géostatistiques à la biologie des pêches maritimes. Thèse de doctorat, ENSMP, Fontainebleau, 157 p.

877. **Petitgas P. (1993)**. Geostatistics for fish stock assessments: a review and an acoustic application. *ICES Journal of Marine Science*, **50** : 285-298.

878. **Petitgas P. & J.J. Levenez (1996)**. Spatial organization of pelagic fish: echogram structure, spatio-temporal condition, and biomass in Senegalese waters. *ICES Journal of Marine Science*, **53** : 147-153.

879. **Pettitt A.N. (1982)**. Durbin-Watson test. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 2, pp. 426-428. Wiley, New York, USA.

880. **Pettitt A.N. & A.B. McBratney (1993)**. Sampling designs for estimating spatial variance components. *Applied Statistics*, **42** : 185-209.

881. **Pfeifer P. (1998)**. Is nature fractal? *Science*, **279** : 785-785.

882. **Pfleiderer S., D.G.A. Ball & R.C. Bailey (1993)**. Auto: a computer program for the determination of the two-dimensional autocorrelation function of digital images. *Computers & Geosciences*, **19** : 825-829.

883. **Philip G.M. & D.F. Watson (1986)**. Matheronian geostatistics. Quo vadis? *Mathematical Geology*, **18** : 93-117.

884. **Philip G.M. & D.F. Watson (1987)**. An open letter to G. Matheron. *Mathematical Geology*, **19** : 453-454.

885. **Phillips D.L. & D.G. Marks (1996)**. Spatial uncertainty analysis: propagation of interpolation errors in spatially distributed models. *Ecological Modelling*, **91** : 213-229.

886. **Phillips J.D. (1985)**. Measuring complexity of environmental gradients. *Vegetatio*, **64** : 95-102.

887. **Phinn S., J. Franklin, A. Hope, D. Stow & L. Huenneke (1996)**. Biomass distribution mapping using airbone digital video imagery and spatial statistics in a semi-arid environment. *Journal of Environmental Management*, **47** : 139-164.

888. **Piazza A., P. Menozzi & L. Cavalli-Sforza (1981)**. The making and testing of geographic gene-frequency maps. *Biometrics*, **37** : 635-659.

889. **Pielou E.C. (1969)**. An introduction to mathematical ecology. Wiley, New York, USA, 279 p.

890. **Pont D. (1986)**. Structure spatiale d'une population du Cyclopide *Acanthocyclops robustus* dans une rizière de Camargue (France). *Acta Oecologica - Oecologia Generalis*, **7** : 289-302.

891. **Porter D.E., D. Edwards, G. Scott, B. Jones & W.S. Street (1997)**. Assessing the impacts of anthropogenic and physiographic influences on grass shrimp in localized salt-marsh estuaries. *Aquatic Botany*, **58** : 289-306.

892. **Posa D. (1989)**. Conditioning of the stationary kriging matrices for some well-known covariance models. *Mathematical Geology*, **21** : 755-765.

893. **Posa D. & D. Marcotte (1992)**. Robustness of kriging weights to non-bias conditions. *Mathematical Geology*, **24** : 759-773.

894. **Postaire J.G. (1987)**. De l'image à la décision. Dunod, Paris, France, 186 p.

895. **Preparata F.P. & M.I. Shamos (1985)**. Computational geometry. An introduction. Springer-Verlag, New York, USA, 398 p.

896. **Press W.H., B.P. Flannery, S.A. Teukolsky & W.T. Vetterling (1989)**. Numerical recipes in Pascal. The art of scientific computing. Cambridge University Press, Cambridge, USA, 759 p.

897. **Pressense L. & F. Salge (1987)**. Les bases de données localisées à l'Institut Géographique National. *La Jaune et la Rouge*, **426** : 45-50.

898. **Prim R.C. (1957)**. Shortest connection networks and some generalizations. *Bell System Technical Journal*, **36** : 1389-1401.

899. **Prins C. (1994)**. Algorithmes de graphes. Avec programmes en Pascal. Eyrolles, Paris, France, 382 p.

900. **Prodon R. & J.D. Lebreton (1994)**. Analyses multivariées des relations espèces-milieu : structure et interprétation écologique. *Vie Milieu*, **44** : 69-91.

901. **Prokhorov A.V. (1995)**. Stochastic equivalence. *In* : Hazewinkel M. (Ed.), *Encyclopedia of mathematics*, vol. 5, p. 375. Kluwer Academic Publishers, Dordrecht, The Netherlands.

902. **Pyper B.J. & R.M. Peterman (1998a)**. Comparison of methods to account for autocorrelation in correlation analyses of fish data. *Canadian Journal of Fisheries and Aquatic Sciences*, **55** : 2127-2140.

903. **Pyper B.J. & R.M. Peterman (1998b)**. Erratum to "Comparison of methods to account for autocorrelation in correlation analyses of fish data" [Can. J. Fisheries Aquat. Sci. 55 (1998) 2127-2140]. *Canadian Journal of Fisheries and Aquatic Sciences*, **55** : 2710.

904. **Qi Y. & J. Wu (1996)**. Effects of changing spatial resolution on the results of landscape pattern analysis using spatial autocorrelation indices. *Landscape Ecology*, **11** : 39-49.
905. **Quenouille M.H. (1949)**. Problems in plane sampling. *Annals of Mathematical Statistics*, **20** : 355-375.
906. **Quenouille M. (1956)**. Notes on bias in estimation. *Biometrika*, **43**: 353-360.
907. **Rahman S., L.C. Munn, R. Zhang & G.F. Vance (1996)**. Rocky mountain forest soils: evaluating spatial variability using conventional statistics ans geostatistics. *Canadian Journal of Soil Science*, **76** : 501-507.
908. **Ramstein G. & M. Raffy (1990)**. Algorithme d'analyse fractale de contours en télédétection et applications. *International Journal of Remote Sensing*, **11** : 191-208.
909. **Reeve R. (1992)**. A warning about standard errors when estimating the fractal dimension. *Computers & Geosciences*, **18** : 89-91.
910. **Reich R.M., R.L. Czaplewski & W.A. Bechtold (1994)**. Spatial cross-correlation of undisturbed, natural shortleaf pine stands in northern Georgia. *Environmental and Ecological Statistics*, **1** : 201-217.
911. **Rhynsburger D. (1973)**. Analytic delineation of Thiessen polygon. *Geographical Analysis*, **5** : 133-144.
912. **Richardson S. & D. Hemon (1981)**. On the variance of the sample correlation between two independent lattice processes. *Journal of Applied Probability*, **18** : 943-948.
913. **Rigaut J.P. (1987)**. Fractals, semi-fractals et biométrie. *In* : Cherbit G. (Ed.), *Fractals. Dimensions non entières et applications*, pp. 231-281. Masson, Paris, France.
914. **Rigaut J.P., D. Schoevaert-Brossault, A.M. Downs & G. Landini (1998)**. Asymptotic fractals in the context of grey-scale images. *Journal of Microscopy - Oxford*, **189** : 57-63.
915. **Riolo R. (1992)**. La recherche de l'absolu. *Pour la Science*, **179** : 101-103.
916. **Riordan J. (1958)**. An introduction to combinatorial analysis. Wiley, New York, USA, 244 p.
917. **Ripley B.D. (1981)**. Spatial statistics. Wiley, New York, USA, 252 p.
918. **Ripley B.D. (1983)**. Computer generation of random variables: a tutorial. *International Statistical Review*, **51** : 301-319.
919. **Ripley B.D. (1987)**. Stochastic simulation. Wiley, New York, USA, 237 p.
920. **Ripley B.D. (1988a)**. Statistical inference for spatial processes. Cambridge University Press, New York, USA, 148 p.
921. **Ripley B.D. (1988b)**. Spatial data analysis. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 8, pp. 570-573. Wiley, New York, USA.
922. **Rivoirard J. (1990)**. A review of lognormal estimators for in situ reserves. *Mathematical Geology*, **22** : 213-221.
923. **Rivoirard J. (1991)**. Introduction au krigeage disjonctif et à la géostatistique non linéaire. ENSMP, Fontainebleau, France, 99 p.
924. **Rivoirard J. (1994)**. Introduction to disjunctive kriging and non-linear geostatistics. Clarendon Press, London, UK, 182 p.
925. **Rizzo D.M. & D.E. Dougherty (1994)**. Characterization of aquifer properties using artificial neural networks: neural kriging. *Water Resources Research*, **30** : 483-497.
926. **Roach D.E. & A.D. Fowler (1993)**. Dimensionality analysis of patterns: fractal measurements. *Computers & Geosciences*, **19** : 849-869.
927. **Robertson G.P. (1987)**. Geostatistics in ecology: interpolation with known variance. *Ecology*, **68** : 744-748.

928. **Robertson G.P., M.A. Huston, F.C. Evans & J.M. Tiedje (1988)**. Spatial variability in a successional plant community: patterns of nitrogen availability. *Ecology*, **69** : 1517-1524.

929. **Robinson G.K. (1982)**. Confidence intervals and regions. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 2, pp. 120-127. Wiley, New York, USA.

930. **Robson D.S. (1982)**. Ecological statistics. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 2, pp. 434-440. Wiley, New York, USA.

931. **Rodriguez M.A. & W.M. Lewis (1997)**. Structure of fish assemblages along environmental gradients in floodplain lakes of the Orinoco river. *Ecological Monographs*, **67** : 109-128.

932. **Rodriguez R.N. (1982)**. Correlation. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 2, pp. 193-204. Wiley, New York, USA.

933. **Rosenbaum M.S. & M. Söderström (1996)**. Cokriging of heavy metals as an aid to biogeochemical mapping. *Acta Agriculturae Scandinavica Section B - Soil and Plant Science*, **46** : 1-18.

934. **Rosenfeld A. (1983)**. Quadtrees and pyramids: hierarchical representation of images. *In* : Haralick R.M. (Ed.), *Pictorial data analysis*, pp. 29-42. Springer-Verlag, Berlin, Germany.

935. **Rossi J.P. (1996)**. Statistical tool for soil biology. 11. Autocorrelogram and Mantel test. *European Journal of Soil Biology*, **32** : 195-203.

936. **Rossi J.P., L. Delaville & P. Quenehervé (1996)**. Microspatial structure of a plant-parasitic nematode community in a sugarcane field in Martinique. *Applied Soil Ecology*, **3** : 17-26.

937. **Rossi J.P. & P. Quenehervé (1998)**. Relating species density to environmental variables in presence of spatial autocorrelation: a study case on soil nematodes distribution. *Ecography*, **21** : 117-123.

938. **Rossi J.P., P. Lavelle & J.E. Tondoh (1995)**. Statistical tool for soil biology. X. geostatistical analysis. *European Journal of Soil Biology*, **31** : 173-181.

939. **Rossi R.E., P.W. Borth & J.J. Tollefson (1993)**. Stochastic simulation for characterizing ecological spatial patterns and appraising risk. *Ecological Applications*, **3** : 719-735.

940. **Rossi R.E., D.J. Mulla, A.G. Journel & E.H. Franz (1992)**. Geostatistical tools for modeling and interpreting ecological spatial dependence. *Ecological Monographs*, **62** : 277-314.

941. **Rouhani S. (1985)**. Variance reduction analysis. *Water Resources Research*, **21** : 837-846.

942. **Rouhani S. & T.J. Hall (1988)**. Geostatistical schemes for groundwater sampling. *Journal of Hydrology*, **103** : 85-102.

943. **Roux M. (1985)**. Algorithmes de classification. Masson, Paris, France, 151 p.

944. **Royall R.M. (1970)**. On finite population sampling theory under certain linear regression models. *Biometrika*, **57** : 377-387.

945. **Royall R.M. (1983)**. Comment on M.H. Hansen *et al.* "An evaluation of model-dependent and probability-sampling inferences in samples surveys". *Journal of the American Statistical Association*, **78** : 794-796.

946. **Royaltey H.H., E. Astrachan & R.R. Sokal (1975)**. Tests for patterns in geographic variation. *Geographical Analysis*, **7** : 369-395.

947. **Rozenberg G. & A. Salomaa (1995)**. Complexity theory. *In* : Hazewinkel M. (Ed.), *Encyclopedia of mathematics*, vol. 1, pp. 768-771. Kluwer Academic Publishers, Dordrecht, The Netherlands.

948. **Rubinstein R.Y. (1981)**. Simulation and the Monte Carlo method. Wiley, New York, USA, 278 p.

949. **Russo D. (1984)**. Design of an optimal sampling network for estimating the variogram. *Soil Science Society of America Journal*, **48** : 708-716.

950. **Russo D. & W.A. Jury (1987)**. A theoretical study of the estimation of the correlation scale in spatially variable fields. I. Stationary fields. *Water Resources Research*, **23** : 1257-1268.

951. **Russo D. & W.A. Jury (1988)**. Effect of the sampling network on estimates of the covariance function of stationary fields. *Soil Science Society of America Journal*, **52** : 1228-1234.

952. **Ruuska R. & J. Helenius (1996)**. GIS analysis of change in a agricultural landscape in Central Finland. *Agricultural and Food Science in Finland*, **5**: 567-576.

953. **Sacks J. & S. Schiller (1988)**. Spatial designs. *In* : Gupta S.S. & J.O. Berger (Eds.), *Statistical decision theory and related topics IV*, pp. 385-399. Springer-Verlag, New York, USA.

954. **Sakarovitch M. (1984)**. Optimisation combinatoire. Méthodes mathématiques et algorithmiques. Programmation Discrète. Hermann, Paris, France, 269 p.

955. **Salomon K.B. (1978)**. An efficient point-in-polygon algorithm. *Computers & Geosciences*, **4** : 173-178.

956. **Samet H. (1981)**. Connected component labeling using quadtrees. *Journal of the Association for Computing Machinery*, **28** : 487-501.

957. **Samet H. (1990)**. Applications of spatial data structures. Computer graphics, image processing, and GIS. Addison-Wesley, Reading, Massachusetts, USA, 505 p.

958. **Santaló L.A. (1953)**. Introduction to integral geometry. Hermann, Paris, France, 127 p.

959. **Sapozhenko A.A. (1995)**. Graph, connectivity of a. *In* : Hazewinkel M. (Ed.), *Encyclopedia of mathematics*, vol. 2, pp. 873-875. Kluwer Academic Publishers, Dordrecht, The Netherlands.

960. **Särndal C.E. (1978)**. Design-based and model-based inference in survey sampling. *Scandinavian Journal of Statistics*, **5** : 27-52.

961. **Särndal C.E., B. Swensson & J.H. Wretman (1992)**. Model assisted survey sampling. Springer-Verlag, New York, USA, 694 p.

962. **Sarnelle O., K.W. Kratz & S.D. Cooper (1993)**. Effects of an invertebrate grazer on the spatial arrangement of a benthic microhabitat. *Oecologia*, **96**: 208-218.

963. **Saupe D. (1988)**. Algorithms for random fractals. *In* : Peitgen H.O. & D. Saupe (Eds.), *The science of fractal images*, pp. 71-136. Springer-Verlag, New York, USA.

964. **Scheiner S.M. (1992)**. Measuring pattern diversity. *Ecology*, **73** : 1860-1867.

965. **Scherrer B. (1983)**. Techniques de sondage en écologie. *In* : Frontier S. (Ed.), *Stratégies d'échantillonnage en écologie*, pp. 63-162. Masson, Paris, France.

966. **Schlesinger W.H., J.A. Raikes, A.E. Hartley & A.F. Cross (1996)**. On the spatial pattern of soil nutrients in desert ecosystems. *Ecology*, **77** : 364-374.

967. **Schotzko D.J. & L.E. O'Keeffe (1989)**. Geostatistical description of the spatial distribution of *Lygus hesperus (*Heteroptera: Miridae) in lentils. *Journal of Economic Entomology*, **82** : 1277-1288.

968. **Schotzko D.J. & L.E. O'Keeffe (1990)**. Effect of sample placement on the geostatistical analysis of the spatial distribution of *Lygus hesperus* (Heteroptera: Miridae) in lentils. *Journal of Economic Entomology*, **83**: 1888-1900.

969. **Schreiber T. (1991)**. A Voronoi diagram based adaptative k-means-type clustering algorithm for multidimensional weighted data. *In* : Bieri H. & H. Noltemeier (Eds.), *Computational geometry. Methods, algorithms and applications*, pp. 265-275. Springer-Verlag, Berlin, Germany.

970. **Schumaker L.L. (1993)**. Computing optimal triangulations using simulated annealing. *Computer Aided Geometric Design*, **10** : 329-345.

971. **Schwertman N.C. & D.M. Allen (1979)**. Smoothing an indefinite variance-covariance matrix. *Journal of Statistical Computation and Simulation*, **9** : 183-194.

972. **Sedgewick R. (1991)**. Algorithmes en langage C. InterEditions, Paris, France, 685 p.

973. **Selkow S.M. (1977)**. The tree-to-tree editing problem. *Information Processing Letters*, **6** : 184-186.

974. **Sen A. (1976)**. Large sample-size distribution of statistics used in testing for spatial correlation. *Geographical Analysis*, **8** : 175-184.

975. **Sen A. (1977)**. "Large sample-size distribution of statistics used in testing for spatial correlation" : a reply. *Geographical Analysis*, **9** : 300.

976. **Şen Z. (1989)**. Cumulative semivariogram models of regionalized variables. *Mathematical Geology*, **21** : 891-903.

977. **Şen Z. (1992)**. Standard cumulative semivariograms of stationary stochastic processes and regional correlation. *Mathematical Geology*, **24** : 417-435.

978. **Şen Z. (1994)**. Reply to comments by Donald E. Myers. *Mathematical Geology*, **26** : 417-418.

979. **Sénéchal B. (1979)**. Groupes et géométries. Hermann, Paris, France, 125 p.

980. **Serfling R.J. (1988)**. U-statistics. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 9, pp. 436-444. Wiley, New York, USA.

981. **Serra J. (1982)**. Image analysis and mathematical morphology. Volume 1. Academic Press, London, UK, 610 p.

982. **Shafer J.M. & M.D. Varljen (1990)**. Approximation of confidence limits on sample semivariograms from single realizations of spatially correlated random fields. *Water Resources Research*, **26** : 1787-1802.

983. **Shapiro A. & J.D. Botha (1991)**. Variogram fitting with a general class of conditionally nonnegative definite functions. *Computational Statistics & Data Analysis*, **11** : 87-96.

984. **Sharov A.A., A.M. Liebhold & E.A. Roberts (1996)**. Spatial variation among counts of Gypsy Moths (Lepidoptera: Lymantriidae) in pheromone-baited traps at expanding population fronts. *Environmental Entomology*, **25**: 1312-1320.

985. **Sharov A.A., E.A. Roberts, A.M. Liebhold & F.W. Ravlin (1995)**. Gypsy moth (Lepidoptera: Lymantriidae) spread in the central appalachians: three methods for species boundary estimation. *Environmental Entomology*, **24** : 1529-1538.

986. **Sharp W.E. & C. Bays (1992)**. A review of portable random number generators. *Computers & Geosciences*, **18** : 79-87.

987. **Shen Q. (1994)**. An application of GIS to the measurement of spatial autocorrelation. *Computers Environment and Urban Systems*, **18** : 167-191.

988. **Sheshinski R. (1979)**. Interpolation in the plane: the robustness to misspecified correlation models and different trend functions. *In* : Patil G.P. & M. Rosenzweig (Eds.), *Contemporary Quantitative Ecology and Related Ecometrics*, pp. 399-420. International Co-operative Publishing House, Fairland, Maryland, USA.

989. **Shimrat M. (1962)**. Algorithm 112. Position of point relative to polygon. *Communications of the ACM*, **5** : 434.

990. **Shinozuka M. & C.M. Jan (1972)**. Digital simulation of random processes and its applications. *Journal of Sound and Vibration*, **25** : 111-128.

991. **Shiryaev A.N. (1995)**. Stochastic indistinguishability. *In* : Hazewinkel M. (Ed.), *Encyclopedia of mathematics*, vol. 5, p. 377. Kluwer Academic Publishers, Dordrecht, The Netherlands.

992. **Shook K. & D.M. Gray (1996)**. Small-scale spatial structure of shallow snowcovers. *Hydrological Processes*, **10** : 1283-1292.

993. **Siarry P. & G. Dreyfus (1988)**. La méthode du recuit simulé. IDSET, Paris, France, 125 p.

994. **Simard Y., P. Legendre, G. Lavoie & D. Marcotte (1992)**. Mapping, estimating biomass, and optimizing sampling programs for spatially autocorrelated data: case study of the northern shrimp (*Pandalu borealis*). *Canadian Journal of Fisheries and Aquatic Sciences*, **49** : 32-45.

995. **Simard Y., D. Marcotte & G. Bourgault (1993)**. Exploration of geostatistical methods for mapping and estimating acoustic biomass of pelagic fish in the Gulf of St. Lawrence: size of echo-integration unit and auxiliary environmental variables. *Aquatic Living Resources*, **6** : 185-199.

996. **Simon G. (1997)**. An angular version of spatial correlations, with exact significance tests. *Geographical Analysis*, **29** : 267-278.

997. **Sircar J.K. & J.A. Cebrian (1991)**. An automated approach for labeling raster digitized contour maps. *Photogrammetric Engineering and Remote Sensing*, **57** : 965-971.

998. **Slatkin M. & H.E. Arter (1991)**. Spatial autocorrelation methods in population genetics. *American Naturalist*, **138** : 499-517.

999. **Smith J.L., J.J. Halvorson & R.I. Papendick (1993)**. Using multiple-variable indicator kriging for evaluating soil quality. *Soil Science Society of America Journal*, **57** : 743-749.

1000. **Smith T.M.F. (1976)**. The foundation of survey sampling: a review (with discussion). *Journal of the Royal Statistical Society Series A*, **139** : 183-204.

1001. **Smith T.M.F. (1994)**. Sample surveys 1975-1990; an age of reconciliation? (with discussion). *International Statistical Review*, **62** : 5-34.

1002. **Smouse P.E., J.C. Long & R.R. Sokal (1986)**. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic Zoology*, **35** : 627-632.

1003. **Snover M.L. & J.A. Commito (1998)**. The fractal geometry of *Mytilus edulis* L. spatial distribution in a soft-bottom system. *Journal of Experimental Marine Biology and Ecology*, **223** : 53-64.

1004. **Soares A. (1998)**. Sequential indicator simulation with correction for local probabilities. *Mathematical Geology*, **30** : 761-765.

1005. **Soares A., J. Tavora, L. Pinheiro, C. Freitas & J.A. Almeida (1993)**. Predicting probability maps of air pollution concentration: a case study on Barreiro/Seixal industrial area. *In* : Soares A. (Ed.), *Geostatistics Tróia '92*, pp. 625-635. Kluwer Academic Publishers, Dordrecht, The Netherlands.

1006. **Sobol I.M., R.B. Statnikov & N.F. Ovchinnikova (1973)**. Location of the characteristic roots of a matrix. *USSR Computational Mathematics and Mathematical Physics*, **13** : 255-258.

1007. **Söderström M. (1992)**. Geostatistical modeling of salinity as a basis for irrigation management and crop selection. A case study in central Tunisia. *Environmental Geology and Water Sciences*, **20** : 85-92.

1008. **Söderström M. & J.E. Eriksson (1996)**. Cadmium in soil and winter wheat grain in southern Sweden. II. Geographical distribution and its relation to substratum. *Acta Agriculturae Scandinavica Section B - Soil and Plant Science*, **46** : 249-257.

1009. **Sokal R.R. (1979a)**. Ecological parameters inferred from spatial correlograms. *In* : Patil G.P. & M. Rosenzweig (Eds.), *Contemporary quantitative ecology and related ecometrics*, pp. 167-196. International Co-operative Publishing House, Fairland, Maryland, USA.

1010. **Sokal R.R. (1979b)**. Testing statistical significance of geographic variation patterns. *Systematic Zoology*, **28** : 227-232.

1011. **Sokal R.R. (1986)**. Spatial data analysis and historical processes. *In* : Diday E. (Ed.), *Data analysis and informatics. Volume IV*, pp. 29-43. North-Holland, Amsterdam, The Netherlands.

1012. **Sokal R.R. & G.M. Jacquez (1991)**. Testing inferences about microevolutionary processes by means of spatial autocorrelation analysis. *Evolution*, **45** : 152-168.

1013. **Sokal R.R., G.M. Jacquez & M.C. Wooten (1989)**. Spatial autocorrelation analysis of migration and selection. *Genetics*, **121** : 845-855.

1014. **Sokal R.R. & N.L. Oden (1978a)**. Spatial autocorrelation in biology. I. Methodology. *Biological Journal of the Linnean Society*, **10** : 199-228.

1015. **Sokal R.R. & N.L. Oden (1978b)**. Spatial autocorrelation in biology. II. Some biological implications and four applications of evolutionary and ecological interest. *Biological Journal of the Linnean Society*, **10** : 229-249.

1016. **Sokal R.R. & N.L. Oden (1991)**. Spatial autocorrelation analysis as an inferential tool in population genetics. *American Naturalist*, **138** : 518-521.

1017. **Sokal R.R., N.L. Oden & J.S.F. Barker (1987)**. Spatial structure in *Drosophila buzzatii* populations: simple and directional spatial autocorrelation. *American Naturalist*, **129** : 122-142.

1018. **Sokal R.R., N.L. Oden, P. Legendre, M.J. Fortin, J. Kim & A. Vaudor (1989)**. Genetic differences among language families in Europe. *American Journal of Physical Anthropology*, **79** : 489-502.

1019. **Sokal R.R., N.L. Oden & B.A. Thomson (1997)**. A simulation study of microevolutionary inferences by spatial autocorrelation analysis. *Biological Journal of the Linnean Society*, **60** : 73-93.

1020. **Sokal R.R., N.L. Oden, B.A. Thomson & J. Kim (1993)**. Testing for regional differences in means: distinguishing inherent from spurious spatial autocorrelation by restricted randomization. *Geographical Analysis*, **25** : 199-210.

1021. **Sokal R.R. & B.A. Thomson (1998)**. Spatial genetic structure of human populations in Japan. *Human Biology*, **70** : 1-22.

1022. **Sokal R.R. & J.D. Thomson (1987)**. Applications of spatial autocorrelation in ecology. *In* : Legendre P. & L. Legendre (Eds.), *Developments in numerical ecology*, pp. 431-466. Springer-Verlag, Berlin, Germany.

1023. **Sokal R.R. & F.J. Rohlf (1995)**. Biometry. The principles and practice of statistics in biological research. Third edition. Freeman, New York, USA, 887 p.

1024. **Sokal R.R. & D.E. Wartenberg (1983)**. A test of spatial autocorrelation analysis using an isolation-by-distance model. *Genetics*, **105** : 219-237.
1025. **Solow A.R. (1985)**. Bootstrapping correlated data. *Mathematical Geology*, **17** : 769-775.
1026. **Solow A.R. (1990)**. Geostatistical cross-validation: a cautionary note. *Mathematical Geology*, **22** : 637-639.
1027. **Solow A.R. (1994)**. Discussion of A.G. Journel "Resampling from stochastic simulations". *Environmental and Ecological Statistics*, **1** : 90.
1028. **Solow A.R. & S. Polasky (1994)**. Measuring biological diversity (with discussion). *Environmental and Ecological Statistics*, **1** : 95-107.
1029. **Sowey E.R. (1978)**. A second classified bibliography on random number generation and testing. *International Statistical Review*, **46** : 89-102.
1030. **Sowey E.R. (1986)**. A third classified bibliography on random number generation and testing. *Journal of the Royal Statistical Society Series A*, **149** : 83-107.
1031. **Späth H. (1967)**. Algorithm 298. Determination of the square-root of a positive definite matrix. *Communications of the ACM*, **10** : 182.
1032. **Srivastava R.M. (1986)**. Philip and Watson. Quo vadunt? *Mathematical Geology*, **18** : 141-146.
1033. **Starks T.H. & J.H. Fang (1982)**. The effect of drift on the experimental semivariogram. *Mathematical Geology*, **14** : 309-319.
1034. **Stein A. (1994)**. The use of prior information in spatial statistics. *Geoderma*, **62** : 199-216.
1035. **Stein A. & L.C.A. Corsten (1991)**. Universal kriging and cokriging as a regression procedure. *Biometrics*, **47** : 575-587.
1036. **Stein M.L. (1987)**. Minimum norm quadratic estimation of spatial variograms. *Journal of the American Statistical Association*, **82** : 765-772.
1037. **Stein M.L. & M.S. Handcock (1989)**. Some asymptotic properties of kriging when the covariance function is misspecified. *Mathematical Geology*, **21** : 171-190.
1038. **Steinhaus H. (1954)**. Length, shape and area. *Colloquium Mathematicum*, **3** : 1-13.
1039. **Stewart G.W. (1985)**. Matrix, ill-conditioned. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 5, pp. 325-326. Wiley, New York, USA.
1040. **Stewart I. (1990)**. La physique des courbes. *Pour la Science*, **155** : 106-117.
1041. **Stewart-Oaten A. (1995)**. Rules and judgments in statistics: three examples. *Ecology*, **76** : 2001-2009.
1042. **Stigler S.M. (1991)**. Stochastic simulation in the nineteenth century. *Statistical Science*, **6** : 89-97.
1043. **Stiteler W.M. & G.P. Patil (1971)**. Variance-to-mean ratio and Morisita's index as measures of spatial patterns in ecological populations (with discussion). *In* : Patil G.P., E.C. Pielou & W.E. Waters (Eds.), *Statistical Ecology. Volume 1. Spatial patterns and statistical distributions*, pp. 423-459. The Pennsylvania State University Press, University Park, Pennsylvania, USA.
1044. **Stoline M.R. (1991)**. An examination of the lognormal and Box and Cox family of transformations in fitting environmental data. *Environmetrics*, **2** : 85-106.
1045. **Strachan I.B. & L.E. Harvey (1996)**. Quantifying the effects of temporal autocorrelation on climatological regression models using geostatistical techniques. *Canadian Journal of Forest Research*, **26** : 864-871.
1046. **Strahler A.H. (1978)**. Binary discriminant analysis: a new method for investigating species-environment relationships. *Ecology*, **59** : 108-116.

1047. **Student (1914)**. The elimination of spurious correlation due to position in time or space. *Biometrika*, **10** : 179-180.

1048. **Su T.H. & R.C. Chang (1991)**. Computing the constrained relative neighborhood graphs and constrained Gabriel graphs in Euclidean plane. *Pattern Recognition*, **24** : 221-230.

1049. **Sugihara G. & R.M. May (1990)**. Applications of fractals in ecology. *Trends in Ecology & Evolution*, **5** : 79-86.

1050. **Sukhatme S. (1989)**. Kriging with perturbed variogram. *Journal of the Indian Statistical Association*, **27** : 79-88.

1051. **Supowit K.J. (1983)**. The relative neighborhood graph with an application to minimum spanning trees. *Journal of the Association for Computing Machinery*, **30** : 428-448.

1052. **Switzer P. (1967)**. Reconstructing patterns from sample data. *Annals of Mathematical Statistics*, **38** : 138-154.

1053. **Switzer P. (1971)**. Mapping a geographically correlated environment (with discussion). *In* : Patil G.P., E.C. Pielou & W.E. Waters (Eds.), *Statistical Ecology. Volume 1. Spatial patterns and statistical distributions*, pp. 235-269. The Pennsylvania State University Press, University Park, Pennsylvania, USA.

1054. **Switzer P. (1979)**. Statistical consideration in network design. *Water Resources Research*, **15** : 1712-1716.

1055. **Switzer P. (1983)**. Geography, statistics in. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 3, pp. 387-392. Wiley, New York, USA.

1056. **Takayasu H. (1990)**. Fractals in the physical sciences. Manchester University Press, Manchester, UK, 170 p.

1057. **Takenaka Y., H. Matsuda & Y. Iwasa (1997)**. Competition and evolutionary stability of plants in a spatially structured habitat. *Researches on Population Ecology*, **39** : 67-75.

1058. **Tal-Ezer H. (1989)**. Polynomial approximation of functions of matrices and applications. *Journal of Scientific Computing*, **4** : 25-60.

1059. **Tal-Ezer H. (1991)**. High degree polynomial interpolation in Newton form. *SIAM Journal on Scientific and Statistical Computing*, **12** : 648-667.

1060. **Tao S. (1995)**. Kriging and mapping of copper, lead, and mercury contents in surface soil in Shenzhen area. *Water Air and Soil Pollution*, **83** : 161-172.

1061. **Taylor L.R. (1961)**. Aggregation, variance and the mean. *Nature*, **189** : 732-735.

1062. **Taylor L.R. (1971)**. Aggregation as a species characteristic (with discussion). *In* : Patil G.P., E.C. Pielou & W.E. Waters (Eds.), *Statistical Ecology. Volume 1. Spatial patterns and statistical distributions*, pp. 357-377. The Pennsylvania State University Press, University Park, Pennsylvania, USA.

1063. **Taylor L.R. (1984)**. Assessing and interpreting the spatial distributions of insect populations. *Annual Review of Entomology*, **29** : 321-357.

1064. **Taylor L.R., I.P. Woiwod & J.N. Perry (1978)**. The density-dependence of spatial behaviour and the rarity of randomness. *Journal of Animal Ecology*, **47** : 383-406.

1065. **Ter Braak C.J.F. (1987)**. The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio*, **69** : 69-77.

1066. **Ter Braak C.J.F. & S. Juggins (1993)**. Weighted averaging partial least squares regression (WA-PLS): an improved method for reconstructing environmental variables from species assemblages. *Hydrobiologia*, **269-270** : 485-502.

1067. **Tercan A.E. & P.A. Dowd (1995)**. Approximate local confidence intervals under change of support. *Mathematical Geology*, **27** : 149-172.

1068. **Terui N. & M. Kikuchi (1994)**. The size-adjusted critical region of Moran's I test statistics for spatial autocorrelation and its application to geographical areas. *Geographical Analysis*, **26** : 213-227.

1069. **Theraulaz G., E. Bonabeau, S. Goss & J.L. Deneubourg (1994)**. L'intelligence collective. Comment les fourmis recherchent leur nourriture et organisent leur nid. *Pour la Science*, **198** : 90-95.

1070. **Thioulouse J., D. Chessel & S. Champely (1995)**. Multivariate analysis of spatial patterns: a unified approach to local and global structures. *Environmental and Ecological Statistics*, **2** : 1-14.

1071. **Thom R. (1972)**. Stabilité structurelle et morphogénèse. Essai d'une théorie générale des modèles. W. A. Benjamin Inc., Reading, Massachusetts, USA, 362 p.

1072. **Thom R. (1991)**. Prédire n'est pas expliquer. Editions Eshel, Paris, France, 175 p.

1073. **Tiefelsdorf M. & B. Boots (1997)**. A note on the extremities of local Moran's I_is and their impact on the global Moran's I^*. *Geographical Analysis*, **29** : 248-257.

1074. **Tilman D. (1994)**. Competition and biodiversity in spatially structured habitats. *Ecology*, **75** : 2-16.

1075. **Tipper J.C. (1991)**. FORTRAN programs to construct the planar Voronoi diagram. *Computers & Geosciences*, **17** : 597-632.

1076. **Titeux P. (1989)**. Automatisation de problèmes de positionnement sous contraintes : une méthode intégrant des techniques de systèmes experts, de compilation et de bases de données. Application en cartographie. Thèse de doctorat, Université Pierre & Marie Curie - Paris 6, 414 p.

1077. **Tjøstheim D. (1978)**. A measure of association for spatial variables. *Biometrika*, **65** : 109-114.

1078. **Todini E. & M. Ferraresi (1996)**. Influence of parameter estimation uncertainty in kriging. *Journal of Hydrology*, **175** : 555-566.

1079. **Tomassone R., C. Dervin & J.P. Masson (1993)**. Biométrie. Modélisation de phénomènes biologiques. Masson, Paris, France, 553 p.

1080. **Tomlin C.D. (1990)**. Geographic information systems and cartographic modeling. Prentice Hall, Englewood Cliffs, New Jersey, USA, 249 p.

1081. **Toompuu A. & F. Wulff (1996)**. Optimum spatial analysis of monitoring data on temperature, salinity and nutrient concentrations in the Baltic proper. *Environmental Monitoring and Assessment*, **43** : 283-308.

1082. **Tough J.G. (1988)**. The computation of the area, centroid, and principal axes of a polygon. *Computers & Geosciences*, **14** : 715-717.

1083. **Tough J.G. & R.G. Miles (1984)**. A method for characterizing polygons in terms of the principal axes. *Computers & Geosciences*, **10** : 347-350.

1084. **Tournassoud P. (1988)**. Géométrie et intelligence artificielle pour les robots. Hermès, Paris, France, 312 p.

1085. **Toussaint G.T. (1980)**. The relative neighbourhood graph of a finite planar set. *Pattern Recognition*, **12** : 261-268.

1086. **Trochu F. (1993)**. A contouring program based on dual kriging interpolation. *Engineering with Computers*, **9** : 160-177.

1087. **Tsai V.J.D. (1993)**. Fast topological construction of Delaunay triangulations and Voronoi diagrams. *Computers & Geosciences*, **19** : 1463-1474.

1088. **Tukey J.W. (1958)**. Bias and confidence in not-quite large samples (abstract). *Annals of Mathematical Statistics*, **29** : 614.

1089. **Turner M.G. (1989)**. Landscape ecology: the effect of pattern on process. *Annual Review of Ecology and Systematics*, **20** : 171-197.

1090. **Turner M.G. & R.H. Gardner (1991)**. Quantitative methods in landscape ecology: an introduction. *In* : Turner M.G. & R.H. Gardner (Eds.), *Quantitative methods in landscape ecology*, pp. 3-14. Springer-Verlag, New York, USA.

1091. **Turner S.J., R.V. O'Neill, W. Conley, M.R. Conley & H.C. Humphries (1991)**. Pattern and scale: statistics for landscape ecology. *In* : Turner M.G. & R.H. Gardner (Eds.), *Quantitative methods in landscape ecology*, pp. 17-49. Springer-Verlag, New York, USA.

1092. **Turner M.G., R.V. O'Neill, R.H. Gardner & B.T. Milne (1989)**. Effects of changing spatial scale on the analysis of landscape pattern. *Landscape Ecology*, **3** : 153-162.

1093. **Unwin D.J. (1975)**. Numerical errors in a familiar technique: a case study of polynomial trend surface analysis. *Geographical Analysis*, **7** : 197-203.

1094. **Upton G.J.G. (1984)**. On Mead's test for pattern. *Biometrics*, **40** : 759-766.

1095. **Upton G.J.G. & B. Fingleton (1985)**. Spatial data analysis by example. Point pattern and quantitative data. Wiley, Chichester, UK, 410 p.

1096. **Urquhart N.S. (1997)**. Discussion of D.J. Brus and J.J. de Gruijter "Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil" . *Geoderma*, **80** : 52-53.

1097. **Urquhart R. (1982a)**. Graph theoretical clustering based on limited neighbourhood sets. *Pattern Recognition*, **15** : 173-187.

1098. **Urquhart R. (1982b)**. Graph theoretical clustering based on limited neighbourhood sets. Erratum. *Pattern Recognition*, **15** : 427-428.

1099. **Valdez-Cepeda R.D. & E. Olivares-Sáenz (1998)**. Fractal analysis of Mexico's annual mean yields of maize, bean, wheat and rice. *Field Crops Research*, **59** : 53-62.

1100. **Van Groenigen J.W., W. Siderius & A. Stein (1999)**. Constrained optimisation of soil sampling for minimisation of the kriging variance. *Geoderma*, **87** : 239-259.

1101. **Van Groenigen J.W. & A. Stein (1998)**. Constrained optimization of spatial sampling using continuous simulated annealing. *Journal of Environmental Quality*, **27** : 1078-1086.

1102. **Van Groenigen J.W., A. Stein & R. Zuurbier (1997)**. Optimization of environmental sampling using interactive GIS. *Soil Technology*, **10** : 83-97.

1103. **Van Hees W.W.S. (1994)**. A fractal model of vegetation complexity in Alaska. *Landscape Ecology*, **9** : 271-278.

1104. **Van Leeuwen E.P., G.P.J. Draaijers & J.W. Erisman (1996)**. Mapping wet deposition of acidifying components and base cations over Europe using measurements. *Atmospheric Environment*, **30** : 2495-2511.

1105. **Van Oosterom P. (1994)**. An R-tree based map-overlay algorithm. *In* : Harts J.J., H.F.L. Ottens, H.J. Scholten & J. van Arragon (Eds.) *EGIS/MARI '94. Fifth european conference and exhibition on geographical information systems*, pp. 318-327. EGIS Foundation, Utrecht, The Netherlands.

1106. **Vardeman S.B. (1987)**. Discussion of G. Casella and R.L. Berger "Reconciling bayesian and frequentist evidence in the one-sided testing problem" and of J.O. Berger and T. Sellke "Testing a point null hypothesis: the irreconcilability of P values and evidence". *Journal of the American Statistical Association*, **82** : 130-131.

1107. **Varekamp C., A.K. Skidmore & P.A.B. Burrough (1996)**. Using public domain geostatistical and GIS software for spatial interpolation. *Photogrammetric Engineering and Remote Sensing*, **62** : 845-854.

1108. **Vedyushkin M.A. (1994)**. Fractal properties of forest spatial structure. *Vegetatio*, **113** : 65-70.

1109. **Veltkamp R.C. (1992)**. The gamma-neighborhood graph. *Computational Geometry - Theory and Applications*, **1** : 227-246.

1110. **Ver Hoef J.M. & N. Cressie (1993)**. Multivariable spatial prediction. *Mathematical Geology*, **25** : 219-240.

1111. **Ver Hoef J.M., N. Cressie & D.C. Glenn-Lewin (1993)**. Spatial models for spatial statistics: some unification. *Journal of Vegetation Science*, **4** : 441-452.

1112. **Verly G.W. (1993)**. Sequential Gaussian cosimulation: a simulation method integrating several types of information. *In* : Soares A. (Ed.), *Geostatistics Tróia '92*, pp. 543-554. Kluwer Academic Publishers, Dordrecht, The Netherlands.

1113. **Vieira C., P. Aubry, D. Lepetit & C. Biemont (1998)**. A temperature cline in copy number for 412 but not roo/B104 retrotransposons in populations of *Drosophila simulans*. *Proceedings of the Royal Society of London Series B*, **265** : 1161-1165.

1114. **Vieira C. & C. Biemont (1996)**. Geographical variation in insertion site number of retrotransposon 412 in *Drosophila simulans*. *Journal of Molecular Evolution*, **42** : 443-451.

1115. **Vieu L. (1991)**. Sémantique des relations spatiales et inférences spatio-temporelles : une contribution à l'étude des structures formelles de l'espace en Langage Naturel. Thèse de doctorat, Université Paul Sabatier, 358 p.

1116. **Villard M.A. & B.A. Maurer (1996)**. Geostatistics as a tool for examining hypothesized declines in migratory songbirds. *Ecology*, **77** : 59-68.

1117. **Vincent P.J., J.M. Haworth, J.G. Griffiths & R. Collins (1976)**. The detection of randomness in plant patterns. *Journal of Biogeography*, **3** : 373-380.

1118. **Vischetti C., M. Businelli, M. Marini, E. Capri, M. Trevisan, A.A.M. Delre, L. Donnarumma, E. Conte & G. Imbroglini (1997)**. Characterization of spatial variability structure in three separate field trials on pesticide dissipation. *Pesticide Science*, **50** : 175-182.

1119. **Voĭtsekhovskiĭ M.I. (1995)**. Jordan theorem. *In* : Hazewinkel M. (Ed.), *Encyclopedia of mathematics*, vol. 3, pp. 363-364. Kluwer Academic Publishers, Dordrecht, The Netherlands.

1120. **Voltz M., P. Lagacherie & X. Louchart (1997)**. Predicting soil properties over a region using sample information from a mapped reference area. *European Journal of Soil Science*, **48** : 19-30.

1121. **Von Neumann J. (1941)**. Distribution of the ratio of the mean square successive difference to the variance. *Annals of Mathematical Statistics*, **12** : 367-395.

1122. **Von Neumann J., R.H. Kent, H.R. Bellinson & B.I. Hart (1941)**. The mean square successive difference. *Annals of Mathematical Statistics*, **12** : 153-162.

1123. **Von Steiger B., R. Webster, R. Schulin & R. Lehmann (1996)**. Mapping heavy metals in polluted soil by disjuntive kriging. *Environmental Pollution*, **94** : 205-215.

1124. **Voss R.F. (1988)**. Fractals in nature: from characterization to simulation. *In* : Peitgen H.O. & D. Saupe (Eds.), *The science of fractal images*, pp. 21-70. Springer-Verlag, New York, USA.

1125. **Wackernagel H. (1993)**. Cours de géostatistique multivariable. ENSMP, Fontainebleau, France, 80 p.

1126. **Wackernagel H. (1998)**. Multivariate geostatistics: an introduction with applications. Second edition. Springer-Verlag, Berlin, Germany, 291 p.

1127. **Wackernagel H. & C. Butenuth (1989)**. Caractérisation d'anomalies géochimiques par la géostatistique multivariable. *Journal of Geochemical Exploration*, **32** : 437-444.

1128. **Wahlberg N., A. Moilanen & I. Hanski (1996)**. Predicting the occurrence of endangered species in fragmented landscapes. *Science*, **273** : 1536-1538.

1129. **Waldhör T. (1996)**. The spatial autocorrelation coefficient Moran's I under heteroscedasticity. *Statistics in Medicine*, **15** : 887-892.

1130. **Walker D.D., J.C. Loftis & P.W. Mielke (1997)**. Permutation methods for determining the significance of spatial dependence. *Mathematical Geology*, **29** : 1011-1024.

1131. **Walker P.A. (1996)**. Spatial modelling and population ecology. *In* : Floyd R.B., A.W. Sheppard & P.J. Debarro (Eds.), *Frontiers of population ecology* , pp. 419-429. CSIRO, Melbourne, Australia.

1132. **Wallace M.K. & D.M. Hawkins (1994)**. Applications of geostatistics in plant nematology. *Journal of Nematology*, **26** : 626-634.

1133. **Walter S.D. (1974)**. On the detection of household aggegation of disease. *Biometrics*, **30** : 525-538.

1134. **Warnes J.J. (1986)**. A sensitivity analysis for universal kriging. *Mathematical Geology*, **18** : 653-676.

1135. **Warnes J.J. & B.D. Ripley (1987)**. Problems with likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika*, **74** : 640-642.

1136. **Warrick A.W. & D.E. Myers (1987)**. Optimization of sampling locations for variogram calculations. *Water Resources Research*, **23** : 496-500.

1137. **Wartenberg D. (1985a)**. Multivariate spatial correlation: a method for exploratory geographical analysis. *Geographical Analysis*, **17** : 263-283.

1138. **Wartenberg D. (1985b)**. Canonical trend surface analysis: a method for describing geographic patterns. *Systematic Zoology*, **34** : 259-279.

1139. **Watkins A.J. (1992)**. On models of spatial covariance. *Computational Statistics & Data Analysis*, **13** : 473-481.

1140. **Watson G.S. (1983)**. Geology, statistics in. *In* : Kotz S. & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, vol. 3, pp. 392-396. Wiley, New York, USA.

1141. **Weber D.D. & E.J. Englund (1992)**. Evaluation and comparison of spatial interpolator. *Mathematical Geology*, **24** : 381-391.

1142. **Weber D.D. & E.J. Englund (1994)**. Evaluation and comparison of spatial interpolator II. *Mathematical Geology*, **26** : 589-603.

1143. **Webster R. (1991)**. Local disjunctive kriging of soil properties with change of support. *Journal of Soil Science*, **42** : 301-318.

1144. **Webster R. & B. Boag (1992)**. Geostatistical analysis of cyst nematodes in soil. *Journal of Soil Science*, **43** : 583-595.

1145. **Webster R. & B. Boag (1992)**. The spatial distribution of cyst nematodes in soil. *In* : Bárdossy A. (Ed.) *Geostatistical methods: recent developments and applications in surface and subsurface hydrology*, pp. 131-141. UNESCO, Paris, France.

1146. **Webster R. & T.M. Burgess (1980)**. Optimal interpolation and isarithmic mapping of soil properties. III. Changing drift and universal kriging. *Journal of Soil Science*, **31** : 505-524.

1147. **Webster R. & T.M. Burgess (1984)**. Sampling and bulking strategies for estimating soil properties in small regions. *Journal of Soil Science*, **35**: 127-140.

1148. **Webster R. & A.B. McBratney (1989)**. On the Akaike information criterion for choosing models for variograms of soil properties. *Journal of Soil Science*, **40** : 493-496.

1149. **Webster R. & M.A. Oliver (1989)**. Optimal interpolation and isarithmic mapping of soil properties. VI. Disjunctive kriging and mapping the conditional probability. *Journal of Soil Science*, **40** : 497-512.

1150. **Webster R. & M.A. Oliver (1992a)**. Sample adequately to estimate variograms of soil properties. *Journal of Soil Science*, **43** : 177-192.

1151. **Webster R. & M.A. Oliver (1992b)**. Confidence intervals on variograms for samples of various sizes. *Sciences de la Terre, Série Informatique Géologique*, **31** : 11-23.

1152. **Webster R. & M.A. Oliver (1993)**. How large a sample is needed to estimate the regional variogram adequately. *In* : Soares A. (Ed.), *Geostatistics Tróia '92*, pp. 155-166. Kluwer Academic Publishers, Dordrecht, The Netherlands.

1153. **Webster R., M.A. Oliver, K.R. Muir & J.R. Mann (1994)**. Kriging the local risk of a rare disease from a register of diagnoses. *Geographical Analysis*, **26** : 168-185.

1154. **Weisz R., S. Fleischer & Z. Smilowitz (1995a)**. Site-specific Iintegrated pest management for high value crops: sample units for map generation using the Colorado Potato Beetle (Coleoptera: Chrysomelidae) as a model system. *Journal of Economic Entomology*, **88** : 1069-1080.

1155. **Weisz R., S. Fleischer & Z. Smilowitz (1995b)**. Map generation in high-value horticultural integrated pest management: appropriate interpolation methods for site-specific pest management of Colorado Potato Beetle (Coleoptera: Chrysomelidae). *Journal of Economic Entomology*, **88** : 1650-1657.

1156. **Wen R. & R. Sinding-Larsen (1997)**. Uncertainty in fractal dimension estimated from power spectra and variograms. *Mathematical Geology*, **29** : 727-753.

1157. **Weseloh R.M. (1996)**. Developing and validating a model for predicting Gypsy Moth (Lepidoptera: Lymantriidae) defoliation in Connecticut. *Journal of Economic Entomology*, **89** : 1546-1555.

1158. **Western A.W., G. Bloschl & R.B. Grayson (1998)**. How well do indicator variograms capture the spatial connectivity of soil moisture? *Hydrological Processes*, **12** : 1851-1868.

1159. **Whalley W.B. & J.D. Orford (1989)**. The use of fractals and pseudofractals in the analysis of two-dimensional outlines: review and further exploration. *Computers & Geosciences*, **15** : 185-197.

1160. **Whitaker R.A. (1977)**. Three algorithms for calculating some or all of the shortest paths in a sparse network. *Geographical Analysis*, **9** : 266-277.

1161. **White D.L., D. Bushek, D.E. Porter & D. Edwards (1998)**. Geographic information systems (GIS) and kriging: analysis of the spatial and temporal distributions of the oyster pathogen *Perkinsus marinus* in a developed and an undeveloped estuary. *Journal of Shellfish Research*, **17** : 1473-1476.

1162. **White L.V. & W.J. Welch (1981)**. A method for constructing valid restricted randomization schemes using the theory of D-optimal design of experiments. *Journal of the Royal Statistical Society Series B*, **43** : 167-172.

1163. **Wichmann B.A. & I.D. Hill (1982)**. Algorithm AS 183. An efficient and portable pseudo-random number generator. *Applied Statistics*, **31** : 188-190.

1164. **Wiens J.A. (1989)**. Spatial scaling in ecology. *Functional Ecology*, **3** : 385-397.

1165. **Wiens J.A., T.O. Crist, K.A. With & B.T. Milne (1995)**. Fractal patterns of insect movement in microlandscape mosaics. *Ecology*, **76** : 663-666.

1166. **Williams J.D. (1941)**. Moments of the ratio of the mean square successive difference to the mean square difference in samples from a normal universe. *Annals of Mathematical Statistics*, **12** : 239-241.

1167. **Williams R.M. (1956)**. The variance of the mean of systematic samples. *Biometrika*, **43** : 137-148.

1168. **Wingle W.L. & E.P. Poeter (1993)**. Uncertainty associated with semivariograms used for site simulation. *Ground Water*, **31** : 725-734.

1169. **With K.A. (1994)**. Using fractal analysis to assess how species perceive landscape structure. *Landscape Ecology*, **9** : 25-36.

1170. **Wolter K.M. (1984)**. An investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association*, **79** : 781-790.

1171. **Wopereis M.C.S., A. Stein, M.J. Kropff & J. Bouma (1996)**. Spatial interpolation of soil hydraulic properties and simulated rice yield. *Soil Use and Management*, **12** : 158-166.

1172. **Yao A.C.C. (1975)**. An $O(|E| \log\log |V|)$ algorithm for finding minimum spanning trees. *Information Processing Letters*, **4** : 21-23.

1173. **Yfantis E.A., G.T. Flatman & J.V. Behar (1987)**. Efficiency of kriging estimation for square, triangular and hexagonal grids. *Mathematical Geology*, **19** : 183-205.

1174. **Yoccoz N.G. (1991)**. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America*, **72** : 106-111.

1175. **Yost R.S., G. Uehara & R.L. Fox (1982a)**. Geostatistical analysis of soil chemical properties of large land areas. I. Semi-variograms. *Soil Science Society of America Journal*, **46** : 1028-1032.

1176. **Yost R.S., G. Uehara & R.L. Fox (1982b)**. Geostatistical analysis of soil chemical properties of large land areas. II. Kriging. *Soil Science Society of America Journal*, **46** : 1033-1037.

1177. **Zahl S. (1974)**. Application of the S-method to the analysis of spatial pattern. *Biometrics*, **30** : 513-524.

1178. **Zeisel H. (1986)**. A remark on algorithm AS 183. An efficient and portable pseudo-random number generator. *Applied Statistics*, **35** : 89.

1179. **Zhang C.S. & O. Selinus (1997)**. Spatial analyses for copper, lead and zinc contents in sediments of the Yangtze River basin. *Science of the Total Environment*, **204** : 251-262.

1180. **Zhang J.T. (1994)**. A combinaison of fuzzy set ordination with detrended correspondence analysis: one way to combine multi-environmental variables with vegetation data. *Vegetatio*, **115** : 115-121.

1181. **Zhang R., A.W. Warrick & D.E. Myers (1990)**. Variance as a function of sample support size. *Mathematical Geology*, **22** : 107-121.

1182. **Zhang R.D., P. Shouse & S. Yates (1999)**. Estimates of soil nitrate distributions using cokriging with pseudo-crossvariograms. *Journal of Environmental Quality*, **28** : 424-428.

1183. **Zhang X.F., J.C.H. van Eijkeren & A.W. Heemink (1995)**. On the weighted least-squares method for fitting a semivariogram model. *Computers & Geosciences*, **21** : 605-608.

1184. **Zhou M. (1998)**. An objective interpolation method for spatiotemporal distribution of marine plankton. *Marine Ecology - Progress Series*, **174** : 197-206.

1185. **Zhu A.X. (1997)**. A similarity model for representing soil spatial information. *Geoderma*, **77** : 217-242.

1186. **Zhu A.X., L.E. Band, B. Dutton & T.J. Nimlos (1996)**. Automated soil inference under fuzzy logic. *Ecological Modelling*, **90** : 123-145.

1187. **Zimmerman D.L. (1989a)**. Computationally exploitable structure of covariance matrices and generalized covariance matrices in spatial models. *Journal of Statistical Computation and Simulation*, **32** : 1-15.

1188. **Zimmermann D.L. (1989b)**. Computationally efficient restricted maximum likelihood estimation of generalized covariance functions. *Mathematical Geology*, **21** : 655-672.

1189. **Zimmerman D.L. (1993)**. Another look at anisotropy in geostatistics. *Mathematical Geology*, **25** : 453-470.

1190. **Zimmerman D.L. & M.B. Zimmerman (1991)**. A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors. *Technometrics*, **33** : 77-91.

1191. **Zinger A. (1964)**. Systematic sampling in forestry. *Biometrics*, **20** : 553-565.

1192. **Zubrzycki S. (1958)**. Remarks on random, stratified and systematic sampling in a plane. *Colloquium Mathematicum*, **6** : 251-264.