GeoSiPol

Les pratiques de la géostatistique dans le domaine des sites et sols pollués

Géostatistique appliquée aux sites et sols pollués

- Manuel méthodologique et exemples d'applications -

INERIS

ADEME

Agence de l'Environnement et de la Maîtrise de l'Energie

GeoSiPol

Version 1, Novembre 2005

Table des matières

Introduction générale *7*

1 Le groupe de travail GeoSiPol « Géostatistique et Sites Pollués » **9**

2 Le manuel méthodologique comme premiere action du groupe **9**

3 Contributions à la réalisation du manuel méthodologique **10**

4 Avertissement au lecteur **11**

I Thème n°1 : Comment passer des analyses chimiques à la cartographie des teneurs en polluant ? *13*

1 Présentation du thème **17**

1.1 Introduction 17

1.2 Cadre métier 17

1.3 Pratiques habituelles 18

1.4 Intérêt d'une approche géostatistique 18

1.5 Objectifs 19

1.6 Méthodologie 19

2 L'analyse exploratoire des données **20**

2.1 Contexte de l'analyse exploratoire 20

2.2 Information sur le site et sur l'échantillonnage 20

2.3 Les étapes de l'analyse exploratoire 22

3 Cartographie **33**

3.1 Introduction au krigeage comme méthode d'estimation 33

3.2 Exploitation de la variance de krigeage 36

3.3 Cas particuliers 38

3.4 Intérêt et limites de l'estimation 40

4 Illustration **41**

4.1 Histogrammes et statistiques élémentaires 41

4.2 Variogrammes expérimentaux 42

4.3 Ajustement du variogramme 44

4.4 Krigeage des teneurs 45

4.5 Conclusions 46

II Thème n°2 : Comment délimiter et quantifier les zones à dépolluer ? *47*

1 Présentation du thème **51**

1.1 Introduction 51

1.2 Cadre réglementaire 51

1.3 Pratiques habituelles 52

1.4 Intérêt d'une approche géostatistique 53

1.5 Objectifs 54

2 Position du problème et hypothèses de travail **54**

2.1 Seuils de teneurs 55

2.2 Teneurs de maille 55
2.3 Critères de classification 56
2.4 Critères de décision 56
2.5 Tri sélectif des sols en dépollution 57
3 Démarche géostatistique détaillée 59
3.1 Méthodologie 59
3.2 Analyse des données 60
3.3 Génération de simulations conditionnelles 61
3.4 Estimation des zones et des volumes contaminés 65
3.5 Prise en compte de critères financiers dans la classification des sols 70
3.6 Dimensionnement d'un tri sélectif en dépollution 71
3.7 Evaluation financière du projet de dépollution 72
4 Illustration 73
4.1 Contexte 73
4.2 Formalisation des critères de classification et décision 74
4.3 Résultats 76
III Thème n°3 : Comment intégrer toute l'information, de l'historique aux analyses sur site ? 81
1 Présentation du thème 85
1.1 Introduction 85
1.2 Cadre métier 85
1.3 Position du problème et hypothèses de travail 87
1.4 Objectifs 88
1.5 Pratiques habituelles 89
1.6 Intérêt d'une approche géostatistique 90
1.7 Méthodologie 90
2 Démarche géostatistique détaillée 90
2.1 Considérations pratiques : analyse des données disponibles 90
2.2 Estimation de teneurs avec prise en compte de l'information indirecte 93
2.3 Cas de plusieurs informations indirectes 95
2.4 Application à la délimitation et la quantification des zones à dépolluer 96
3 Illustration 96
3.1 Contexte 96
3.2 Analyse des données 97
3.3 Structure spatiale bivariable 99
3.4 Cartographie des teneurs en hydrocarbures totaux 100
3.5 Conclusions 102
IV Thème n°4 : Quel est l'apport de la géostatistique à l'échantillonnage des sites pollués ? 103
1 Présentation du thème 107
1.1 Introduction 107
1.2 Cadre réglementaire 108

1.3 Pratiques habituelles 108
1.4 Intérêt d'une approche géostatistique 109
1.5 Objectifs 109
2 Questions les plus fréquentes sur l'échantillonnage 110
2.1 Combien d'échantillons pour une approche géostatistique ? 110
2.2 Quelles conditions de prélèvement respecter ? 111
2.3 Comment échantillonner sur la verticale? 111
2.4 Faut-il faire des analyses précises, mais chères et donc peu nombreuses, ou des analyses moins précises, mais peu coûteuses et donc plus nombreuses ? 111
2.5 Comment procéder pour que des données collectées dans le cadre des évaluations de risques puissent être utilisées aussi pour une quantification des volumes à traiter ? Un échantillonnage spécifique doit-il être envisagé ? 112
2.6 Doit-on envisager un échantillonnage préférentiel de certaines parties du site ? 112
2.7 Suffit-il d'échantillonner à la maille définie par la portée du variogramme ? 113
3 Approche géostatistique détaillée 113
3.1 Méthodologie 113
3.2 Analyse exploratoire des données 113
3.3 Définition de critères de rationalisation de l'échantillonnage 113
3.4 Estimation des incertitudes 114
3.5 Calcul d'un plan d'échantillonnage rationalisé (localisation, nombre) 114
3.6 Production de résultats pour l'aide à la décision 115
4 Illustration 117
4.1 Contexte 117
4.2 Analyse exploratoire des données existantes 117
4.3 Définition de critères de rationalisation de l'échantillonnage 120
4.4 Production de cartes de la contamination et estimation de l'incertitude 120
4.5 Calcul d'un plan d'échantillonnage rationalisé (nombre, localisation) 121
4.6 Valorisation des résultats d'échantillonnage complémentaire 123
Conclusion 125
Bibliographie 129
Sites internet 133
Index des termes géostatistiques 137

INTRODUCTION GÉNÉRALE

1 LE GROUPE DE TRAVAIL GEOSIPOL « GEOSTATISTIQUE ET SITES POLLUES »

Comparées aux techniques d'interpolation classiques, les méthodes géostatistiques présentent une valeur ajoutée importante : cartographie des teneurs en polluants en place intégrant la nature physique du phénomène, quantification du terme source dans le cas de la contamination possible d'une nappe par des sols pollués, estimation de quantités de terres à dépolluer et évaluation d'un budget associé qui tient compte de l'incertitude d'estimation. Elles permettent de diminuer les coûts d'investigation en intégrant des informations indirectes dans les estimations (données géologiques, géophysiques, relatives à d'autres substances, organoleptiques, etc.). Utilisée à bon escient, la géostatistique fournit aussi des outils précieux pour l'optimisation des phases essentielles que sont l'échantillonnage, le tri sélectif de terres à dépolluer, ou bien encore le contrôle de l'état de contamination résiduelle d'un site après dépollution. En Europe, et notamment en France, l'utilisation de ces méthodes reste pourtant marginale dans le domaine des sites et sols pollués.

Dans un tel contexte, il est apparu intéressant de susciter une réflexion sur les possibilités pratiques offertes par la géostatistique dans la gestion des sites et sols pollués, en organisant un groupe de travail national sur ce thème. La vocation de ce groupe, baptisé ***GeoSiPol***, est triple :

1. Faire connaître et reconnaître l'intérêt pratique et la pertinence de la géostatistique pour le diagnostic et la dépollution de sites contaminés.
2. Donner accès à la géostatistique à un plus grand nombre de praticiens, auprès de bureaux d'étude et de sociétés de dépollution : vers l'établissement et la validation « d'approches standards » définies par rapport à des objectifs d'étude simples et précis.
3. Fournir aux Autorités environnementales les éléments d'information et les conditions pour une intégration possible de la géostatistique dans la méthodologie nationale.

GeoSiPol s'est juridiquement constitué en **association Loi 1901** créée en juin 2004. Outre les membres experts à l'origine du projet (FSS International, Géovariances, Ecole des Mines de Paris), elle comprend :

- des institutionnels (ADEME, INERIS),
- des bureaux d'études,
- des centres de recherche,
- des industriels concernés par la gestion des sites et sols pollués.

2 LE MANUEL METHODOLOGIQUE COMME PREMIERE ACTION DU GROUPE

Afin de permettre aux organismes publics et aux professionnels des sites pollués de mieux se positionner par rapport à ce type de démarches et de le proposer, la première action de ***GeoSiPol*** a été d'élaborer un manuel méthodologique présentant, de façon didactique, le principe et l'intérêt d'une approche géostatistique sur quatre thèmes d'étude liés aux sites et sols pollués. Ce manuel repose sur l'expérience des membres experts dans ce domaine.

Le manuel est centré sur **quatre thèmes** identifiés comme prioritaires par le groupe :

1. Comment passer des analyses chimiques à la cartographie des teneurs en polluant ?
2. Comment délimiter et quantifier les zones à dépolluer ?
3. Comment intégrer toute l'information, de l'historique aux analyses sur site ?
4. Quel est l'apport de la géostatistique à l'échantillonnage des sites pollués ?

Pour chaque thème, le manuel donne les lignes directrices et décrit les étapes à suivre pour traiter le thème considéré selon une approche géostatistique fondée sur des méthodes connues et reconnues. Une bibliographie, une liste de sites internet et de forums, sur la géostatistique et son application dans

différents domaines, et un index des termes géostatistiques employés complètent le manuel (voir annexes).

Ce manuel est destiné à un public sans connaissances particulières en géostatistique. A l'attention des lecteurs désireux d'en savoir davantage sur les méthodes proposées, certains paragraphes plus techniques peuvent préciser des points théoriques utiles. Ces paragraphes apparaissent en italique et peuvent être omis en première lecture sans perte de compréhension du reste du texte.

L'ordre de lecture du manuel importe peu. Il peut cependant être recommandé aux lecteurs non avertis de commencer par le premier thème qui aborde quelques aspects fondamentaux de la géostatistique, tels que l'analyse exploratoire des données, auxquels font référence les autres thèmes.

Chaque thème débute par une fiche technique recto verso synthétique de présentation du thème.

Chaque thème a été élaboré selon le protocole suivant prévu pour en assurer la cohérence et la qualité.

- Mise au point d'un canevas revu et validé par les membres experts et l'ensemble des membres adhérents.
- Rédaction des thèmes par les membres experts.
- Relecture critique croisée entre membres experts.
- Relecture critique par plusieurs membres adhérents institutionnels et privés volontaires.
- Réunion de présentation et validation des thèmes auprès de l'ensemble des membres de *GeoSiPol*.
- Edition de la version finale du manuel dans le présent document.

3 CONTRIBUTIONS A LA REALISATION DU MANUEL METHODOLOGIQUE

La réalisation du manuel méthodologique a bénéficié d'un financement ADEME et INERIS.

Le manuel a été co-écrit par les membres experts du groupe de travail GeoSiPol :

Jean-Paul Chilès (Ecole des Mines de Paris)
Hélène Demougeot-Renard (FSS International)
Dimitri D'Or (FSS International)
Chantal de Fouquet (Ecole des Mines de Paris)
Michel Garcia (FSS International)
Nicolas Jeannée (Géovariances)
Gaëlle Le Loc'h (Ecole des Mines de Paris)
Jean-Jacques Peraudin (Géovariances)

Il a bénéficié d'une relecture et d'une validation des membres du groupe :

Eric Antoinet (ANTEA)
Raymonde Blanchin (BRGM)
Thierry Blondel (CCB)
Olivier Bour (INERIS)
Giovanni Cardenas (INERIS)
Patrick Charbonnier (Bail Industrie)
Fabien Comby (IKOS)
Jean Marie Côme (BURGEAP)
Christian Cornet (CSD Azur)
Véronique Croze (ICF Environnement)
Didier Dubot (CEA)
Nadine Dueso (ADEME)
Frédérique Fabre (ADEME)
Valérie Guérin (CNRSSP)
Frank Haeseler (IFP)
Martine Louvrier (ADEME)
Franck Marot (ADEME)
Emmanuel Steyer (CNRSSP)
Sylvie Traverse (BURGEAP)

4 AVERTISSEMENT AU LECTEUR

Il est rappelé que ce document se veut être une introduction à la géostatistique replacée dans un cadre métier d'application aux sites et sols pollués. Pour chacun des thèmes abordés, le principe d'une démarche géostatistique y est présenté, justifié et illustré sur un cas d'étude réel.

Les démarches présentées n'ont rien d'universel. Elles sont supposées fournir des lignes directrices pour atteindre des objectifs d'étude à l'aide de la géostatistique. Ces approches demandent à être adaptées au cas par cas en fonction des sites étudiés.

Ce document n'a pas pour but de former un ingénieur d'études ou un responsable de projet à la pratique de la géostatistique. Si le manuel précise certains points théoriques jugés utiles pour la compréhension des démarches géostatistiques, il n'a pas vocation à décrire dans le détail toutes les méthodes, géostatistiques ou autres, qui y sont discutées. Le lecteur intéressé pour en savoir davantage ou pour passer à l'application d'approches géostatistiques est invité à consulter les références bibliographiques ou les sites internet recommandés en annexes. Il peut aussi se faire connaître auprès de ***GeoSiPol*** qui pourra l'orienter dans sa démarche.

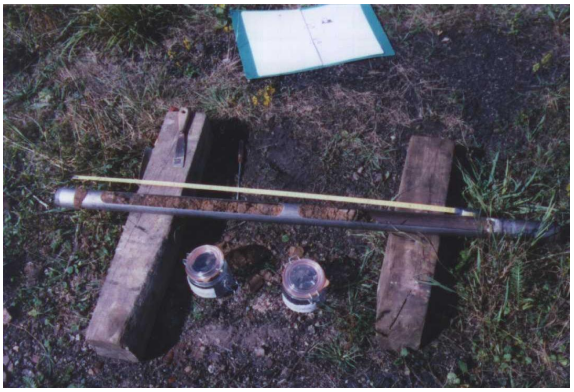
I THÈME N°1 : COMMENT PASSER DES ANALYSES CHIMIQUES À LA CARTOGRAPHIE DES TENEURS EN POLLUANT ?

Comment passer des analyses chimiques à la cartographie des teneurs en polluant ?

Fiche technique N°1

NOUS SOMMES TOUS CONFRONTES A CE PROBLEME

Le prélèvement d'échantillons et l'analyse chimique de leur teneur en polluant sont des opérations coûteuses. Aussi le nombre de données disponibles pour la caractérisation d'un site potentiellement contaminé est-il toujours limité. Il est certes plus élevé au stade d'un diagnostic approfondi en vue d'une Évaluation Détaillée des Risques que pour un diagnostic initial (Évaluation Simplifiée des Risques), mais les données sont fréquemment hétérogènes du fait de campagnes espacées dans le temps, de protocoles d'échantillonnage différents ou d'analyses chimiques de différentes qualités. **Dans tous les cas il importe de tirer le meilleur parti possible de l'information disponible** pour :



- comprendre la répartition de la pollution dans l'espace, et notamment la localisation des fortes teneurs ;
- cartographier cette pollution au mieux et l'assortir d'un indicateur de précision.

La méthodologie nationale juge que l'analyse des incertitudes est « essentielle et indispensable », mais l'approche à mettre en œuvre pour caractériser ces incertitudes reste floue.

LES PRATIQUES ACTUELLES

Une analyse exploratoire et statistique des données permet d'analyser les incertitudes qui leur sont attachées, de comprendre la structure spatiale des teneurs, d'orienter le choix de la méthode d'interpolation. Bien que cruciale, cette phase est habituellement traitée de façon rudimentaire, quand elle n'est pas tout simplement ignorée. La tendance est de passer directement à une cartographie des teneurs dans les sols (toutes données confondues) en utilisant des méthodes d'interpolation ou d'estimation plus ou moins conventionnelles.

Ainsi, le traitement des données manque souvent de fondements méthodologiques :

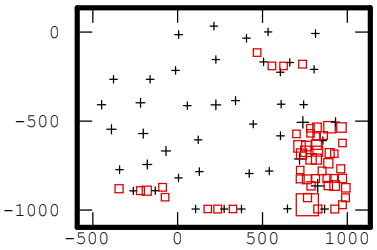
- absence de différenciation des données des différentes campagnes alors que les méthodes d'échantillonnage (volume et mode de prélèvement des échantillons) et d'analyse (précision) ont évolué ;
- traitement indifférencié alors que les données font ressortir des zones aux caractéristiques distinctes (liées par exemple à la lithologie) ;
- manque de validation et de traitement spécifique des données anomales, qui peuvent correspondre à des pics de pollution ou à des erreurs de mesure ;
- recours à des méthodes d'interpolation inadaptées à la structuration spatiale des teneurs ;
- appréciation erronée de l'incertitude associée aux résultats, incertitude qui est bien souvent tout simplement ignorée.

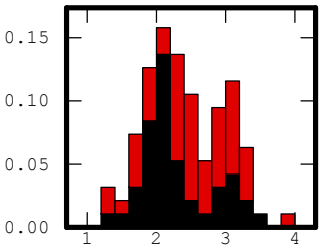
Cela peut avoir pour conséquence une évaluation erronée de la répartition de la pollution, conduisant à des décisions inappropriées.

CE QUE NOUS APPORTE LA GEOSTATISTIQUE

La géostatistique permet une analyse cohérente des données, des incertitudes et des erreurs qui les entachent, ainsi que de la structuration spatiale de la teneur. Cette analyse permet à son tour de quantifier l'incertitude sur les cartes de teneur en polluant qui sont produites.

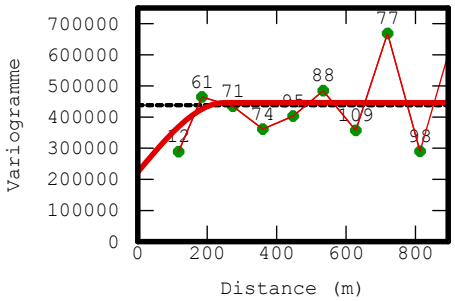
En pratique, il s'agit d'abord de réaliser une analyse exploratoire des données (plan de localisation, histogramme, etc.) pour **identifier les tendances** générales, **repérer les données anomales**, délimiter des zones aux propriétés différentes.





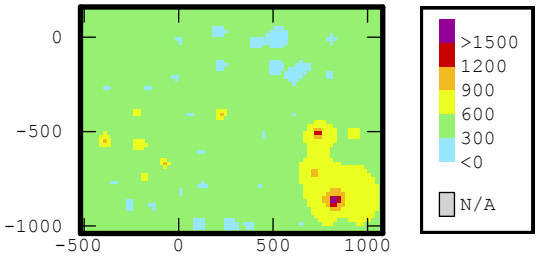
Il s'agit ensuite de **caractériser la variabilité spatiale** de la teneur, notamment la portée ou "distance de corrélation" et l'anisotropie du milieu.

Cette caractérisation permet d'**orienter le choix de la méthode d'interpolation :** on retiendra une méthode qui cale l'interpolation sur les données si les teneurs varient de façon continue (retombée de fumées d'usine par exemple) ou une méthode qui lisse les données si celles-ci comportent des erreurs de mesure ou des microstructures importantes. Ou mieux, on emploiera l'estimateur proposé par la géostatistique (le krigeage), qui s'adapte à l'un et l'autre cas ou à toute situation intermédiaire, car il fait intervenir la variabilité spatiale identifiée dans l'étape précédente. Cet estimateur permet de plus d'**assortir toute carte interpolée d'une carte de précision.**



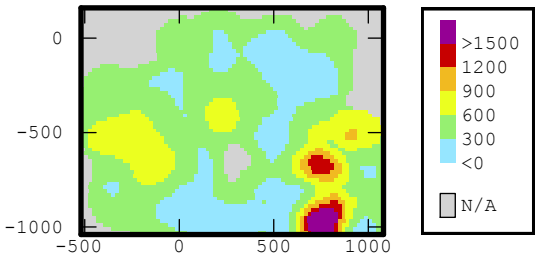
En évitant les biais, en améliorant et en quantifiant la précision de la cartographie de la teneur, l'approche géostatistique permet de mieux fonder les prises de décision.

Pour en savoir plus…

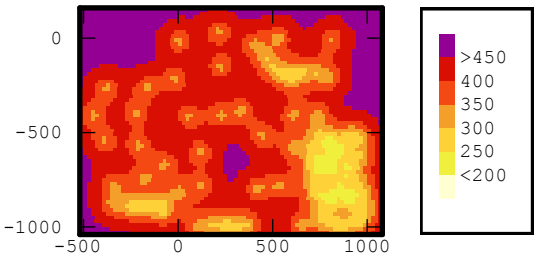


*Carte de la teneur **sans** géostatistique*

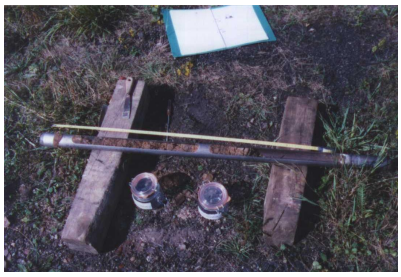
Concernant ce thème, le manuel méthodologique ***GeoSiPol*** fournit les lignes conductrices de l'analyse exploratoire des données, rappelle les notions élémentaires nécessaires à la cartographie des teneurs et en précise les conditions de mise en œuvre et le champ d'application. Cette approche, dont les performances ont déjà été largement testées et validées, est illustrée par un cas réel de pollution par des hydrocarbures dans des sols superficiels.



*Carte de la teneur **avec** géostatistique*



Incertitude associée à la carte géostatistique



Comment passer des analyses chimiques à la cartographie des teneurs en polluant ?

1 PRESENTATION DU THEME

1.1 Introduction

Le prélèvement d'échantillons sur un site potentiellement pollué et l'analyse chimique de leurs teneurs sont des opérations coûteuses. Aussi le nombre de données disponibles pour la caractérisation d'un site est-il toujours limité, même s'il est bien sûr plus élevé au stade d'un diagnostic approfondi en vue d'une **Evaluation Détaillée des Risques (EDR)** que pour un diagnostic initial (**Evaluation Simplifiée des Risques / ESR**). Dans les deux cas il importe de tirer le meilleur parti possible de l'information disponible pour :

- comprendre la structure de la pollution dans l'espace, et notamment la répartition des fortes teneurs ;
- cartographier cette pollution au mieux compte tenu de sa structure spatiale ;
- assortir cette cartographie d'un indicateur de précision.

Les données sont fréquemment hétérogènes du fait de campagnes espacées dans le temps, de protocoles d'échantillonnage différents ou d'analyses chimiques de différentes qualités. Il est donc nécessaire de mettre en cohérence les différents types d'investigation.

La méthodologie nationale juge que l'analyse des incertitudes est « essentielle et indispensable », mais l'approche à mettre en œuvre pour caractériser ces incertitudes reste trop souvent relativement floue. La géostatistique propose une méthodologie éprouvée pour y parvenir. En effet, elle permet une analyse cohérente des données, des incertitudes et des erreurs qui les entachent, ainsi que de la structuration spatiale de la teneur. Cette analyse permet à son tour de quantifier l'incertitude sur les cartes de teneur en polluant qui sont produites.

Dans les applications pratiques, on s'intéresse habituellement à plusieurs polluants et les analyses chimiques sont complétées par des informations auxiliaires (lithologie ; informations qualitatives ou organoleptiques ; mesures de terrain : prospection géophysique, mesure des gaz contenus dans le sol, kit chimique, etc.). Généralement moins précises mais plus nombreuses, ces dernières apportent une information indirecte sur la localisation de la contamination. La valorisation de ces informations auxiliaires faisant l'objet du thème N° 3, nous considérons ici le cas de données concernant une seule substance, ou du moins des substances qui peuvent être étudiées séparément.

1.2 Cadre métier

Le problème posé peut être formulé comme suit : **« Comment caractériser la répartition spatiale de la teneur en polluant sur le site et en déduire une cartographie de la teneur assortie de sa précision ? »**

Par le biais de la cartographie, on peut mieux identifier les zones polluées (ou non) et avoir une idée de leur extension, sans toutefois aller jusqu'à des calculs de volumes à dépolluer.

Ce problème s'inscrit bien sûr dans une problématique plus générale :

- En amont se pose la question de la conception de la campagne d'échantillonnage qui fournira les données nécessaires à l'étude. Ce point fait l'objet du thème N° 4.
- En aval on cherche à délimiter les zones à excaver ou à dépolluer et à évaluer leur volume. Ce point fait l'objet du thème N° 2.

1.3 Pratiques habituelles

Dans les études réalisées actuellement, l'analyse exploratoire et statistique des données est habituellement rudimentaire, quand elle n'est pas tout simplement ignorée. La tendance est de passer directement à une cartographie des teneurs dans les sols (toutes données confondues) en utilisant des méthodes d'interpolation ou d'estimation plus ou moins conventionnelles (pondération par l'inverse de la distance, moyenne mobile...). Ces pratiques sont la conséquence d'une démarche purement empirique, qui s'explique par le manque de méthodes ou de moyens des ingénieurs en environnement pour aborder l'analyse de données spatiales.

Plus généralement, l'analyse des données est mal perçue car elle ne constitue pas une fin en soi mais apparaît comme une simple étape intermédiaire. De cette valorisation des données dépendent pourtant pleinement la qualité de la cartographie et la quantification de sa précision.

On constate ainsi souvent un manque de fondement méthodologique dans l'exploitation des données :

- absence de différenciation des données de campagnes différentes alors que les méthodes d'échantillonnage (taille de la carotte,…) et d'analyse (précision) ont évolué ;
- absence de compréhension ou de prise en compte de la source initiale et de la mise en place de la pollution ;
- traitement indifférencié alors que les données font ressortir des zones aux caractéristiques distinctes (liées par exemple à la lithologie) ;
- manque de validation des données présentant des anomalies, qui peuvent correspondre à des pics de pollution ou à des erreurs de mesure ;
- absence de schéma conceptuel des facteurs de dispersion de la pollution dans les sols (solubilité...), préalable requis pour toute analyse géostatistique ;
- recours à des méthodes d'interpolation inadaptées à la structuration spatiale des teneurs ;
- choix inadéquat des paramètres d'interpolation ;
- appréciation erronée de l'incertitude associée aux résultats.

Cela peut avoir pour conséquence une évaluation erronée de la répartition de la pollution et conduire à des décisions inappropriées.

1.4 Intérêt d'une approche géostatistique

La démarche géostatistique comporte trois phases :

1. rechercher et comprendre la variabilité spatiale,
2. et la quantifier,
3. pour estimer les teneurs là où elles n'ont pas été mesurées.

L'analyse exploratoire (phase 1) consiste à :

- visualiser les caractéristiques de l'échantillonnage ;

- examiner les relations entre méthode d'échantillonnage, milieu et teneur pour éventuellement subdiviser les données en sous-ensembles plus homogènes ;
- élaborer des statistiques de base sur les teneurs ;
- valider les données présentant des anomalies ;
- caractériser la variabilité spatiale des teneurs, et en particulier préciser l'homogénéité ou l'hétérogénéité du site, ainsi que la portée ou "distance de corrélation" des variogrammes expérimentaux.

La phase 2 consiste à caler les paramètres synthétisant la distribution statistique et spatiale des teneurs, notamment par l'ajustement d'un modèle aux variogrammes expérimentaux.

Ces phases préliminaires permettent (phase 3) de réaliser une cartographie par krigeage, adaptée à la situation étudiée. Le krigeage s'adapte à la variabilité spatiale exprimée par le variogramme, ce qui lui confère une grande flexibilité, contrairement aux méthodes conventionnelles de cartographie. Le krigeage assortit de plus toute interpolation d'un écart-type qui caractérise sa précision.

Le bénéfice attendu, en améliorant la précision de l'estimation de la teneur en polluant, est de réduire à moindre coût l'incertitude qui lui est associée et par suite l'incertitude sur la délimitation et la quantification des volumes contaminés. Un sous-produit important de l'emploi de méthodes d'analyse spatiale des données est également une validation fine des données, qui permet d'éviter des biais dans l'estimation du niveau de pollution.

1.5 Objectifs

Ce chapitre a pour objectifs de :

- Fournir des lignes conductrices pour l'analyse exploratoire et statistique de données provenant d'un site pollué.
- Introduire les notions de base nécessaires à l'exploitation de ces résultats pour la cartographie des teneurs dans les sols.
- Illustrer la méthodologie proposée.

Elle est axée sur l'application de techniques géostatistiques **monovariables**. Lorsque plusieurs polluants sont mesurés aux mêmes points, ces techniques peuvent leur être appliquées séparément, les techniques multivariables ne permettant une amélioration de la cartographie que dans des cas très particuliers. Si les différents polluants ne sont pas tous mesurés sur les mêmes échantillons, les outils monovariables ici présentés gagnent à être complétés par les outils **multivariables** qui font l'objet du thème N° 3.

1.6 Méthodologie

Nous nous intéressons tout d'abord à l'analyse exploratoire et statistique des données, à la base de toute approche géostatistique et qui devrait être un passage obligé dans toute étude de site pollué. La réalisation de l'EDR santé oblige d'ailleurs pour partie à une analyse statistique des données. A ce titre, l'analyse des données constitue une phase préliminaire à laquelle font référence tous les autres thèmes. La cartographie de la teneur en polluant est présentée ensuite.

Sont donc traités ici :

- L'analyse exploratoire des données (teneur, ou autre variable comme par exemple une épaisseur) au moyen d'outils statistiques et géostatistiques très simples (plans de localisation, histogrammes, variogrammes, etc.) mais potentiellement très puissants lorsqu'ils sont utilisés de façon interactive.
- Le choix et l'ajustement de modèles statistiques qui décrivent la structure spatiale de la contamination.

- L'exploitation de ces modèles pour une cartographie des teneurs dans les sols (estimation des teneurs par krigeage) et une quantification de l'incertitude locale sur ces teneurs (écart-type de krigeage).

2 L'ANALYSE EXPLORATOIRE DES DONNEES

Nous nous plaçons dans le cadre de l'étude : un polluant dont la teneur est connue en un nombre limité de points et dont nous souhaitons faire la cartographie.

2.1 Contexte de l'analyse exploratoire

L'analyse exploratoire des données intervient à l'issue des différentes phases d'échantillonnage, avec leurs lacunes et leurs imperfections ou leurs améliorations : changements du mode de prélèvement ou de la méthode d'analyse, évolution du plan d'échantillonnage, notamment avec la profondeur. La compréhension de la répartition de la pollution sur le site nécessite de tenir compte de l'ensemble des informations disponibles, en particulier :

- du cadre géologique et hydrogéologique : localisation et propriétés des formations présentes sur le site ;
- de l'historique du site : activités successives, traitement du site lors de sa fermeture.

L'objectif est de comprendre et de représenter (en particulier graphiquement) comment la pollution s'organise sur le site : implantation des différents niveaux de teneur, régularité ou contrastes des valeurs en des points d'échantillonnage voisins, variation des teneurs avec la profondeur ou avec les types de sols, etc. L'analyse exploratoire prépare ainsi l'étape de cartographie. La méthode consiste en une interrogation interactive des données.

Sous réserve que l'échantillonnage initial soit adapté aux caractéristiques du site, la structure spatiale synthétisée par le variogramme indiquera la possibilité ou non d'une délimitation précise des zones à dépolluer, sur la base de l'information actuellement disponible ou future. Tout autant que le krigeage des teneurs et sa cartographie, la possibilité d'adapter la maille de reconnaissance aux caractéristiques de la pollution particulière étudiée constitue un résultat utile et immédiat de l'analyse exploratoire des données.

2.2 Information sur le site et sur l'échantillonnage

Une description du site est indispensable, à la fois sur le milieu initial (cadre géologique et hydrogéologique) et sur ses transformations ultérieures (étude historique). Elle doit être complétée par une description précise de l'échantillonnage.

2.2.1 Cadre géologique

La façon dont le site est constitué explique pour partie la répartition spatiale des teneurs.

Une ou plusieurs formations (sols, roches) sont-elles présentes sur le site ? Comment sont-elles disposées : suivant une stratification verticale régulière par niveaux subhorizontaux, ou le site recouvre-t-il latéralement différentes formations ? Quelles sont les propriétés de ces formations : par exemple un niveau imperméable (argiles, substratum rocheux) limite-t-il l'extension verticale de la pollution ?

Une description succincte du fonctionnement hydrogéologique et hydrogéochimique du site est également utile.

- A quelle profondeur se situe la nappe ?
- Si la nappe est peu profonde, baigne-t-elle ou a-t-elle baigné les niveaux pollués, et sur quelle épaisseur ? Quelle est la hauteur de battement de la nappe ? Quel en est le sens d'écoulement ?

- Le site présente-t-il des conditions géochimiques particulières (par exemple pH) pouvant expliquer la mise en solution, ou au contraire la précipitation de certaines substances ?

2.2.2 Modifications du site

L'industrialisation d'un site provoque des remaniements plus ou moins importants des sols avec l'apport de matériaux divers, notamment de remblais. Par ailleurs, les polluants recherchés sont liés aux activités passées. Le niveau des teneurs et leur localisation dépendent donc des activités successives, puis des transformations ultérieures du site.

- Durant l'occupation du site : quelles ont été les principales matières premières utilisées, et les principales productions ? Où se situaient les infrastructures majeures (transports, manutention, stockages) ?
- Comment le site a-t-il été remanié lors de sa fermeture ou du démantèlement des installations : nivellement, mélange des sols, remblaiement des fossés, etc. ?

Un cheminement préférentiel le long de certaines infrastructures (canalisations enterrées, fondations) peut donner lieu à de fortes teneurs très localisées, pas toujours détectables à partir d'un échantillonnage un peu lâche. Il est donc utile de savoir si les infrastructures ont été remaniées lors du démantèlement ; par exemple, jusqu'à quelle profondeur les sols ont-ils été creusés ou remblayés ?

L'étude variographique peut en effet mettre en évidence des remaniements locaux, des contrastes entre teneurs d'échantillons voisins explicables par des apports ou par une couverture de sols « propres » etc. Toutefois, il est dommage d'effectuer un examen exploratoire détaillé des données pour conclure finalement à l'absence de structure spatiale, et donc à l'impossibilité de toute sélection efficace, avant de « se souvenir » que le site en question a été totalement remanié et les sols mélangés avec différents remblais lors du démantèlement des installations. Il aurait été possible de se rendre compte très tôt que la démarche était inutile !

2.2.3 Informations sur le schéma d'échantillonnage

L'examen exploratoire s'effectue sur les données disponibles, principalement les résultats d'analyses des prélèvements. Les particularités de cet échantillonnage ont des conséquences sur les conclusions de cet examen.

A quel objectif répond l'échantillonnage disponible : rechercher et caractériser les différents sols ou types de remblais présents sur le site ? Ou a-t-il été effectué dans le but de cartographier l'épaisseur d'un niveau de remblais, ou bien de quantifier les teneurs ? En particulier, l'échantillonnage est-il **systématique** (cas par exemple d'une maille régulière) ou bien **préférentiel**, par exemple pour identifier une source de contamination suspectée suite à l'historique du site ? Nous verrons qu'un échantillonnage irrégulier mais non préférentiel ne pose en général pas de problème particulier. Un échantillonnage préférentiel impose en général des corrections statistiques, et peut à l'extrême être inutilisable pour un traitement statistique.

Dans le cas où les échantillons sont implantés irrégulièrement, il convient de préciser si cet échantillonnage est ou non préférentiel.

Un échantillonnage lacunaire peut rendre impossible une bonne caractérisation de la pollution. Par exemple, un échantillonnage limité à 1,50 m de profondeur, ne fournit pas d'information sur les valeurs des teneurs au-delà de cette profondeur.

2.2.4 Caractéristiques des prélèvements et des analyses

La description précise des échantillons est également indispensable : coordonnées x,y dans le plan, profondeur ou cote. Lorsque seule la profondeur est indiquée, il faut savoir si la surface du site est « suffisamment » horizontale ou plane : des échantillons à même profondeur sont-ils situés au même niveau ?

On montre en géostatistique que la variabilité des teneurs mesurées dépend du **support** des échantillons, c'est-à-dire de leur volume et de leur géométrie. Pour les calculs statistiques (variance, histogrammes, variogramme), on ne peut donc pas mélanger des échantillons de taille trop variable,

sauf cas particulier de l'étude de l'accumulation sur une couche, produit de la teneur par l'épaisseur de cette couche (cf § 2.3.4.).

Il faut donc savoir si les prélèvements sont tous identiques, et préciser leur support. Est-il quasi « ponctuel » (quelques dm^3), ou bien l'analyse est-elle effectuée sur un tronçon de carotte ? Quels sont le volume, et plus précisément la longueur et le diamètre des échantillons ?

Quelle partie des échantillons est analysée ? Est-ce l'ensemble de l'échantillon, après broyage et homogénéisation, ou bien y a-t-il eu un tamisage préalable, et dans ce cas à quelle granulométrie ? Les débris divers (scories ; matériaux de construction, notamment les briques qui absorbent certaines substances) ont-ils été enlevés ?

Lorsque différentes formations (types de sols, remblais) sont présentes sur le site et correspondent à différents niveaux de teneurs, il est nécessaire de préciser la **matrice** de chaque échantillon (argile, fragments de briques …). En effet, on est parfois amené à traiter séparément certaines formations. Ainsi, dans un milieu stratifié, lorsque la pollution se localise préférentiellement dans une formation, il faut cartographier cette formation pour déterminer les volumes pollués : cotes du toit et du mur ou épaisseur de la couche.

Lorsqu'un niveau imperméable bloque la migration verticale de la pollution, il est nécessaire d'estimer la topographie du toit de ce niveau, ou plutôt l'épaisseur de la zone polluée, pour évaluer le volume des sols contaminés. Ce niveau imperméable a-t-il ou non été recherché sur chaque sondage ?

Lorsque **plusieurs campagnes** se sont succédées, avec des intervenants différents, le mode d'échantillonnage peut avoir varié, et la précision des analyses s'être améliorée. Dans ce cas, il est nécessaire de distinguer les différentes campagnes lors de l'examen exploratoire ; l'ensemble des données est ensuite pris en compte de façon différenciée pour l'estimation.

Lorsque les teneurs visées sont très faibles, comme pour certains métaux, les **limites de détection** ou de quantification ont-elles évolué entre les différentes campagnes ?

*Enfin, l'estimation s'effectue généralement sur les teneurs, alors que le risque sanitaire dépend aussi de la **spéciation**, c'est-à-dire de la phase ou de la forme chimique de la substance. A même teneur, ce risque n'est pas le même suivant la solubilité de la substance ou de manière plus générale sa biodisponibilité. Actuellement, il n'est guère possible d'en tenir compte pour l'estimation, les mesures correspondantes n'étant pas disponibles. Bien que très important, cet aspect reste en effet négligé lors de l'échantillonnage.*

2.3 Les étapes de l'analyse exploratoire

Dans la suite, l'échantillonnage est supposé régilier ou modérément irrégulier, et non préférentiel. Le cas d'un échantillonnage très irrégulier, plus problématique, sera simplement évoqué.

2.3.1 Localisation des données

La première étape consiste à visualiser la localisation des données sur le site, afin de :

- vérifier l'absence d'erreurs grossières dans les coordonnées des échantillons ;
- préciser les caractéristiques de l'échantillonnage (régularité, etc.…) ;
- fournir une première indication de la répartition des teneurs sur le site, en représentant les différentes classes de teneur par des symboles différents.

Cette visualisation devrait être effectuée systématiquement, en codant les différentes informations comme des variables supplémentaires. Ceci permet de visualiser comment s'organisent spatialement :

- les classes de teneur des substances étudiées ;
- la topographie, lorsque le site n'a pas été nivelé ;
- les différents types de sols : localisation, profondeur, épaisseur ;
- les différentes campagnes successives de reconnaissance et les différents types de prélèvement (fosses, sondages …).

Il est utile de faire apparaître différemment les points d'échantillonnage renseignés ou non pour la variable représentée. Pour améliorer la lisibilité des graphiques, on utilise différents symboles, en faisant varier leurs dimensions ou leurs couleurs. La représentation devrait être complétée par le tracé des caractéristiques majeures du site : limites, cours d'eau, infrastructures actuelles (accès routiers, ferroviaires …) ou anciennes (implantations des différentes unités industrielles).

Une visualisation par plans, complétée par des sections verticales est utile dans le cas d'une reconnaissance en trois dimensions, en particulier pour repérer :

- la régularité ou l'irrégularité de l'échantillonnage vertical ;
- d'éventuelles variations de profondeur ou de cote des échantillons suivant les endroits ;
- la présence de niveaux non reconnus ou au contraire échantillonnés préférentiellement.

Les effectifs donnés par les statistiques élémentaires sont utilement complétés par les histogrammes de variables synthétiques (exemple : année de la campagne, codage numérique des différentes formations pédologiques, y compris les différents types de remblais) ; certains logiciels permettant de lier les histogrammes aux cartes d'implantation, aident à repérer la localisation des différentes campagnes, les profondeurs échantillonnées, la localisation des différentes formations, etc…

Cette visualisation permet de mieux comprendre le site. Elle pose généralement de nouvelles questions :

- Si l'échantillonnage est localement plus dense, pour quelles raisons la maille des prélèvements a-t-elle été localement resserrée ?
- Si la reconnaissance verticale est irrégulière, certains niveaux sont-ils peu ou pas reconnus ?
- Verticalement, les échantillons sont-ils prélevés au sein d'un niveau lithologique bien défini, ou selon un découpage indépendant de la lithologie ?
- La profondeur échantillonnée varie-t-elle selon les campagnes ? Est-elle liée à l'extension verticale des pollutions, ou à la présence d'éléments particuliers : formation pédologique, présence de la nappe ?

Exemple : La Figure I-1 présente l'histogramme de la profondeur des échantillons d'une première campagne de reconnaissance sur un site pollué par des Hydrocarbures Aromatiques Polycycliques (HAP) ainsi que le nuage de corrélation entre profondeur et teneur. La moyenne du logarithme de la teneur par classe de la profondeur (la régression empirique) n'indique aucune décroissance de la teneur avec la profondeur : la valeur plus faible pour la classe de profondeur supérieure à 3 mètres n'est pas significative, cette classe de profondeur étant réduite à un unique échantillon. En l'absence d'information complémentaire (présence d'un substratum rocheux par exemple) limiter l'échantillonnage à moins de trois mètres parait donc arbitraire et rend impossible la quantification du volume total des sols pollués. La quantification des teneurs n'est possible que sur une épaisseur de sols inférieure à 2 mètres environ, seuls sept échantillons ayant été prélevés à une profondeur supérieure à deux mètres.

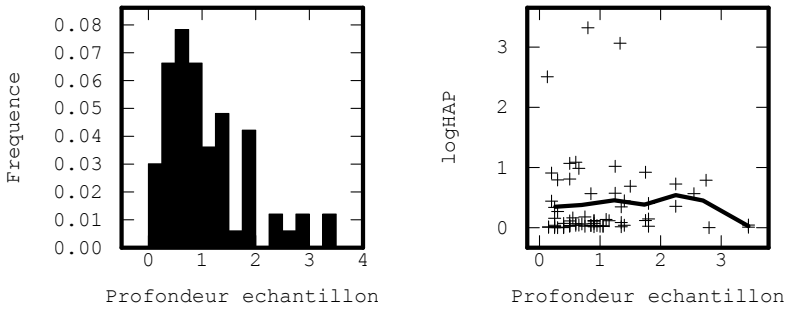


Figure I-1 Histogramme de la profondeur des 69 échantillons et nuage de corrélation entre logarithme de la teneur en HAP et profondeur.

Notons que l'utilisation du logarithme permet d'obtenir des statistiques moins sensibles aux rares teneurs très élevées. Le logarithme décimal présente également l'intérêt de fournir directement l'ordre de grandeur des teneurs mesurées (le logarithme décimal de 10^n vaut n). Un problème se pose lorsque certaines données sont

nulles, leur logarithme valant alors $-\infty$; on le résout en translatant les mesures d'une petite quantité, par exemple de 1 mg/kg ms (la transformée des teneurs faibles est alors positive).

2.3.2 Localisation spatiale des teneurs

Une fois les échantillons localisés, avec indication des différentes campagnes, des formations pédologiques ou autres informations utiles, nous nous intéressons à la localisation des différentes classes de teneur, en relation éventuellement avec une valeur limite réglementaire.

Les **statistiques élémentaires** des teneurs (minimum, maximum, médiane, moyenne, voir plus loin) sont utiles. On reprendra ces statistiques en distinguant suivant les cas selon la profondeur échantillonnée (entre 0 et 1m, 1 et 2m, etc…), selon la localisation (de part et d'autre d'un accès routier ou ferroviaire, d'un canal ; à proximité d'une ancienne unité de production …), pour faire apparaître des niveaux ou des zones de valeurs plus fortes.

Un choix adapté des classes de l'histogramme fournit directement la proportion des échantillons appartenant à des intervalles de teneur fixés (par exemple inférieure à 500 mg/kg ms (mg/kg), ou comprise entre 500 et 1000). Si le support des échantillons change significativement avec les campagnes de reconnaissance, les échantillons des différentes campagnes n'ont pas la même variabilité et ne doivent pas être mélangés sans étude préalable pour le calcul des statistiques élémentaires ou d'histogrammes.

Lorsque l'étendue des valeurs est importante, l'histogramme des teneurs est parfois peu lisible : l'effectif des classes des teneurs les plus fortes étant généralement réduit, ces classes ne sont pas toujours visibles sur l'histogramme, qui semble se limiter aux seules premières classes, celles des valeurs faibles. Le passage au logarithme des teneurs améliore la lisibilité de l'histogramme, faisant éventuellement apparaître plusieurs modes.

La présence de plusieurs modes sur l'histogramme des teneurs est un indice important. Un histogramme multimodal est symptomatique de la présence de différentes « populations statistiques », pouvant correspondre :

- à différentes origines de la pollution ou à des zones de niveaux de teneur bien différenciées (procédés de production ou types de rejets variables au cours du temps, présence au cours du temps d'unités de production différentes ou localisées en plusieurs endroits…)
- de l'influence des différentes formations pédologiques, selon que la pollution y est plus ou moins mobile ;
- Un mode peut cependant constituer un artefact, résultant d'un resserrement local de la maille de reconnaissance, qui augmente artificiellement la proportion des échantillons autour de certaines valeurs. Lier l'histogramme à la carte d'implantation, et examiner l'échantillonnage dans des coupes verticales, aide à détecter cette situation (voir l'exemple en fin de chapitre).

Lorsque l'histogramme des teneurs est fortement dissymétrique, la multimodalité apparaît plus clairement sur l'histogramme du logarithme des teneurs.

Exemple : les figures suivantes présentent un exemple d'échantillonnage irrégulier d'un site pollué par des HAP, avec des prélèvements plus resserrés au Nord-Ouest du site. L'histogramme des teneurs mesurées (Figure I-2) indique qu'environ 80% des valeurs mesurées sont inférieures à 250, et environ 15% sont comprises entre 250 et 500. Sur les 66 teneurs mesurées, les trois valeurs restantes se répartissent entre 2000 et 6500. Ainsi, avec un choix pertinent de la largeur des classes, la simple lecture des axes d'un histogramme donne de l'information sur l'étendue des valeurs mesurées et sur leur « distribution ».

En liant interactivement l'histogramme et la carte d'implantation, on observe que les maxima sont plutôt regroupés vers le centre de la zone échantillonnée de façon dense, les minima se situant vers le pourtour du site. Cependant, des valeurs fortes se retrouvent à peu près partout sur la carte, notamment vers l'est, échantillonné de façon moins dense.

Les choses sont plus claires en échelle logarithmique. L'histogramme du logarithme translaté (Figure I-2) apparaît bimodal, avec un premier mode autour de 1 (teneur proche de 10) et un second autour de 2,5 (teneur proche de 300). Visualisées sur la carte d'implantation, les teneurs du mode supérieur se situent plutôt vers le centre de la zone échantillonnée, avec un échantillon isolé vers l'est ; les teneurs du mode faible se rencontrent partout sur le site, y compris à proximité immédiate de valeurs plus fortes

On voit ainsi que les teneurs supérieures à 1000 mg/kg ms sont regroupées vers le centre de la zone où les valeurs sont supérieures à 100 mg/kg ms.

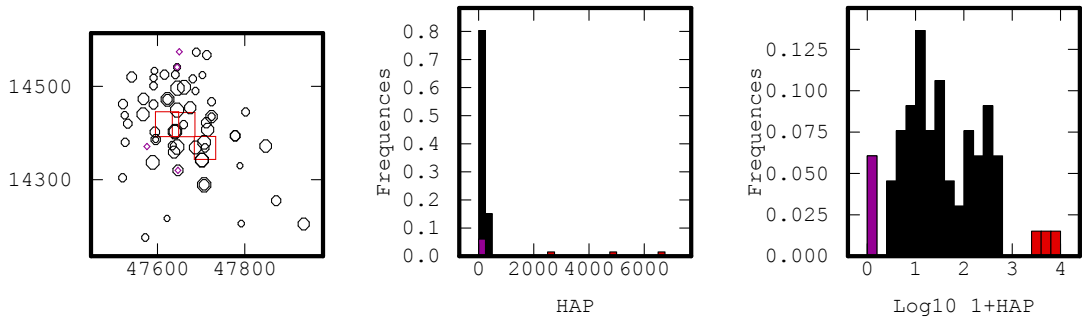


Figure I-2 Carte d'implantation des prélèvements. Histogramme des teneurs en HAP et histogramme de leur logarithme décimal translaté. La dimension des symboles est proportionnelle au logarithme de la teneur. Les symboles superposés correspondent à plusieurs prélèvements le long de la verticale. Les carrés indiquent les valeurs supérieures à 1000 mg/kg ms, et les losanges les valeurs égales au minimum mesuré (0,2). La largeur des classes de l'histogramme des teneurs est de 250 mg/kg ms.

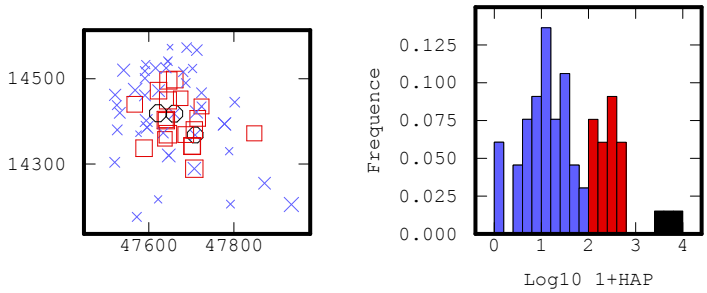


Figure I-3 Bimodalité de l'histogramme du logarithme des teneurs, et localisation des deux modes. Les cercles correspondent aux teneurs supérieures à 1000 mg/kg ms (logarithme supérieur à 3), les carrés aux teneurs comprises entre 100 et 1000 mg/kg ms (logarithme supérieur à 2), et les croix aux teneurs inférieures à 100 mg/kg ms.

En généralisant cet exemple, l'examen simultané de l'histogramme des teneurs, de la carte d'implantation, ainsi que les nuages de corrélation entre teneur et coordonnées (avec tracé de la moyenne par classe) permettent entre autres de répondre aux questions suivantes :

- Les fortes teneurs sont-elles localisées sur le site, éventuellement à une certaine cote ou à une certaine profondeur ? Apparaissent-elles regroupées localement en une ou quelques zones, ou au contraire sont-elles disséminées sur une partie ou tout le site ?
- Observe-t-on une proximité des faibles et fortes teneurs, ou des transitions par des valeurs intermédiaires ?
- Observe-t-on une variation systématique dans une direction, notamment suivant la verticale. Dans ce cas, le contraste est-il mieux marqué si on le visualise en fonction de la cote d'altitude ou de la profondeur ?

L'histogramme peut faire apparaître quelques **valeurs particulièrement fortes** (ou faibles), très différentes des autres valeurs mesurées. De telles « anomalies » devraient toujours être contrôlées : correspondent-elles à la réalité, et dans ce cas, de quoi sont-elles éventuellement symptomatiques (remblai local, enfouissement de déchets ….) ? Ou bien correspondent-t-elles à des erreurs, par exemple de report des valeurs ?

L'intérêt d'un examen précoce des données est évident : pouvoir retourner sur le site pour contrôler les anomalies détectées est parfois indispensable.

Une fois précisées les caractéristiques de l'échantillonnage, la localisation des classes de teneur sur le site, y compris en profondeur, et leurs relations avec « le milieu », notamment les différents types de sols, orientent la suite de la modélisation :

- Le site est-il considéré comme **homogène**, ou est-il nécessaire de **distinguer différentes zones**, et dans ce cas, suivant quels critères ?
- Verticalement, doit-on considérer une seule formation ou plusieurs couches ? L'épaisseur de chaque couche est-elle à peu près constante ou au contraire variable ? Dans le cas d'une seule « couche », les teneurs présentent-elles une variation systématique avec la profondeur (croissance ou décroissance, présence d'un niveau plus fortement pollué) ?
- La géométrie de la zone à étudier est-elle fixée ou fait-elle partie des inconnues ? Par exemple, veut-on caractériser les teneurs dans les trois premiers niveaux de 0,5 m d'épaisseur d'un site nivelé, ou veut-on d'abord évaluer l'épaisseur variable d'une couche de remblais ?

Les réponses à ces questions orientent l'étude vers une modélisation à 2 ou 3 dimensions (notée 2D ou 3D).

Les nuages de corrélation entre teneurs et coordonnées (en particulier la coordonnée verticale) peuvent utilement faire apparaître des zones ou des niveaux (des « strates ») de valeurs plus fortes.

2.3.3 Statistiques élémentaires

Les statistiques élémentaires : nombre d'échantillons, minimum, maximum, moyenne, variance ou écart-type, ainsi que la médiane et les quartiles, complétés par le coefficient de variation (rapport de l'écart-type à la moyenne), fournissent un résumé statistique de la teneur. Par exemple, la comparaison des moyennes, minimums et maximum pour deux types de sols peut faire apparaître des valeurs systématiquement plus élevées pour l'un d'eux.

Ces statistiques sont cependant sensibles :

- aux irrégularités de l'échantillonnage ;
- à l'extension de la zone échantillonnée (c'est le cas en particulier pour la variance et l'écart-type) ;
- au support, c'est-à-dire au volume et à la géométrie de l'échantillon : toutes choses égales par ailleurs, la variance diminue lorsque le support de l'échantillon augmente. Comme pour tout autre calcul, on ne peut simplement mélanger des données de support différent.

Ces statistiques, toujours utiles, sont à relier à l'implantation des données, aux différents ensembles mis en évidence (types de sondages, types de sols, etc.) et aux histogrammes.

2.3.4 Choix des variables d'étude

Des variables doivent être choisies pour évaluer et modéliser l'état de contamination des sols. Ces variables (par exemple des teneurs en polluant) peuvent être définies de multiples façons. Le choix dépend :

- des caractéristiques du site (géométrie des formations présentes sur le site, caractéristiques des teneurs, variations avec la profondeur) ;
- des caractéristiques de l'échantillonnage ;
- éventuellement, de contraintes techniques sur les méthodes de dépollution.

Si les sols sont pollués sur une épaisseur importante, telle qu'il soit possible de différencier plusieurs niveaux lors du traitement du site, on utilisera plutôt la teneur sur un volume constant. Par exemple, avec des sondages verticaux découpés et analysés par tronçons de 0,50 m de hauteur, la variable étudiée sera la teneur moyenne sur ces tronçons, ou sur une épaisseur multiple de cette hauteur (1 m, 1,50 m d'épaisseur). Une modélisation tridimensionnelle sera alors nécessaire.

Dans le cas d'un niveau d'épaisseur constante et assez faible, s'il n'est pas prévu de différencier les teneurs verticalement lors du traitement, la variable retenue sera la teneur sur l'épaisseur du niveau (sous réserve que l'échantillonnage ait été effectué de façon systématique sur toute la hauteur de la couche). Elle pourra être modélisée en deux dimensions.

Si les sols constituent une couche fine d'épaisseur variable, on étudiera plutôt :

- l'épaisseur de la couche ou la cote de sa limite inférieure et celle de sa limite supérieure ;

- « l'accumulation », produit de la teneur moyenne par la hauteur de la couche correspondante. En effet, la teneur moyenne sur un niveau d'épaisseur variable n'est pas additive car elle doit être pondérée par la hauteur variable de la couche (§3.3).

Dans la suite, la variable étudiée sera principalement la teneur et les modèles pourront être à 2 ou 3 dimensions. Les méthodes se transposent à l'estimation d'une cote topographique ou d'une épaisseur.

Lorsque de nombreuses substances sont présentes simultanément, on peut :

- soit modéliser chacune d'elles ; c'est par exemple le cas pour une pollution métallique multiple, en particulier d'origines ou de propriétés (solubilité) différentes.
- Modéliser la somme, par exemple les « hydrocarbures totaux », pour une première caractérisation globale. Lorsque la méthode de traitement du site varie suivant la teneur de certaines substances, une modélisation multivariable de plusieurs substances ou groupes de substances, plus précise, est parfois nécessaire (voir le thème N° 3).
- Modéliser un composant particulier, choisi pour ses propriétés. Par exemple, le benzo (a)pyrène B(a)P est parfois retenu comme indicateur de la famille usuelle des HAP.

2.3.5 Variabilité spatiale des teneurs : vers le variogramme

Quelle information la teneur mesurée en un point apporte-t-elle sur la valeur en un point voisin ? Par exemple le zinc, très soluble, présente des teneurs variant plutôt régulièrement dans l'espace, tandis que les teneurs en HAP peu ou pas solubles peuvent dans certains cas présenter de très forts contrastes.

La **variabilité spatiale** des teneurs (ou de l'épaisseur des remblais) est donc une caractéristique importante d'un site. Elle est étudiée à l'aide des outils variographiques.

Nuée variographique

La corrélation décrit le degré de liaison entre deux variables, par exemple la teneur mesurée en deux échantillons distants de 100m. A la notion de corrélation, on préfère celle plus générale de variabilité, qui mesure l'amplitude de l'écart entre les valeurs.

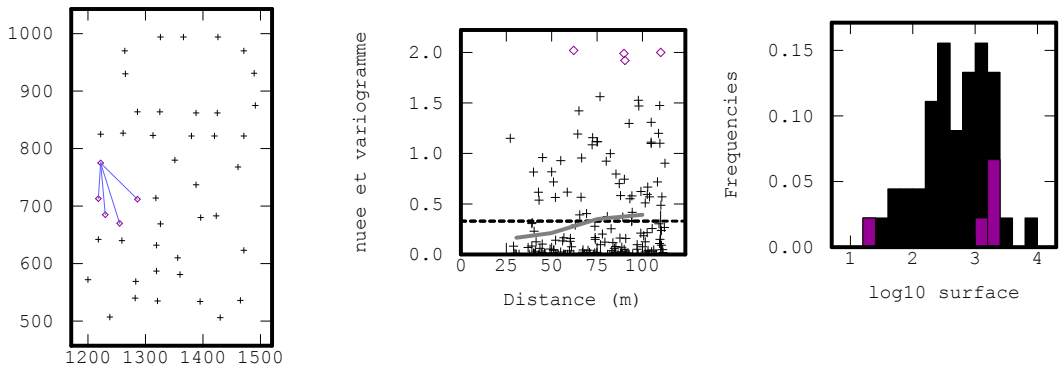


Figure I-4 Carte d'implantation, nuée variographique et histogramme. Logarithme décimal des teneurs en HAP. Les quatre écarts importants repérés en haut de la nuée variographique correspondent aux quatre couples de points reliés par un segment sur la carte d'implantation. Leur forte variabilité s'explique par le fort contraste des teneurs : le point commun aux quatre segments correspond sur l'histogramme au minimum et les quatre autres points à des valeurs fortes. Les cinq échantillons associés aux quatre écarts sur la nuée variographique sont reportés en mauve sur les graphiques.

La nuée variographique est le nuage de corrélation du demi écart quadratique $\gamma_{ij}=½[z(x_i)-z(x_j)]^2$ fonction de la distance d_{ij} entre les points de prélèvements x_i et x_j (Figure I.4).

Une nuée variographique se présente comme un nuage de points. Le moyen le plus simple pour la résumer consiste à en calculer la moyenne par classe de distance (voir la courbe tracée sur la nuée,

Figure I-4). Cette courbe croit généralement avec la distance h et se stabilise parfois à partir d'une certaine distance. C'est le variogramme.

Variogramme

Le variogramme permet de quantifier la variabilité en fonction de la distance.

Le variogramme est défini expérimentalement comme la moyenne sur le champ étudié, du demi carré de l'écart entre les valeurs mesurées en deux points, en fonction de la distance et de l'orientation séparant ces deux points

Le variogramme est la moitié de l'écart quadratique moyen en fonction de la distance h

$$\gamma(h) = \frac{1}{2}\frac{1}{n(\vec{h})}\sum_i \left(z(x_i+\vec{h}) - z(x_i)\right)^2$$

avec : - $n(\vec{h})$, nombre de couples de points distants de $\vec{h}$

- $\left(z(x_i+\vec{h}) - z(x_i)\right)^2$ écart quadratique des teneurs aux points x et $x+\vec{h}$

- $\frac{1}{n(\vec{h})}\sum_i \left(z(x_i+\vec{h}) - z(x_i)\right)^2$ moyenne de ces écarts quadratiques pour tous les couples de points distants de $\vec{h}$ disponibles

Le coefficient $\frac{1}{2}$ simplifie les calculs ultérieurs.

Le variogramme fournit beaucoup d'information sur la structure spatiale de la variable étudiée. Pour une même distance h, une variabilité plus forte entraîne une plus forte valeur du variogramme. Le comportement du variogramme, par exemple sa croissance plus ou moins rapide, dépendent donc de la variabilité spatiale de la variable étudiée.

2.3.6 Calcul du variogramme expérimental

Par le calcul du variogramme expérimental, nous chercherons à mettre en évidence les propriétés structurales (§ 2.3.7) que nous allons modéliser par la suite (§ 2.3.8).

Lorsque la maille est régulière, on calcule d'abord le variogramme suivant les directions principales de la grille, y compris les bissectrices dans le plan. Lorsque la distance augmente, le nombre de couples de points disponibles pour une distance donnée tend à diminuer (à maille régulière sur un segment, ce nombre diminue de 1 à chaque pas).

Lorsque la maille est irrégulière, on introduit les notions de tolérance sur le pas et de tolérance angulaire.

Pour l'estimation, seules interviennent en pratique les valeurs du variogramme jusqu'à une certaine distance. Pour cette raison, on limite le calcul du variogramme expérimental aux distances pour lesquelles le nombre de couples de points reste « suffisant ».

Le choix des paramètres pour le calcul des variogrammes expérimentaux (distances, orientations) est important, des conditions de calcul inadéquates pouvant masquer la structure spatiale. Ce choix dépend du phénomène étudié et de la maille de reconnaissance.

Nous verrons par la suite (§2.3.7) que le comportement du variogramme aux petites distances reflète la régularité de la variable étudiée. Pour cette raison, il est utile de pouvoir disposer d'une information sur la variabilité à petites distances, en rapport avec la résolution souhaitée pour la cartographie.

En pratique, on recommande donc de **resserrer localement la maille d'échantillonnage** dès les premières campagnes, en quelques endroits du site, dans des zones de teneurs fortes et de teneurs intermédiaires. Ceci permet d'affiner l'ajustement du variogramme, et également de préciser la présence d'un éventuel effet proportionnel (voir l'exemple au §4).

2.3.7 Le variogramme, outil structural

Le variogramme expérimental fournit des indications sur la régularité de la variable étudiée, sur sa stationnarité, son éventuelle anisotropie et plus généralement sur ses caractéristiques spatiales.

Stationnarité

La variable est stationnaire lorsque certaines de ses statistiques (moyenne, variance, et variogramme) peuvent être supposés comme variant peu sur le domaine étudié. Cela revient à considérer que la teneur fluctue autour d'une moyenne identique sur tout le champ. Si la teneur augmente avec la profondeur, elle n'est pas stationnaire verticalement. Si la teneur augmente fortement à proximité d'anciennes infrastructures (aires de stockage ou de manutention), elle n'est pas stationnaire latéralement.

Lorsqu'à partir d'une certaine distance le variogramme est croissant ou décroissant la variable n'est pas stationnaire sur le champ considéré.

La stationnarité dépend de l'échelle d'observation. Les teneurs peuvent ne pas être stationnaires sur l'ensemble du champ, avec la présence de quelques zones de fortes teneurs bien localisées, et être stationnaires à l'intérieur de ces zones, fluctuant localement autour d'une moyenne beaucoup plus élevée que sur le reste du site.

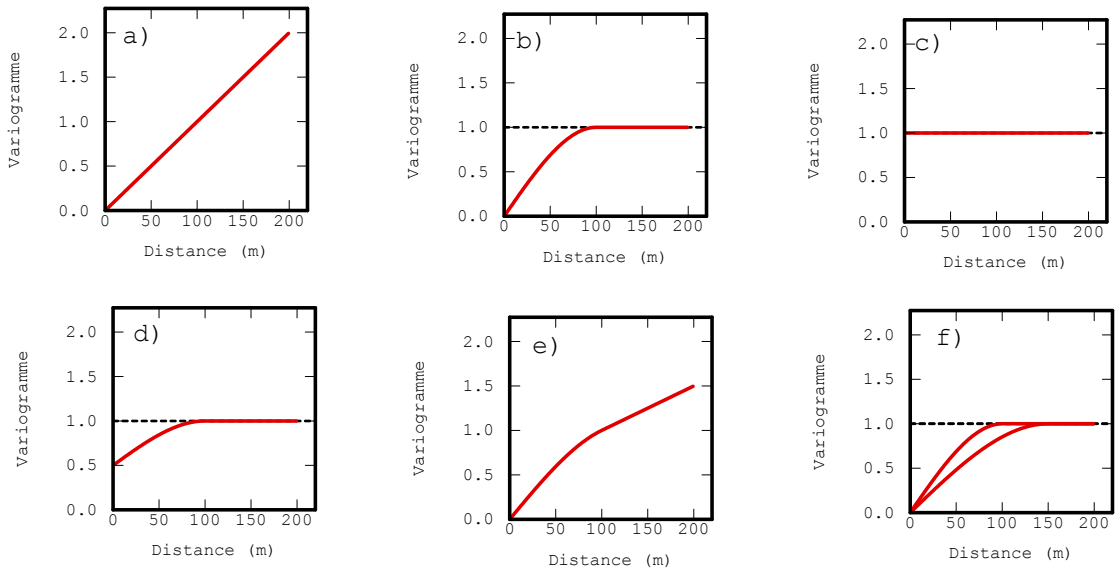


Figure I-5 Différents modèles de variogramme. a) modèle linéaire, correspondant à une variable non stationnaire ; b) modèle « sphérique » de portée a=100m et de palier 1 ; c) effet de pépite pur de palier 1 ; d) modèle gigogne, somme d'un effet de pépite et d'un schéma sphérique ; e) modèle gigogne, somme d'un schéma sphérique et d'un schéma linéaire ; f) modèle d'anisotropie géométrique : la portée varie selon les directions.

Palier, portée, structures gigognes

Lorsque la variable est stationnaire, le variogramme oscille, au-delà d'une certaine distance, autour d'une valeur constante dénommée **palier**. La **portée**, ou « distance de corrélation », est la distance à partir de laquelle ce palier est atteint. Elle correspond à la zone d'influence d'un échantillon.

Le plus souvent, plusieurs échelles de variabilité se superposent ; le variogramme peut se décomposer en plusieurs **structures gigognes**, avec chacune sa propre portée et son propre palier pour les composantes stationnaires.

Isotropie

Il y a **isotropie** lorsque la variabilité présente les mêmes caractéristiques dans toutes les directions.

En général, un milieu stratifié n'est pas isotrope, la variabilité verticale étant différente de la variabilité horizontale. La règle générale est l'**anisotropie**, entre le plan horizontal et la direction verticale, mais

également dans le plan horizontal. En revanche, lorsqu'un site a été remanié, et les matériaux mélangés sur une profondeur suffisante, la teneur peut être isotrope, au moins aux petites distances.

Si les variogrammes calculés dans les différentes directions se superposent, on supposera le phénomène isotrope (dans le plan, par exemple). On recalcule alors un variogramme moyen, en mélangeant les couples de points d'orientations différentes et en tenant compte seulement de la distance.

La **carte variographique** (Figure I-6) est un moyen très commode pour vérifier la présence d'éventuelles anisotropies. Elle consiste à représenter à l'aide de couleurs la valeur du variogramme en fonction de la distance et de l'orientation. Lorsque le phénomène est isotrope, le variogramme dépendant seulement de la distance mais pas de l'orientation, la carte variographique se présente sous forme de cercles concentriques centrés sur l'origine.

La carte variographique met facilement en évidence des anisotropies différentes pour les structures aux différentes échelles. Les directions principales d'anisotropie, correspondant aux directions de plus forte et de plus faible variabilité, se lisent parfois directement sur la carte

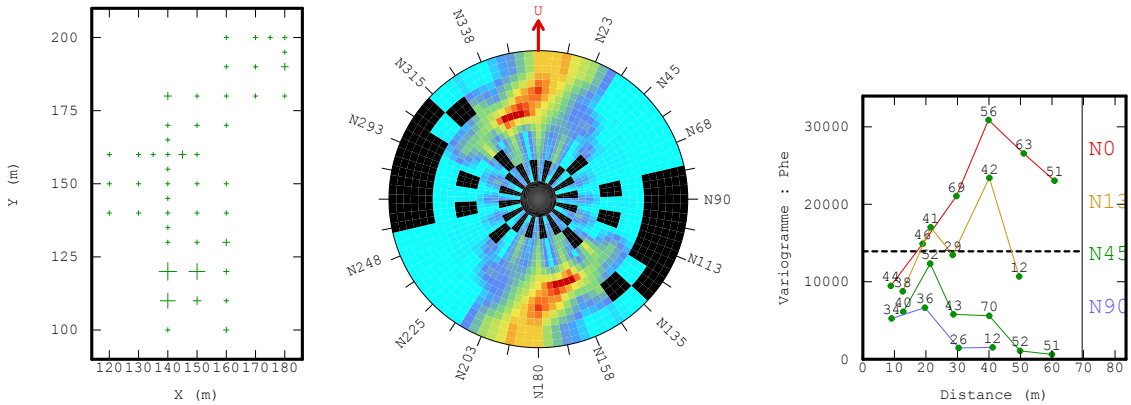


Figure I-6 Teneur en Phénanthrène. Implantation des prélèvements, carte variographique, variogrammes directionnels (nord-sud, est-ouest, et suivant les deux bissectrices).

Les lacunes de la carte variographique reflètent l'irrégularité de l'échantillonnage.

Le nombre de couples de points par pas du variogramme est reporté. La carte variographique et les variogrammes directionnels indiquent une anisotropie, avec une variabilité spatiale plus forte en Nord-Sud qu'en Est-Ouest.

Dans l'exemple précédent, les trois teneurs fortes situées au Sud du site sont-elles responsables de l'anisotropie ? Associées à des teneurs faibles ou moyennes, les teneurs fortes provoquent des écarts quadratiques élevés. A cause de la géométrie des points expérimentaux, ces valeurs fortes sont présentes pour les écarts calculés pour toutes les distances dans la direction Nord-Sud, et seulement pour quelques couples dans la direction est-ouest, ce qui peut suffire à expliquer la variabilité beaucoup plus faible dans cette direction. .

Si l'on supprime du calcul les huit points de mesures situés au sud du site, l'anisotropie disparaît (figure suivante). Dans ce cas, un modèle isotrope est préférable, et on calculera le variogramme moyen omnidirectionnel Figure I-7 et Figure I-8*).*

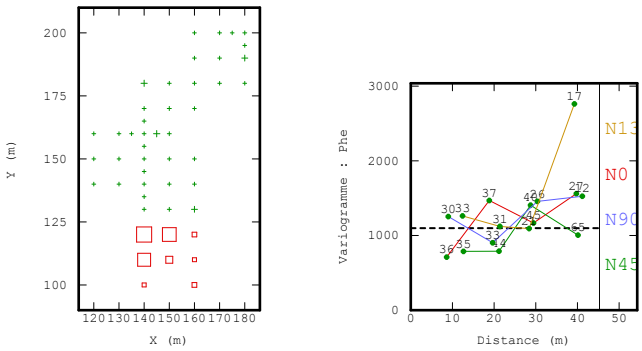


Figure I-7 Pour le Phénanthrène, la suppression des fortes teneurs (points masqués indiqués par des carrés rouges) du calcul des variogrammes fait disparaître l'anisotropie. Celle-ci résultait ici en effet de la géométrie des points expérimentaux.

Régularité

Le variogramme indique, au facteur ½ près, l'amplitude moyenne de l'écart quadratique. A distance nulle, et si l'on ne dispose que d'un seul échantillon par prélèvement, on a $\gamma(0)=0$. Comme un variogramme est toujours positif, sa croissance plus ou moins rapide aux petites distances traduit la détérioration de la corrélation entre teneurs lorsque la distance entre les échantillons augmente. Par ordre de régularité décroissante (Figure I-5) :

- pour une variable très régulière, par exemple la cote d'un substratum très continu, le variogramme présente un comportement légèrement parabolique à l'origine. Ce type de comportement ne s'observant pas pour des teneurs de substances polluantes dans les sols n'est pas présenté ici.
- Pour des variables régulières, comme l'épaisseur d'un niveau de remblais ou la teneur d'une substance soluble, le variogramme présente un comportement linéaire à l'origine (Figure I-5b).
- Le plus souvent, le variogramme présente une forte variation à l'origine, dite « effet de pépite » Figure I-5d), qui intègre toutes les échelles de variabilité inférieures à la maille de reconnaissance. Les erreurs de mesure, les incertitudes sur la localisation des échantillons, etc. se traduisent par une augmentation de l'effet de pépite additionnel sur le variogramme expérimental. A même maille de reconnaissance, plus cet effet de pépite est important, moins l'estimation est précise. Pour des teneurs en pollution de sols, il n'est pas rare que l'effet de pépite représente plus de 50% de la valeur du variogramme aux distances accessibles expérimentalement.
- Le cas limite est l'effet de pépite pur (Figure I-5c) : en l'absence de structure spatiale, c'est-à-dire dans le cas où les teneurs semblent réparties totalement au hasard sur le site, le variogramme est constant en-dehors de l'origine. Dans ce cas, aucune sélection des zones de teneurs fortes ou faibles n'est possible. Ce cas peut se rencontrer, lorsqu'un site a été profondément remanié ou remblayé avec des matériaux mélangés. On l'observe parfois localement, sur une partie de site.

Le comportement du variogramme a des conséquences importantes sur la précision de l'estimation, donc sur la faisabilité effective d'une dépollution.

L'effet proportionnel

Fréquemment, sur les sites un peu vastes, la variabilité spatiale reste modérée dans les zones de teneurs faibles et augmente localement avec la teneur. Calculés sur deux zones de teneurs respectivement fortes ou faibles, les variogrammes expérimentaux peuvent parfois être ajustés avec des modèles identiques à un facteur multiplicatif près. On parle alors d'effet proportionnel.

Loi spatiale

Le variogramme, qui correspond à la moyenne de la nuée variographique par classes de distance, ne caractérise pas complètement la variabilité spatiale. D'autres caractéristiques comme la loi de probabilité conjointe des teneurs en deux points simultanément (la loi spatiale bivariable), correspondent à une description plus précise du phénomène. Une modélisation plus poussée est nécessaire pour les calculs non linéaires, notamment les évaluations de dépassement de seuil.

Une caractérisation encore plus complète est donnée par les probabilités conjointes des teneurs en 3 points, 4 points, etc., qui constituent la « loi spatiale ». Cette caractérisation est nécessaire lorsqu'on veut construire des simulations conditionnelles (thème N°2).

2.3.8 Ajustement du variogramme

Un variogramme expérimental n'étant connu que pour certaines distances et n'ayant pas les propriétés mathématiques requises pour les calculs ultérieurs, il est nécessaire de l'ajuster par une fonction mathématique : le **modèle de variogramme**. Cette fonction fournit les valeurs du variogramme dans toutes les directions et pour toutes les distances. Cet ajustement conclut l'analyse variographique.

L'ajustement du ou des variogrammes expérimentaux retenus constitue la transition des données expérimentales vers le modèle géostatistique. Fixer les paramètres d'un modèle de variogramme

revient à poser des hypothèses sur la structure spatiale de la variable étudiée. Le modèle choisi aura une grande influence sur le résultat de l'estimation :

- stationnarité ou non stationnarité, suivant qu'à l'échelle retenue, le variogramme se stabilise ou non à un palier ;
- isotropie ou anisotropie ;
- valeurs des portées et des paliers.

Il est utile de préciser le domaine de validité du variogramme ajusté (distances, extension de la zone concernée sur le site).

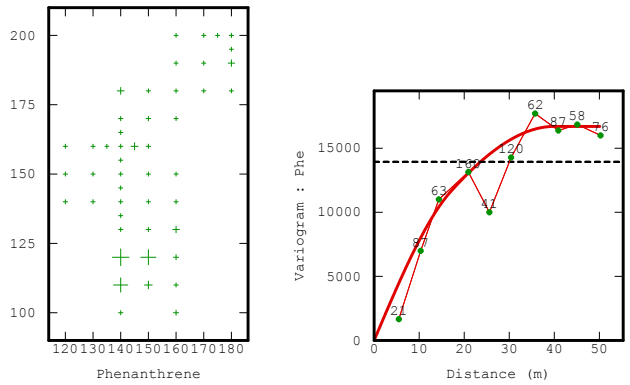


Figure I-8 Teneur en Phénanthrène. Implantation des 47 échantillons, variogramme expérimental omnidirectionnel et ajustement. Le modèle est la somme de deux structures « sphériques », de portées respectives 17m et 40m, et de paliers respectifs 4000 et 12700. Sur la carte d'implantation, les distances sont en mètres et les symboles sont proportionnels à la teneur.

Lorsque les points expérimentaux sont trop peu nombreux, il n'est pas toujours possible d'obtenir un variogramme expérimental « fiable » (voir thème N°4). On peut alors construire une enveloppe raisonnable de modèles à partir de ce variogramme expérimental, en posant différentes hypothèses :

- stationnarité ou absence de stationnarité : choix d'un variogramme à palier, ou comportant une composante non stationnaire, linéaire par exemple ;
- portées fortes ou faibles, relativement aux dimensions du site ou de la maille d'échantillonnage.

On calcule alors l'estimation avec ces différents variogrammes. La comparaison des résultats montre la nécessité, ou non, d'une meilleure connaissance du variogramme, compte tenu des objectifs de l'étude.

A quoi correspond un « trop faible nombre d'échantillons » ? On ne peux répondre à cette question de façon générale, la réponse dépendant des propriétés de la variable étudiée, propriétés que l'on cherche justement à déterminer (voir thème N° 4). Une approche multivariable (utilisation d'une variable auxiliaire par exemple) peut également permettre de tirer parti d'un faible nombre de points de données (voir thème N° 3).

*Est-il possible de valider l'ajustement d'un variogramme ? La **validation croisée**, d'usage assez répandu, consiste à extraire un à un chaque point expérimental du jeu de données et à réestimer la teneur en ces points à partir des données restantes. La comparaison de cette estimation à la teneur mesurée fournit une erreur d'estimation expérimentale. Des critères globaux comme :*

- *la moyenne du carré de ces erreurs (variance d'estimation expérimentale),*
- *la moyenne du carré des erreurs divisées par la variance d'estimation déduite du variogramme (comparaison de la variance d'estimation expérimentale à la variance d'estimation théorique), afin de tenir compte des irrégularités de l'échantillonnage,*

permettent de confronter globalement différents modèles. Il est préférable de toujours regarder les résultats en détail, pour vérifier si les différences observées sont bien représentatives de l'ensemble de la zone étudiée et ne proviennent pas seulement d'un ou de quelques points particuliers (ce qu'il faudra chercher à expliquer). L'interprétation de la validation croisée demande un peu d'expérience.

2.3.9 Irrégularité de l'échantillonnage

Un échantillonnage très irrégulier, en particulier un échantillonnage préférentiel, peut être problématique. Lorsque l'échantillonnage a été prélevé de façon préférentielle, cela crée des biais que l'on doit corriger. Il est donc indispensable de le savoir, d'où la nécessité de rassembler un maximum d'informations sur le processus d'échantillonnage : les prélèvements ont-ils été effectués « n'importe où » c'est-à-dire « au hasard » ou bien d'après des observations locales non consignées, ou encore implantés en fonction d'anciennes infrastructures : bordure de voie, limite d'aire de stockage ou de zone d'épandage, etc.

Il est parfois possible de détecter un échantillonnage préférentiel

- *lorsque la majorité des fortes valeurs, repérables sur l'histogramme, se situe dans la zone échantillonnée de façon plus dense.*
- *en comparant les moyennes locales des teneurs dans la zone échantillonnée de façon dense et ailleurs sur le site ;*
- *en comparant les statistiques des teneurs, calculées « directement » puis en pondérant chaque point par la densité locale de l'information*

Un échantillonnage très irrégulier complique l'analyse exploratoire. Les statistiques élémentaires ne sont plus nécessairement représentatives des caractéristiques de la pollution sur la zone. Il est souvent nécessaire de **pondérer** les valeurs, en attribuant un poids plus faible aux échantillons d'une zone densément informée qu'aux échantillons isolés.

Il ne faut pas confondre un échantillonnage irrégulier et un resserrement local de la maille. Dans ce dernier cas, qui facilite l'analyse géostatistique, statistiques globales et histogrammes sont calculées à partir de la maille régulière. Le resserrement local de la maille est uniquement utilisé pour préciser le comportement du variogramme aux petites distances.

3 CARTOGRAPHIE

Nous considérons désormais le cas usuel d'un polluant dont la teneur est connue en un nombre limité de points. Nous avons réalisé une analyse statistique de ces données, y compris en ce qui concerne leur variabilité spatiale. Nous nous proposons maintenant d'en faire la cartographie.

3.1 Introduction au krigeage comme méthode d'estimation

Cartographier une variable, ici la teneur en un élément polluant, permet de synthétiser sous une forme quantitative et visuelle la connaissance que nous en avons.

Un premier niveau de représentation, présenté dans la section précédente, consiste à reporter sur un plan chaque mesure par un symbole de taille proportionnelle à la teneur ou encore à positionner chaque point de mesure en y indiquant la teneur mesurée. Un tel mode de représentation peut mettre en valeur certaines données mais ne comporte aucune part d'interprétation. En contrepartie il est peu lisible dès que les données sont un peu nombreuses. Et bien entendu il ne s'aventure pas à « prédire » la teneur en un point non mesuré.

La cartographie va au-delà de ce mode de représentation. Elle se traduit par des courbes isovaleurs, des cartes pixellisées où la couleur de chaque pixel correspond à une classe de valeur de la teneur, une grille de blocs colorés en fonction de leur teneur moyenne, etc. Elle contient une part d'interprétation pour interpoler entre les données ou extrapoler au-delà.

Deux voies sont possibles pour spécifier une méthode de cartographie. La première consiste à définir ses propriétés interpolatoires ; c'est le cas des splines par exemple où on demande à la carte d'être lisse (par exemple continue et dérivable partout). La seconde consiste à définir la façon d'estimer la variable en un point où elle n'a pas été mesurée ; c'est le cas par exemple de la méthode de pondération des données par l'inverse de la distance. C'est le cas aussi du krigeage employé en géostatistique. En fait, les deux points de vue se rejoignent, du moins en théorie : tout problème spline peut s'exprimer comme un problème de krigeage et inversement.

3.1.1 Estimateurs linéaires

Pour estimer la teneur Z en un point x_0 à partir des observations $z(x_i)$ en n points de mesure x_i, les méthodes classiques (pondération par l'inverse de la distance, moyenne mobile...) emploient un estimateur linéaire, c'est-à-dire de la forme

$$Z^*(x_0) = \lambda_1 Z(x_1) + \lambda_2 Z(x_2) + \ldots + \lambda_n Z(x_n)$$

où les pondérateurs λ_i sont de somme unité :

[1] $\lambda_1 + \lambda_2 + \ldots + \lambda_n = 1$

Dans le cas de la pondération par l'inverse de la distance par exemple, λ_i est égal à $A / d(x_0, x_i)$, où $d(x_0, x_i)$ représente la distance entre le point à estimer x_0 et le point de mesure x_i, et où A est le facteur de normalisation assurant que la somme des pondérateurs est égale à 1.

Bien évidemment un grand nombre d'estimateurs linéaires peuvent être définis. Comment choisir ? En général on souhaite que la carte passe bien par les données, ce qu'assurent nombre de méthodes conventionnelles, comme par exemple la pondération par l'inverse de la distance, ou géostatistiques avec le krigeage. Dans les deux cas, l'analyse exploratoire des données peut aider dans le choix de la méthode et de ses paramètres. La différence entre les méthodes conventionnelles et géostatistiques se trouve dans la façon d'interpoler les teneurs entre les données ou de les extrapoler au-delà. Interpolation et extrapolation signifient nécessairement continuité ou corrélation spatiale. Elle est implicite avec les méthodes conventionnelles. Elle est explicite à travers le variogramme et donc le krigeage qui utilise ce dernier (voir § 3.1.2).

Il est à noter que lorsque les données sont bruitées sous l'effet d'erreurs de mesure ou autres, on ne souhaite plus forcément que les cartes soient parfaitement ajustées aux données. L'introduction d'un effet de pépite avec le krigeage permet de filtrer ces erreurs et de produire une carte moyenne adaptée au mieux aux données. Dans le cas d'une pondération par l'inverse de la distance, l'introduction d'une composante non nulle a pour obtenir un pondérateur du type $A / [a + d(x_0, x_i)]$ filtrera également un terme d'erreur, mais qu'il est plus difficile de quantifier.

3.1.2 Krigeage

Au lieu de chercher le meilleur estimateur parmi une liste limitée, ne pourrait-on en définir un qui s'adapterait au cas par cas et puisse être alors considéré comme le meilleur (en un sens à préciser) parmi tous les estimateurs linéaires possibles ?

Deux estimateurs peuvent être comparés expérimentalement de la façon suivante : on calcule $Z^*(x_0)$ au point $x_0$pour chacun des estimateurs, puis on réalise la mesure de $Z(x_0)$. On compare alors les erreurs d'estimation $Z^*(x_0) - Z(x_0)$ auxquelles conduisent ces deux estimateurs. Il est difficile de conclure sur la base d'une seule expérience. Mais si on la répète un nombre suffisant de fois pour des configurations similaires de x_0, x_1, x_2, …, x_n, on pourra mesurer lequel des deux estimateurs est le plus efficace (pour ce type de configuration). On pourra retenir par exemple celui qui conduit <u>le plus souvent</u> à l'erreur la plus faible (en valeur absolue), ou celui pour lequel la valeur absolue de l'erreur est <u>en moyenne</u> la plus faible. En géostatistique on retient celui pour lequel le carré de l'erreur est en moyenne le plus faible. Pourquoi ? Parce que, une fois le variogramme modélisé, et à condition de supposer que la variable étudiée se comporte partout de la même manière (hypothèse de stationnarité), on est capable de calculer théoriquement la moyenne du carré de l'erreur correspondant à une infinité de situations analogues, sans avoir besoin de réaliser de nouvelles mesures. Si on adopte ce critère, le krigeage est l'estimateur optimum parmi tous les estimateurs linéaires : en moyenne il conduit à une erreur quadratique plus faible que n'importe quel autre estimateur linéaire.

Rechercher ce meilleur estimateur revient à résoudre un système linéaire d'équations, impliquant le variogramme et dont les solutions, les poids de krigeage, sont les pondérateurs de la combinaison linéaire. Ce système de $n + 1$ équations à $n + 1$ inconnues (les n pondérateurs, plus un paramètre de Lagrange qui s'introduit à cause de la condition [1]) a la forme suivante :

$$\begin{aligned}
&\lambda_1 \gamma(0) + \lambda_2 \gamma(x_2 - x_1) + \ldots + \lambda_n \gamma(x_n - x_1) + \mu = \gamma(x_0 - x_1) \\
&\lambda_1 \gamma(x_1 - x_2) + \lambda_2 \gamma(0) + \ldots + \lambda_n \gamma(x_n - x_2) + \mu = \gamma(x_0 - x_2) \\
&\qquad\qquad\qquad\qquad \vdots
\end{aligned}$$

[2]

$$\lambda_1\,\gamma(x_1 - x_n) + \lambda_2\,\gamma(x_2 - x_n) + \ldots + \lambda_n\,\gamma(0) \qquad + \mu = \gamma(x_0 - x_n)$$

$$\lambda_1 \qquad + \lambda_2 \qquad + \ldots + \lambda_n \qquad = 1$$

La partie gauche du système de krigeage fait intervenir les termes $\gamma(x_j - x_i)$. La partie droite contient les termes $\gamma(x_0 - x_i)$ qui représentent la corrélation spatiale entre les données et le point d'estimation x_0. Le krigeage dépend ainsi de la configuration des données, de la localisation du point x_0 à estimer par rapport à celles-ci et de la variabilité spatiale synthétisée dans le variogramme. Une redondance de l'information se traduit par des pondérateurs plus faibles affectés aux données proches les unes des autres et donc fortement corrélées.

En sous-produit, le krigeage donne la moyenne quadratique de l'erreur ou variance de krigeage σ_K^2, dont la racine carrée est l'écart-type de krigeage σ_K. Sous certaines conditions, cet écart-type peut être utilisé pour définir des intervalles de confiance (voir plus loin).

Cas d'un effet de pépite pur

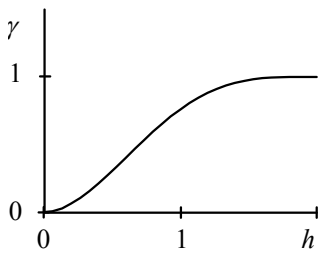
En cas d'effet de pépite pur C_0, les termes $\gamma(x_i - x_j)$ pour $i \neq j$ sont tous égaux à C_0. Il est aisé de voir que les pondérateurs de krigeage d'un point non mesuré sont alors tous égaux à $1 / n$, où n est le nombre de données. En l'absence de structure spatiale, on ne peut faire mieux qu'estimer $Z(x_0)$ par la moyenne des données. La variance de krigeage vaut alors $C_0\,(1 + 1/n)$. On se trouve ramené aux résultats des statistiques classiques. Si on cartographie Z^* ou σ_K, on obtient une carte de couleur uniforme.

3.1.3 Propriétés interpolatoires du krigeage

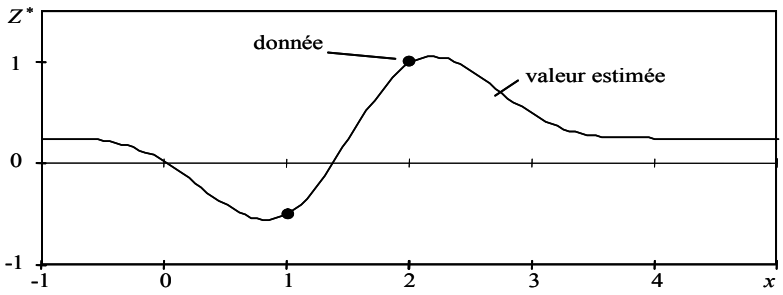
Le krigeage est un interpolateur exact, c'est-à-dire que si le point à estimer fait partie du jeu de données l'estimation en ce point est égale à la valeur mesurée.

Il est à noter que l'objectif du krigeage n'est pas de donner une carte lisse, ou lisible, mais d'estimer au mieux en tout point x_0. Une carte krigée reflète la connaissance que l'on a de la variable étudiée : elle est plutôt lisse dans les zones mal renseignées et présente des détails dans les zones bien échantillonnées : elle gère au mieux l'information disponible **mais ne crée pas d'information**.

Examinons le krigeage en tant que fonction de x_0. Le premier membre du système [2] ne dépendant pas de x_0, sa solution s'exprime comme une combinaison linéaire du second membre, donc des valeurs $\gamma(x_0 - x_i)$ fournies par le variogramme. Il en résulte qu'au voisinage des données la carte krigée présentera la même allure que le variogramme au voisinage de l'origine (voir Figure I-9) ; si le variogramme est parabolique, il en sera de même du krigeage autour d'un point de mesure. S'il est linéaire (pour $|h| > 0$), le krigeage présentera des extrema de même style (c'est-à dire des trous ou pics aigus) autour des données ; s'il présente un effet de pépite, le krigeage présentera une discontinuité à chaque point de donnée : par définition, le krigeage d'un point expérimental redonne la valeur mesurée, mais le krigeage d'un point voisin, aussi proche soit-il, marque une discontinuité.



(a)



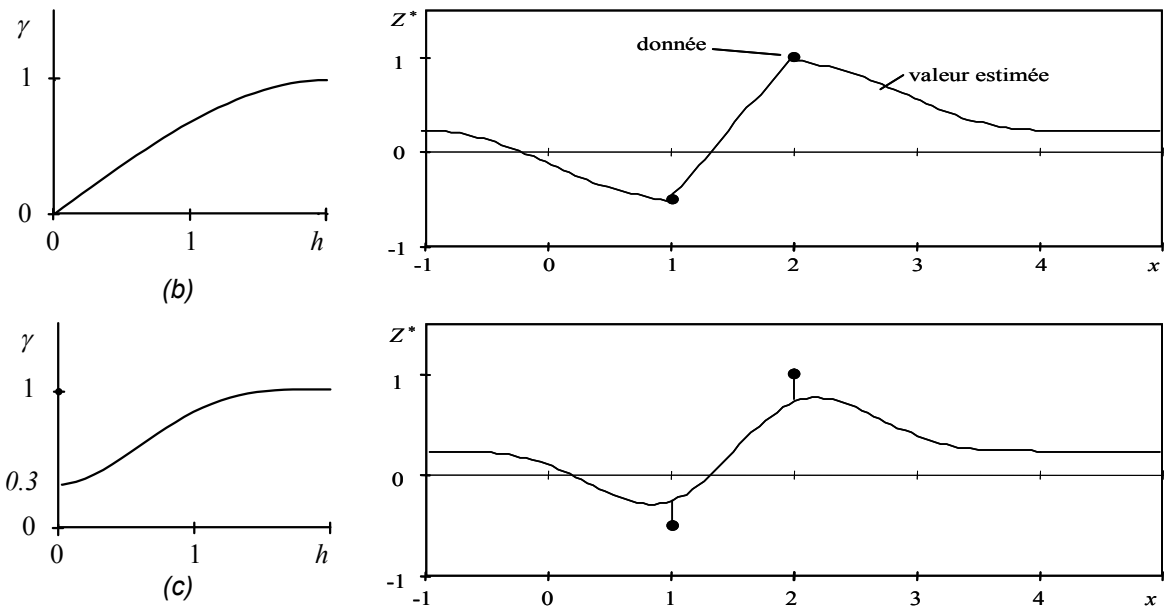


Figure I-9 Influence du comportement à l'origine : valeurs estimées par krigeage à partir de deux points de données. (a) comportement parabolique à l'origine ; (b) comportement linéaire ; (c) effet de pépite partiel. D'après Chilès et Delfiner (1999).

Krigeage en présence d'un effet de pépite

Comme nous venons de le voir (§3.1.2), en présence d'un effet de pépite, le krigeage d'un point de mesure redonne la valeur mesurée, mais le krigeage d'un point voisin, aussi proche soit-il de la donnée, marque une discontinuité. Dans les applications pratiques, lorsqu'on interpole Z en un point x_0, par exemple un nœud de grille, on ne s'intéresse pas tant à l'estimation au point x_0 exact, mais en un point qui coïncide à peu près avec x_0. Autrement dit, si par hasard x_0 coïncide avec un point expérimental, on souhaite obtenir une solution en continuité avec les points proches plutôt que retrouver la mesure expérimentale. Si par exemple l'effet de pépite correspond à une erreur de mesure, il existe des techniques de filtrage permettant de retrouver la composante continue du variogramme (sans le terme d'erreur).

En toute rigueur il s'agit d'un problème multivariable (on a deux variables : le phénomène lui-même et le terme d'erreur), dont la solution est un cokrigeage (voir thème N° 3), mais ce cas particulier est tellement trivial qu'il constitue un cas particulier du krigeage.

3.2 Exploitation de la variance de krigeage

3.2.1 Cas gaussien

Un cas particulier mérite d'être mentionné, celui d'une variable à loi spatiale gaussienne et de moyenne connue. Par "variable à loi spatiale gaussienne" il faut entendre non seulement que l'histogramme de cette variable est gaussien, mais encore qu'il en est de même de l'histogramme de toute variable obtenue par combinaison linéaire ; ainsi par exemple la valeur moyenne dans un bloc v est aussi gaussienne. Dans ce cas l'espérance conditionnelle est tout simplement le krigeage et la variance conditionnelle est la variance de krigeage (elle ne dépend donc pas des valeurs observées z(x1), …, z(xn)). Il convient toutefois de préciser que le krigeage n'est pas l'estimateur présenté plus haut, mais le krigeage simple, où la condition (1) n'est pas imposée. En effet cette condition a pour but d'estimer la moyenne sans erreur lorsqu'elle est inconnue. Si la moyenne m est connue, on prend un estimateur de la forme

$$Z^* = m + \lambda_1 (Z(x_1) - m) + \lambda_2 (Z(x_2) - m) + \ldots + \lambda_n (Z(x_n) - m)$$

Le système de krigeage simple est similaire au système de krigeage ordinaire, sauf qu'il ne comporte comme inconnues que les n pondérateurs λ_i.

Le krigeage simple jouit donc de bonnes propriétés dans le cas gaussien et à moyenne connue. En particulier la variance de krigeage simple σ^2_{KS} est une variance conditionnelle, et on peut en déduire un intervalle de confiance car l'erreur de krigeage est gaussienne. Par exemple l'intervalle de confiance à 95% est

$Z^* - 2\,\sigma_{KS} \leq Z(x_0) \leq Z^* + 2\,\sigma_{KS}$

Dans le cas du krigeage ordinaire et d'une variable non gaussienne, on dit parfois que l'intervalle de confiance conventionnel à 95% est

$Z^* - 2\,\sigma_{KO} \leq Z(x_0) \leq Z^* + 2\,\sigma_{KO}$

Cet intervalle de confiance conventionnel peut être considéré comme une bonne approximation de l'intervalle de confiance réel si l'histogramme de Z est proche d'un histogramme gaussien (histogramme symétrique sans longue queue de distribution).

3.2.2 Signification de la variance de krigeage

La variance de krigeage doit être comprise comme la moyenne des erreurs quadratiques Elle dépend uniquement de la position relative des données (x_1, …, x_n) par rapport au point d'estimation x_0. Elle n'est pas conditionnelle aux valeurs observées $z(x_1)$, …, $z(x_n)$. Supposons ainsi que sur un site on trouve les deux situations de la Figure I-10. Le krigeage leur associera la même estimation, ce qui paraît normal, mais aussi le même écart-type de krigeage, alors que la situation de gauche paraît plus sûre que celle de droite. A géométrie égale, pour un même variogramme, la variance de krigeage est ainsi la même, que les valeurs des données soient faibles, fortes, toutes égales ou très différentes.

C'est une limitation du krigeage, qui pour être levée demanderait de recourir à l'espérance conditionnelle. Cette dernière n'est généralement pas aisée à calculer et suppose une information considérablement plus riche que le variogramme sur la variabilité spatiale du phénomène étudié, inaccessible en pratique.

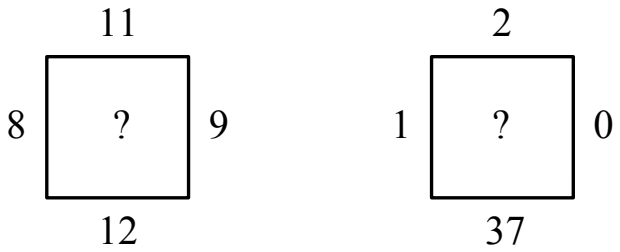


Figure I-10 Deux scénarios avec la même valeur estimée et la même variance de krigeage (lorsqu'on utilise le même modèle de variogramme).

Dans l'exemple précédent, on peut obtenir des écarts-types de krigeage plus proches de ce que suggère l'intuition en subdivisant la zone d'étude en sous-zones de variabilité spatiale plus homogène. Ceci suppose toutefois que ces sous-zones soient aisées à définir et que chacune contienne un nombre suffisant de données pour qu'on puisse y effectuer l'analyse variographique.

3.2.3 Adaptation de la maille de reconnaissance

La variance de krigeage mesure la précision de l'estimation. Or on montre qu'elle ne dépend que du variogramme, de la configuration des données et du point à estimer, mais pas des valeurs en ces points. A l'issue d'une première campagne d'échantillonnage ayant permis de calculer un variogramme, on peut alors calculer la variance d'estimation qui serait obtenue pour différentes mailles de reconnaissance. Ceci permet sous certaines conditions d'adapter la maille en fonction de la précision souhaitée.

On peut aboutir à une maille irréaliste car beaucoup trop dense, en particulier en cas de forte composante pépitique du variogramme, ou en présence d'une structure à très courte portée et fort palier. Cela peut signifier qu'une sélection locale fine des zones polluées risque d'être illusoire en pratique. Dans ce cas, une réflexion s'impose sur la faisabilité d'une dépollution « sélective ».

3.3 Cas particuliers

Ce chapitre concerne des remarques de mise en œuvre de la géostatistique dans certains cas particuliers et peut être sauté en première lecture.

3.3.1 Krigeage d'une variable lognormale

Si Z a une distribution lognormale, $Y = log(Z)$ est gaussienne. Il est souvent préférable de travailler sur Y que sur Z, car une variable lognormale a de mauvaises propriétés statistiques si la variance relative est grande : son histogramme comporte alors une longue queue de distribution, et la présence ou l'absence d'une forte valeur peut marquer fortement le calcul de la moyenne et de la variance, ou le krigeage d'un point. En travaillant sur Y on obtient un estimateur Y^* de $Y(x_0)$ et un écart-type de krigeage σ_K. Y^* est aussi un estimateur médian : la loi de Gauss étant symétrique, on a une chance sur 2 que la vraie valeur Y soit plus petite que Y^*, et une chance sur 2 qu'elle soit plus grande que Y^*. Et on a un intervalle de confiance à 95% sur Y :

$$Y^* - 2\,\sigma_K \leq Y(x_0) \leq Y^* + 2\,\sigma_K$$

Comme $Z = exp(Y)$, on peut en déduire que $Z^* = exp(Y^*)$ est l'estimateur médian de $Z(x_0)$, et que l'intervalle de confiance à 95% est

$$exp(Y^*) / S_K^2 \leq Z(x_0) \leq exp(Y^*)\, S_K^2 \quad \text{avec} \quad S_K = exp(\sigma_K)$$

S_K joue le rôle d'un écart-type multiplicatif. Il est possible également de construire à partir de Y^* un estimateur au sens habituel (c'est-à-dire représentant la moyenne et non la médiane des situations possibles) mais cet estimateur est moins robuste et demande des contrôles appropriés. Il est nécessaire lorsqu'on veut estimer une quantité de polluant mais n'est d'aucune utilité si on s'intéresse plutôt à une probabilité de dépassement de seuil.

3.3.2 Estimation d'un bloc ou d'un domaine

Bien souvent on s'intéresse plutôt à la valeur moyenne de Z dans un bloc v ou dans un domaine V qu'à la valeur de Z en un point x_0 précis. Le krigeage se généralise très simplement à l'estimation d'un volume. En effet, le krigeage étant un estimateur linéaire, il suffit de remplacer les termes faisant intervenir x_0 par leur moyenne lorsque x_0 décrit le bloc v. Le premier membre du système de krigeage (2), où x_0 n'intervient pas, reste donc inchangé, et les termes du second membre, de la forme $\gamma(x_0 - x_i)$, sont à remplacer par les moyennes de $\gamma(x - x_i)$ lorsque x décrit le volume v.

Il en résulte d'ailleurs que le krigeage de v est la moyenne des krigeages des points x_0 décrivant v. Par contre la variance de krigeage n'est pas la moyenne des variances de krigeage des points décrivant v. Elle est en général plus faible, car on se trompe moins sur une valeur moyenne que sur une valeur ponctuelle.

3.3.3 Estimation de la valeur moyenne sur l'épaisseur d'une formation

Les mesures de pollution sont souvent relatives à un prélèvement sur une certaine épaisseur. Si cette profondeur est la même partout (par exemple 1 m), on se ramène à une étude 2D où $Z(x)$ représente la teneur en polluant au point géographique (2D) x sur une épaisseur de 1 m. Si en chaque point de mesure on a réalisé en fait plusieurs prélèvements dans des tranches de profondeur successives, on doit en toute rigueur réaliser une étude à trois dimensions. Souvent on se contente de travailler sur l'accumulation de pollution au droit de tout point x de l'espace géographique 2D, ce qui permet de se ramener à une étude 2D. On peut alors considérer trois variables :

- l'épaisseur de la formation, $E(x)$;
- l'accumulation $A(x)$, c'est-à-dire la somme des produits teneur fois hauteur des échantillons successifs prélevés au point x, ou tout simplement le produit teneur moyenne fois épaisseur de la formation au point x ;
- la teneur moyenne $T(x) = A(x) / E(x)$.

Les méthodes linéaires s'appliquent à des variables additives : si v_1 et v_2 sont deux volumes disjoints identiques, si $Z(x)$ a pour moyenne z_1 dans v_1 et z_2 dans v_2, la moyenne de Z dans la réunion de v_1 et v_2 est $(z_1 + z_2) / 2$. Cela ne va pas toujours de soi. Une variable qualitative comme la blancheur ou la transparence n'est pas additive. Une variable quantitative comme la perméabilité ne l'est pas non plus (sauf cas particulier). Une teneur est additive, mais des teneurs moyennes mesurées sur des épaisseurs différentes ne sont pas additives : il faut les pondérer par les épaisseurs correspondantes. C'est ce que fait une accumulation. La méthodologie à suivre est alors la suivante :

- kriger l'épaisseur ;
- kriger l'accumulation ;
- prendre pour estimation de la teneur le rapport A^* / E^*.

Pour éviter l'apparition d'artéfacts, il est recommandé d'employer les mêmes points de donnée et le même variogramme (à une constante multiplicative près) pour le krigeage de l'épaisseur et pour le krigeage de l'accumulation.

On obtiendra les variances de krigeage de l'épaisseur et de l'accumulation. On ne peut en déduire simplement l'écart-type d'estimation de la teneur moyenne.

Il peut y avoir des exceptions à cette méthodologie : si le nuage de corrélation entre l'épaisseur et la teneur moyenne montre que la variabilité de la teneur moyenne ne diminue pas quand l'épaisseur augmente mais au contraire reste du même ordre de grandeur, il est préférable de kriger séparément l'épaisseur et la teneur moyenne et d'en déduire l'accumulation si on s'y intéresse.

3.3.4 Pondérateurs négatifs

Il est à noter que si le système de krigeage ordinaire impose que la somme des pondérateurs soit égale à 1, il n'impose pas que les pondérateurs soient positifs. À une dimension, lorsqu'on représente une variable mesurée en deux points par la droite passant par ces deux points, c'est précisément le fait d'introduire un pondérateur négatif et un pondérateur supérieur à 1, qui permet d'aller en-deçà des données d'un côté, et au-delà de l'autre côté. De même le krigeage fournit souvent quelques pondérateurs négatifs. Lorsqu'on estime une variable positive (teneur ou épaisseur par exemple), on n'est donc pas à l'abri de résultats de krigeage négatifs. Une exception notable est celle d'un effet de pépite pur, où les n données reçoivent toutes le pondérateur $1 / n$.

Lorsque des variables positives (teneur, épaisseur) sont estimées négativement, on commence toujours par vérifier la variance de krigeage associée. Assez fréquemment, celle-ci se révèle plutôt élevée ; l'estimation négative est alors un indice d'une mauvaise précision de l'estimation, qui n'est pas faisable de façon réaliste à partir de l'information disponible, trop lacunaire.

Parfois cependant, on obtient localement des valeurs négatives, sans que cela corresponde à des lacunes de l'échantillonnage. On se contente généralement de les mettre à zéro. Pour faire mieux, il faut réaliser un krigeage sous la contrainte que le résultat soit positif. Il existe pour cela une solution rigoureuse dans le cas gaussien à moyenne connue ("échantillonneur de Gibbs"), qui reste assez robuste lorsqu'on s'éloigne de ces conditions. Comme elle est assez lourde (la solution s'obtient par une méthode itérative), en pratique elle n'est pas mise en œuvre dans ces circonstances.

3.3.5 Krigeage d'une indicatrice

On peut aussi kriger toute autre variable additive qu'une teneur, et notamment la fonction indicatrice d'une lithologie (type de sols). Si $I_a(x)$ est la fonction indicatrice associée à la lithologie a, cette fonction vaut 1 si au point de mesure x la lithologie observée est a, elle vaut 0 sinon. La valeur krigée est alors interprétée comme la probabilité que le point krigé x_0 appartienne à cette lithologie (krigeage ponctuel), ou comme la proportion de points du domaine krigé v qui appartiennent à cette lithologie (krigeage d'un volume v). Il ne faut pas oublier de tempérer cette interprétation par l'incertitude donnée par la variance de krigeage, qui peut être importante. Il ne faut pas non plus s'étonner de trouver des estimations négatives ou supérieures à 1 : les pondérateurs de krigeage pouvant être négatifs mais de somme unité, lorsque les données sont toutes égales à 0 ou 1, les estimateurs peuvent sortir de l'intervalle [0, 1]. Pour l'éviter il faut recourir à un krigeage sous contrainte, pour lequel il n'existe pas de solution méthodologique indiscutable.

Lorsqu'on s'intéresse au dépassement d'un seuil de teneur s, on peut se ramener au cas précédent en définissant l'indicatrice correspondante $I_s(x) = 1$ si $Z(x)>s$ et 0 sinon et en krigeant cette indicatrice. On perd cependant beaucoup d'information. En effet, de la donnée $Z(x_i)$ on ne retient plus que sa position par rapport à s ; ainsi, si elle est supérieure à s, on traitera de la même manière une donnée peu supérieure au seuil et une donnée à très forte teneur. Pour valoriser pleinement l'information, il faut recourir aux méthodes de géostatistique non linéaire, qui ne sont pas traitées ici.

3.3.6 Erreurs de mesure d'amplitude variable

Lorsque les données proviennent de campagnes différentes, elles sont souvent associées à des erreurs de mesure de niveaux différents. Pour en tenir compte, dans le premier membre du système (2), les termes $\gamma(x_i - x_j)$ pour $i \neq j$ doivent comporter la composante continue du variogramme, plus $(\sigma_i^2 + \sigma_j^2) / 2$. Si les erreurs de mesure sont toutes du même niveau, caractérisé par l'effet de pépite C_0, on retrouve bien l'introduction du terme C_0 sur toute la matrice de krigeage sauf sa diagonale.

3.3.7 Krigeage en voisinage unique ou krigeage local

Tel qu'il a été développé ci-dessus, le krigeage d'un point x_0 ou d'un volume v est effectué à partir de l'ensemble des données. On procède ainsi lorsque le nombre de données n'est pas trop élevé. Dans le cas contraire, on effectue l'estimation à partir d'un sous-ensemble des données : les points situés au voisinage de x_0 ou de v. Deux raisons poussent à procéder ainsi :

- *Si la résolution du système [2] est encore réalisable à partir d'un millier de données, le temps de calcul requis augmente toutefois très rapidement avec n.*
- *L'hypothèse de stationnarité à la base de l'estimateur de krigeage utilisé, est souvent raisonnable à l'échelle locale mais discutable à l'échelle globale (par exemple, la moyenne varie lentement mais sûrement d'une zone à l'autre).*

La sélection des données voisines de x_0 ou de v est un problème délicat. Il est recommandé de retenir des données entourant bien le point ou le bloc krigé. On procède habituellement à une sélection par octants : l'espace est découpé en huit secteurs angulaires centrés sur x_0 ou v, de 45° chacun ; les données sont réparties dans ces huit octants et y sont classées par distance croissante au centre. On retient alors dans chaque octant les p points les plus proches. Des points complémentaires sont retenus lorsque certains octants comportent moins de p données. Dans le cas du krigeage d'un bloc v, on s'assure de plus que tous les points intérieurs à v sont sélectionnés.

Il est à noter que le krigeage en voisinage glissant altère les propriétés interpolatoires du krigeage : l'estimateur peut présenter des discontinuités parasites dues aux changements de voisinage. Pour en limiter l'effet, il est recommandé d'utiliser des voisinages suffisamment grands (par exemple 16 points en tout, soit 2 par octant).

3.4 Intérêt et limites de l'estimation

Intérêt :

- Représentation du phénomène et cartographie, permettant par exemple d'identifier les zones de faibles ou fortes teneurs, et donc les zones à dépolluer.
- Evaluation de l'incertitude locale sur les teneurs par rapport à des seuils de dépassement (cartes de probabilité).
- Lorsqu'on a des doutes sur le modèle de la teneur, la comparaison des cartes obtenues par krigeage avec l'un ou l'autre modèle permet de mettre en évidence certaines des conséquences du choix de l'un ou l'autre modèle.

Limitations :

- Les cartes de valeurs estimées ne sont généralement pas une représentation réaliste des teneurs dans les sols. Les erreurs d'estimation sont invisibles à ce stade, mais toujours présentes. La variance de krigeage ne fournit qu'un résultat statistique sur ces erreurs. De même, la variabilité locale de la teneur vraie n'est pas celle observée sur le résultat de l'estimation (effets de lissage).
- L'estimation ne permet pas de quantifier les volumes à dépolluer (voir thème N°2).

4 ILLUSTRATION

La Figure I-11 présente l'implantation des prélèvements « superficiels » d'une campagne de mesures sur un site pollué par des hydrocarbures, dont l'analyse fournit la teneur en « hydrocarbures totaux », exprimée en mg/kg ms. Le site ayant été nivelé, tous les échantillons, prélevés à peu près à la même profondeur, sont supposés situés à la même cote. L'étude est donc effectuée dans le plan, à deux dimensions. La variable étudiée est la teneur, le support étant supposé approximativement identique pour tous les échantillons.

4.1 Histogrammes et statistiques élémentaires

L'échantillonnage plus dense au nord-est s'explique dans ce cas par les connaissances sur l'historique du site, ce que reflètent les statistiques : la teneur est plus élevée sur cette zone, comme le confirment les statistiques élémentaires (Tableau I-1). Sur l'ensemble du site, on définit une maille à peu près régulière en retenant au nord-est quelques échantillons pour obtenir approximativement la même densité d'échantillonnage que sur le reste du site.

La moyenne sur la zone nord-est est plus de deux fois supérieure à celle sur l'ensemble du site (estimée par la moyenne pour la maille « régulière »), et la médiane pour le nord-est plus de quatre fois supérieure à celle sur tout le site. Le rapport des écart-types est légèrement inférieur à celui des moyennes. La dispersion relative autour de la moyenne, traduite par le coefficient de variation, est voisine pour les deux ensembles.

Le resserrement local de la maille correspond dans ce cas à un échantillonnage préférentiel.

Tableau I-1 Statistiques élémentaires de la teneur en hydrocarbures totaux de sols superficiels. Le coefficient de variation est le rapport de l'écart-type à la moyenne.

Maille	Nombre échantillons	Minimum (mg/kg ms)	Maximum (mg/kg ms)	Médiane (mg/kg ms)	Moyenne (mg/kg ms)	Ecart-type (mg/kg ms)	Coeff. de variation
Toutes données	83	17	8180	243	653	1061	1,62
« régulière »	44	19	3134	119	434	658	1,52
Nord-est	45	17	8180	520	968	1322	1,37

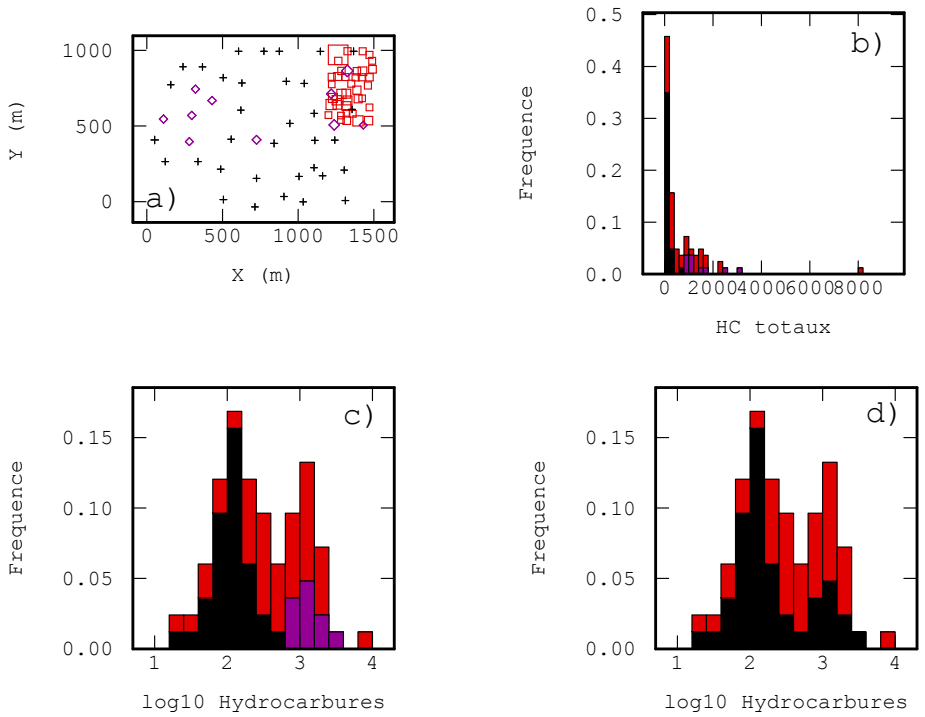


Figure I-11 Carte d'implantation des échantillons (a), histogramme des teneurs (b), et de leur logarithme décimal (c et d). En rouge, les échantillons « supplémentaires » de la maille dense au Nord-est. En mauve, données de la grille « régulière » appartenant au mode des teneurs fortes.

Visualisés sur l'histogramme des teneurs ou de leur logarithme (Figure I-11), les prélèvements de la zone nord-est se retrouvent principalement dans les classes de teneurs fortes. Le suréchantillonnage local n'est cependant pas la cause de la bimodalité des teneurs, bien visible sur l'histogramme des logarithmes. Le premier mode des teneurs se situe autour de 135 mg/kg ms, le second correspond à des teneurs environ dix fois plus élevées (autour de 1450 mg/kg ms). Le mode des teneurs élevées correspond aussi à des données de la grille régulière situées à l'Ouest du site. Comme bien souvent, l'historique ne permet de retrouver qu'une partie des zones de forte pollution.

4.2 Variogrammes expérimentaux

Bien que les zones de teneurs fortes apparaissent toutes situées dans la moitié nord du site, on trouve également dans cette partie des teneurs du mode faible. Comme il n'apparaît pas de limite nette entre zones de teneurs fortes ou faibles, et compte tenu du nombre réduit d'échantillons de la maille « régulière », le site est traité en un seul ensemble pour l'estimation des teneurs. Ce type de modélisation serait inadapté pour des calculs plus complexes de risque de dépassement de seuil ; il faudrait alors modéliser la bimodalité.

Calculée avec les données de la maille « régulière », la carte variographique laisse pressentir une éventuelle anisotropie que l'extension du site, plus réduite suivant la direction nord-sud qu'en est-ouest, et l'irrégularité de l'échantillonnage, ne permettent pas de confirmer (Figure I-12). Par simplicité, un modèle isotrope est donc retenu.

Le variogramme des logarithmes est tracé à titre indicatif, car les variogrammes des teneurs et de leur logarithme sont généralement différents. Le variogramme des logarithmes présente un fort « effet de pépite », et une structure de portée inférieure à 500 m. Classiquement, la structure spatiale est moins marquée sur le variogramme des teneurs, car les écarts quadratiques sont amplifiés par rapport à ceux calculés sur les logarithmes. Le variogramme des teneurs, croissant pour les distances grandes relativement aux dimensions du site, indique une non stationnarité des teneurs à l'échelle du site. Ceci est compatible avec la présence des deux zones de plus fortes teneurs, situées dans la partie nord du site.

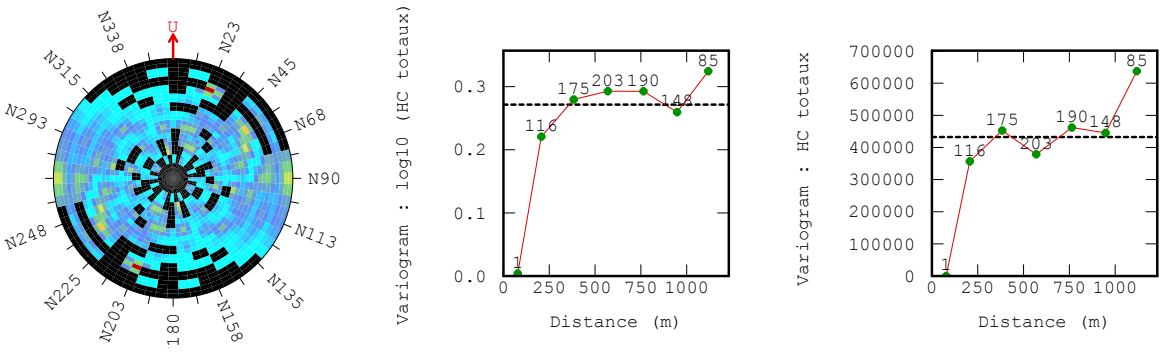


Figure I-12 Carte variographique et variogramme omnidirectionnel du logarithme et des teneurs calculées avec les données de la maille « régulière ». Le nombre de couples de points par pas du variogramme est reporté. Le premier point du variogramme, calculé avec un seul couple, ne peut être considéré comme significatif.

Les coefficients de variation restant voisins au nord-est et sur l'ensemble du site, faisons l'hypothèse que la structure spatiale est analogue pour ces deux ensembles, à un facteur près traduisant l'amplitude des fluctuations. L'échantillonnage plus dense au nord-est permet alors de préciser le comportement du variogramme aux petites distances.

Dans le quart nord-est, la carte variographique (non présentée) et les variogrammes directionnels confirment l'hypothèse d'isotropie aux petites distances (Figure I-13 pour les logarithmes et Figure I-14). La nuée variographique apparaît très dispersée, à cause de la teneur maximum des échantillons(valeur 8200 mg/kg ms), largement supérieure aux autres valeurs (deuxième valeur la plus forte : 3100 mg/kg ms). Supprimer momentanément cette valeur forte montre une nuée variographique assez homogène, et modifie peu l'allure du variogramme, sauf à un facteur près pour l'amplitude des écarts quadratiques.

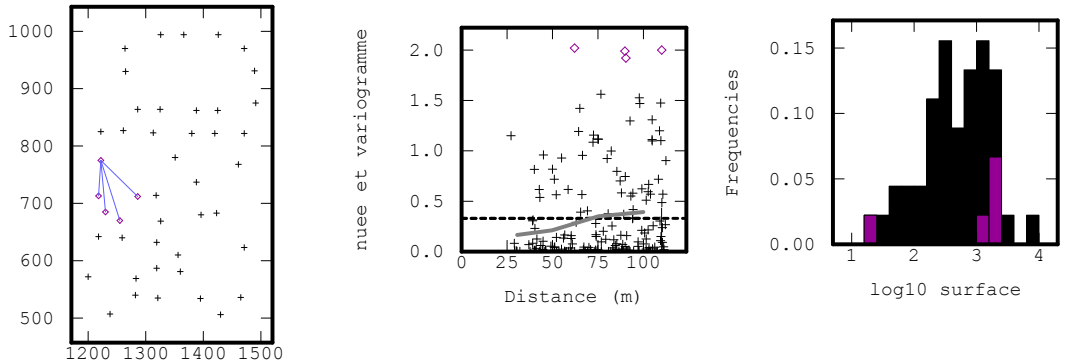


Figure I-13 Carte d'implantation, nuée variographique et histogramme de la zone nord-est. Logarithme décimal des teneurs en HAP. Les quatre écarts importants repérés en haut de la nuée variographique correspondent aux quatre couples de points reliés par un segment sur la carte d'implantation. Leur forte variabilité s'explique par le fort contraste des teneurs : le point commun aux quatre segments correspond sur l'histogramme au minimum et les quatre autres points à des valeurs fortes. Les cinq échantillons associés aux quatre écarts sur la nuée variographique sont reportés en mauve sur les graphiques.

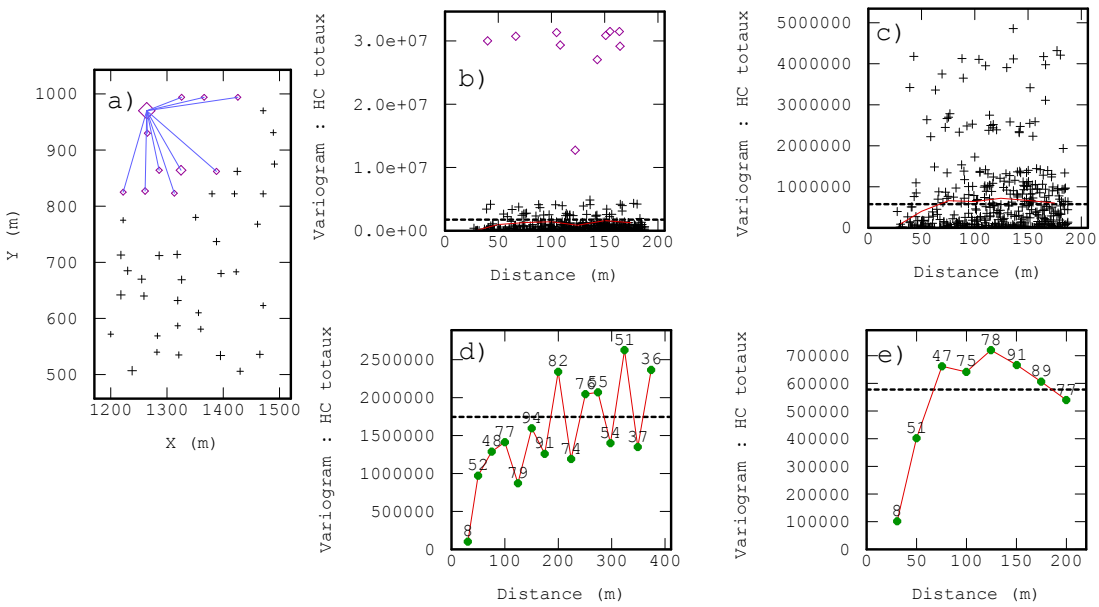


Figure I-14 Variographie pour les données de la grille dense située au nord-est. L'ensemble des écarts quadratiques élevés visibles sur la nuée variographique (b) provient du contraste entre le maximum des teneurs locales, et les données situées à proximité(a). Supprimer ce point fait apparaître une nuée variographique homogène (c). Calculés avec ou sans la teneur maximum (d, e), les variogrammes diffèrent principalement par le palier, plus de trois fois plus élevé lorsque le maximum est conservé (ce qu'il convient de faire).

L'échantillonnage plus dense révèle la présence d'une structure de petite portée, d'environ 250 m, voire même 100m si on enlève la teneur la plus forte de cette zone dense.

4.3 Ajustement du variogramme

L'ajustement du variogramme effectue une synthèse entre la structure à petite distance, inférée sur la grille dense au nord-est du site, et la structure aux distances plus grande, inférée sur la maille quasi régulière. La variabilité étant plus forte au nord-est du site, le variogramme aux petites distances est recalé à un facteur près pour l'ajustement du variogramme pour l'ensemble du site, utilisé pour le krigeage.

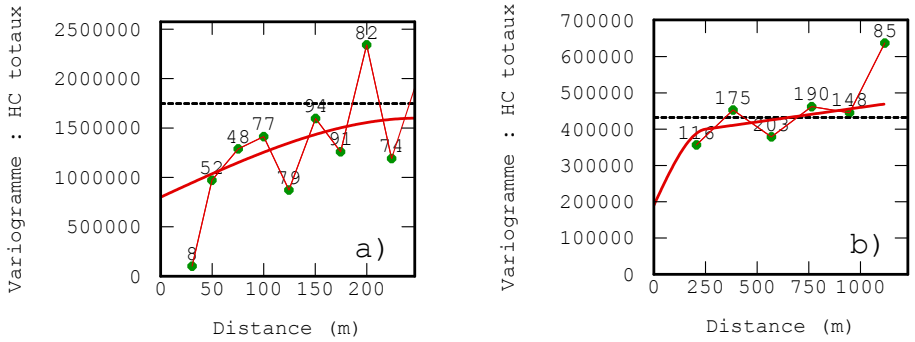


Figure I-15 Ajustement du variogramme des teneurs pour les données de la grille fine au nord-est (a) et de la grille régulière (b). A un facteur près, l'ajustement aux petites distances se retrouve dans le modèle global.

Le variogramme ajusté sur les teneurs de la grille fine au nord-est est la somme d'un effet de pépite et d'une structure sphérique de portée 250 m, les paliers de chacune des structures étant identiques et égaux à 8,5 10^5 mg/kg ms^2. Pour les données de la grille quasi régulière, ces paliers sont ramenés à 1,9 10^5 mg/kg ms^2 ; une composante supplémentaire est constituée par un variogramme linéaire, de coefficient 80 pour les distances exprimées en mètres.

4.4 Krigeage des teneurs

Les données de la grille quasi régulière étant peu nombreuses (44), on retient ici un voisinage unique, le krigeage atténuant l'effet de la redondance d'information portée par les données de la grille fine au Nord-Est. On estime la teneur moyenne sur des blocs de 25 m x 25 m. La superposition des points expérimentaux à la carte des valeurs krigées (Figure I-16) montre clairement :

- que la grille fine au Nord-Est déborde de la tache de pollution vers l'est, ce qui permet d'en préciser l'extension dans cette direction ;
- que, suivant le seuil de dépollution fixé, il peut être nécessaire de préciser l'extension des taches vers l'ouest et vers le centre par un échantillonnage complémentaire.

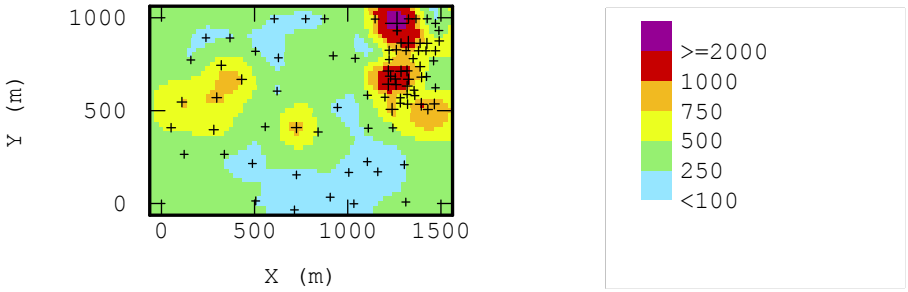


Figure I-16 Carte des teneurs moyennes (mg/kg ms) estimées sur des « blocs » carrés de 25 m de coté. Les points de prélèvements des données sont reportés.

La carte des teneurs estimées dépasse l'enveloppe des points expérimentaux. La carte de l'écart-type de l'erreur de krigeage (Figure 17a), fournie par le modèle, fait clairement apparaître une meilleure précision au Nord-est, due au resserrement de la maille. L'histogramme de l'écart-type d'estimation indique un mode de valeurs faibles, correspondant à cette zone (Figure 17b).

Remarque : Une modélisation plus fine nécessiterait de corriger la carte de la variance de l'erreur d'estimation pour tenir compte de la plus forte variabilité spatiale dans les zones de forte teneur. L'effet proportionnel, mis en évidence lors du calcul des variogrammes expérimentaux, a pour conséquence d'augmenter l'écart-type de l'erreur d'estimation dans cette zone.

La carte des valeurs estimées indique des teneurs plutôt faibles vers le pourtour ouest, sud et sud-est du site, là où les données sont peu nombreuses ; en l'absence de données sur ce pourtour, cette extrapolation peut s'avérer illusoire, et même risquée puisqu'elle indique des valeurs faibles. Il est préférable de limiter la carte des valeurs estimées aux seules zones pour lesquelles la précision est « suffisante ». Pour cela, il suffit de masquer les blocs associés aux valeurs fortes de l'histogramme de l'écart-type. Dans cet exemple, les blocs pour lesquels l'écart-type de krigeage dépasse 450 mg/kg ms ont été effacés de la carte (Figure I-17 et Figure I-18).

En effet, un écart-type de krigeage élevé exprime que l'erreur de krigeage peut être élevée. Ici, un écart-type de krigeage de 450 mg/kg ms correspond à un intervalle de confiance conventionnel de ±900 mg/kg ms autour de la teneur estimée ! A cause de la forte composante pépitique du variogramme, l'écart-type d'estimation est partout supérieur à 200 mg/kg ms, soit un intervalle de probabilité conventionnel de ±400 mg/kg ms autour de la teneur estimée.

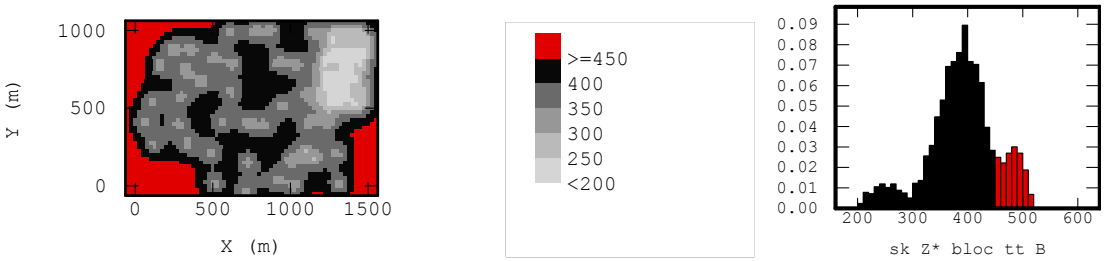


Figure I-17 Carte et histogramme de l'écart-type de krigeage de la teneur moyenne des blocs carrés de 25 m de coté. L'écart-type pour les blocs indiqués en rouge dépasse 450mg/kg ms.

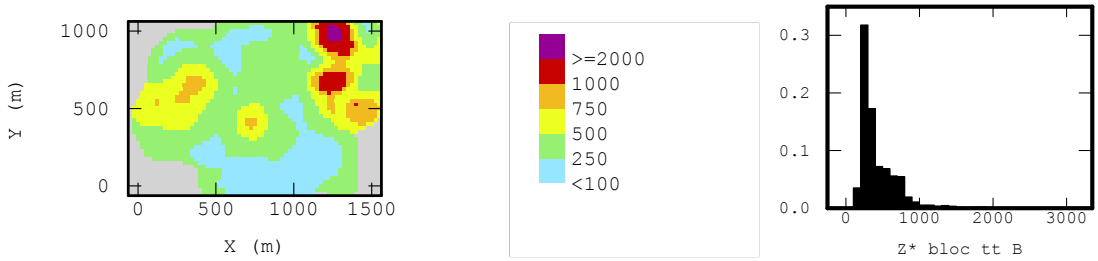


Figure I-18 Carte et histogramme des teneurs estimées sur des blocs carrés de 25 m de coté, après suppression des blocs pour lesquels l'écart-type d'estimation dépasse 450mg/kg ms.

Après suppression des zones estimées en extrapolation, les statistiques des teneurs estimées sont indiquées au tableau. Les statistiques avant suppression de ces zones, plus « optimistes », figurent également, à titre indicatif. La moyenne pour les 44 échantillons de la grille quasi régulière de 434mg/kg ms, diffère peu de la moyenne des teneurs estimées des blocs.

L'écart-type pour ces prélèvements (658 mg/kg ms) est largement supérieur à celui des teneurs estimées des blocs (301 mg/kg ms) : effet de support (la variance diminue quand le support augmente) et effet de lissage du krigeage se conjuguent pour réduire la variance des valeurs estimées par rapport à la variance expérimentale des échantillons.

Tableau I-2 Statistiques de la teneur *estimée* sur des blocs carrés de 25 m de coté, avec éventuelle sélection sur l'écart-type de krigeage. Le quantile à 25% est la valeur dépassée par les ¾ des blocs estimés ; la médiane, la valeur dépassée par la moitié des blocs, et le quantile à 75%, la valeur dépassée par ¼ des blocs.

Sélection	Effectif	Minimum (mg/kg ms)	Maximum (mg/kg ms)	Moyenne (mg/kg ms)	Quantile à 25% (mg/kg ms)	Médiane (mg/kg ms)	Quantile à 75% (mg/kg ms)
Avec	2465	169.	2962.	440.	254.	327.	551.
Sans	2925			431.	265.	340.	503.

4.5 Conclusions

L'analyse exploratoire des données, incluant en particulier l'analyse variographique, est une étape indispensable à une bonne compréhension de l'organisation de la pollution sur le site. Si elle prend du temps, elle en fait également gagner beaucoup plus ultérieurement, grâce à une identification précoce des problèmes. Cette analyse valorise ainsi considérablement l'échantillonnage, souvent très onéreux. Elle est à la base des développements présentés dans les chapitres suivantes.

La connaissance du terrain est un préalable indispensable pour adapter l'échantillonnage au site, et faciliter ainsi l'analyse variographique et l'estimation. L'étude de la variance d'estimation peut aussi sous certaines conditions aider à définir la maille d'échantillonnage (voir le thème N°4 pour plus d'information sur l'échantillonnage). On évitera néanmoins d'utiliser un échantillonnage préférentiel.

La carte d'estimation permet une visualisation des zones de teneur forte ou faible, et est un préalable souvent indispensable aux études complémentaires en vue d'une éventuelle dépollution. La carte de l'écart-type d'estimation s'avère réellement utile pour mettre en évidence les zones où l'estimation est peu précise, en particulier dans les zones estimées en extrapolation ou celles moins bien échantillonnées.

Même dans le cas où l'estimation est suffisamment précise, il est important de garder à l'esprit qu'une carte est une modélisation et non pas la réalité.

Analyse exploratoire, carte des teneurs krigées ou de l'écart-type d'estimation sont autant de résultats qui apportent une réelle connaissance du site.

II THÈME N°2 : COMMENT DÉLIMITER ET QUANTIFIER LES ZONES À DÉPOLLUER ?

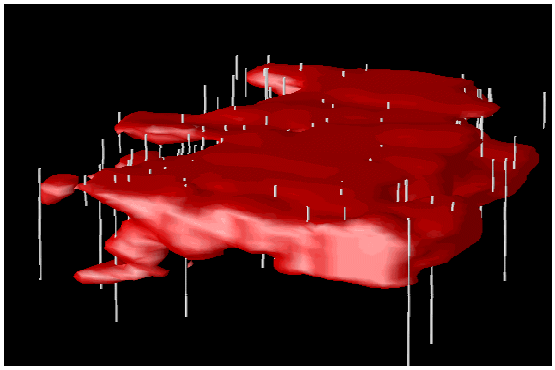
Comment délimiter et quantifier les zones à dépolluer ?

Fiche technique N°2

NOUS SOMMES TOUS CONFRONTES A CE PROBLEME

…Après des années de pratique, la délimitation et l'estimation de quantités de matériaux à dépolluer (sols, parfois déchets) reste un problème majeur pour les professionnels des sites contaminés.

L'information disponible en fin d'évaluation des risques est toujours incomplète et hétérogène, que le nombre de données soit faible ou important, mais il faut pourtant souvent s'en contenter pour localiser les zones contaminées nécessitant un traitement, estimer les volumes ou tonnages de matériaux correspondants, dresser un budget de réhabilitation ou dimensionner un projet de dépollution.



LES PRATIQUES ACTUELLES

Les praticiens ont alors recours à des méthodes d'estimation plus ou moins sophistiquées, allant de la simple délinéation manuelle de contours sur la carte des points renseignés, au calcul par des règles empiriques de voisinage ou par cartographie (interpolation) des teneurs.

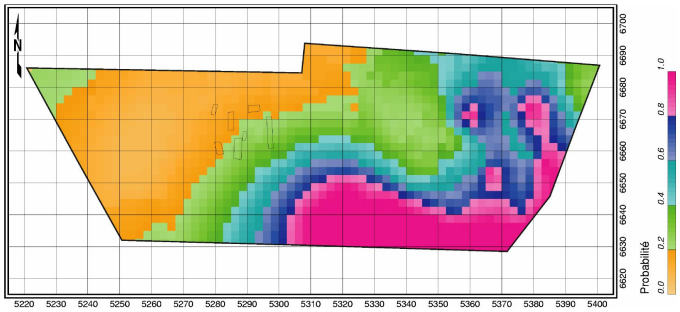
Ces méthodes se sont révélées insuffisantes à maintes reprises au cours des 15 dernières années. En effet, les chantiers où des terres polluées plus nombreuses que prévues ont été découvertes ne se comptent plus. Les conséquences en sont souvent fâcheuses : dépassement du budget initial, retard dans la réception des travaux, contentieux entre le responsable du site contaminé et le bureau d'étude, risques résiduels pour l'homme et son environnement. A l'inverse, combien de projets de réaménagement ont été surdimensionnés quand ils n'ont pas été simplement abandonnés faute de pouvoir évaluer convenablement l'étendue de la pollution.

Deux causes essentielles expliquent ces erreurs d'estimation qui dépendent du nombre et de la qualité des données disponibles. **La première se trouve dans la variabilité spatiale des teneurs dans les sols**, variabilité qui prend des formes différentes selon l'origine de la pollution. Sans analyser et sans reproduire convenablement cette variabilité dans les modèles, les cartes de teneur ou les délimitations de zones polluées ne sont pas des représentations réalistes de l'état de contamination des sols. **La seconde cause provient des incertitudes**. Données et estimations sont nécessairement entachées d'incertitudes qui ne peuvent être ignorées pour juger objectivement de l'état de contamination de sols.

CE QUE NOUS APPORTE LA GEOSTATISTIQUE

La géostatistique fournit le formalisme et les outils nécessaires pour pallier ces inconvénients. Appliqués convenablement à chaque situation, ces outils permettent de modéliser de façon réaliste la répartition spatiale des teneurs ainsi que l'incertitude sur ces teneurs.

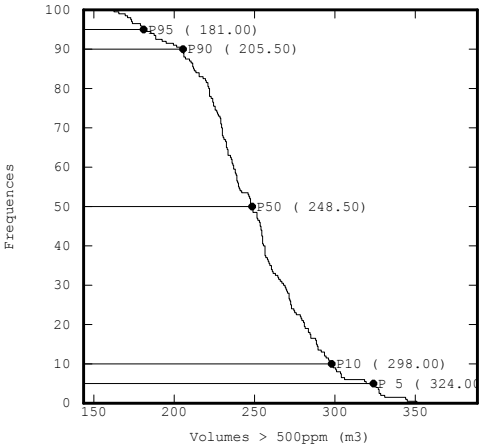
Délimiter et quantifier des zones à dépolluer revient à classer des matériaux comme contaminés ou sains sachant que l'état de connaissance de la pollution est incertain. L'approche géostatistique permet d'établir cette classification selon différents critères compréhensibles par tous. **Ces critères permettent d'apprécier le risque financier, sanitaire ou simplement statistique associé à la réhabilitation.** Il s'agit ensuite de minimiser ce risque en fonction des incertitudes reconnues.



Si l'approche géostatistique s'attache à reproduire la structure (corrélation) spatiale de données de pollution, **l'intérêt d'une telle approche reste entier lorsque cette corrélation est absente, comme dans le cas de sols fortement remaniés.** L'absence **révélée** de continuité spatiale des teneurs est un résultat en soi qui indique que tout modèle prédictif, quel qu'il soit, restera fortement incertain sur l'emplacement de la pollution.

…ET AUSSI UN OUTIL D'AIDE A LA DECISION

L'apport de la géostatistique va au-delà de la simple localisation et quantification des matériaux à dépolluer. A partir des résultats produits, il est en effet possible :



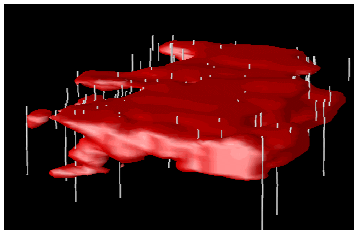
- De quantifier le risque de mauvaise classification des sols comme contaminés ou sains pour en évaluer ensuite les conséquences et agir en connaissance de cause (confirmation globale ou partielle de la dépollution, tri sélectif en dépollution, report après échantillonnage complémentaire, etc.).
- D'appliquer une fonction de coût aux estimations volumiques afin d'estimer le budget de dépollution et l'incertitude sur ce budget.
- De rationaliser le schéma de tri des matériaux en cours de dépollution.
- De comparer différents scénarios, que ce soit au niveau des objectifs de dépollution, des mesures de dépollution ou des réaménagements du site envisagés après réhabilitation.

Il s'en suit une aide globale au projet de réhabilitation, sur ses aspects financiers, techniques et environnementaux.

Pour en savoir plus…

Concernant ce thème, le manuel méthodologique ***GeoSiPol*** présente une approche géostatistique standard permettant de délimiter et de quantifier les matériaux à dépolluer. Elle montre **comment les résultats d'estimation peuvent être utilisés comme éléments d'aide à la décision dans un projet de réhabilitation**. Cette approche, aisée à mettre en œuvre et dont les performances ont déjà été testées et validées, est illustrée sur le cas complexe d'un ancien centre de production de gaz présentant une contamination multiple.

Note: Première figure tirée de Houlding, S.W ., « Practical Geostatistics : Modeling and Spatial Analysis, Springer Verlag, ISBN 3-540-66820-9



Comment délimiter et quantifier les zones à dépolluer ?

1 PRESENTATION DU THEME

1.1 Introduction

Dans de nombreuses situations, les professionnels de la réhabilitation des sites contaminés sont amenés à délimiter et à estimer les quantités de matériaux à dépolluer (volumes ou tonnages de sols ou remblais). Sont à dépolluer les matériaux où des teneurs en polluants, mesurées ou estimées, dépassent des ***seuils***. Ces seuils peuvent être des teneurs critiques établies sur la base d'une Evaluation Détaillée des Risques (EDR) ou bien des critères d'acceptation en décharge ou en filière de dépollution.

L'information disponible sur l'état de contamination d'un site est toujours ***partielle*** : les teneurs en polluants ne sont connues qu'en un nombre fini de points d'échantillonnage, plus ou moins nombreux et plus ou moins bien répartis sur le site selon les contraintes réglementaires et les impératifs budgétaires auxquels doit répondre le propriétaire du site ou le responsable de la contamination. L'information est également ***hétérogène***.

- Ancienneté : les données proviennent souvent de différentes phases de reconnaissance qui peuvent être largement espacées dans le temps.
- Nature : les teneurs en polluants peuvent avoir été mesurées en laboratoire ou sur site et être complétées par d'autres informations qui renseignent indirectement sur la contamination (lithologie, critères organoleptiques).

Nous nous plaçons ici dans la situation où les campagnes de reconnaissance ont été menées à terme. Des évaluations chiffrées, les ***plus précises possibles***, des volumes de sols à dépolluer et des coûts associés doivent être produites. A ces évaluations s'ajoute la localisation sur le site des sols jugés contaminés. Evaluations et localisation demeurent cependant incertaines et leurs ***incertitudes*** doivent pouvoir être quantifiées pour être prises en compte comme éléments de décision. Il s'agit à ce stade de définir des filières de traitement, de stockage ou de revalorisation des matériaux selon leur niveau de contamination.

Dans un tel contexte, les outils méthodologiques employés par les praticiens sont actuellement rudimentaires, du fait bien souvent de fortes contraintes temporelles et budgétaires lors de la réalisation des études, mais aussi de la méconnaissance d'outils adaptés à leurs problèmes. L'objectif de ce chapitre est de définir et de montrer l'intérêt d'une démarche géostatistique pour estimer et localiser des zones à dépolluer en prenant en compte les incertitudes. Une ***démarche géostatistique standard*** et ***aisée à mettre en œuvre*** y est présentée et illustrée sur le cas réel d'un ancien centre de production de gaz.

1.2 Cadre réglementaire

L'évaluation et la réhabilitation des sites contaminés français suivent une démarche réglementaire stricte, présentée dans les Guides méthodologiques « Gestion des sites « potentiellement » pollués » du Ministère de l'Aménagement et du Territoire (2000). La délimitation des zones contaminées et l'estimation des quantités de matériaux à dépolluer interviennent à deux étapes de l'approche réglementaire.

- En fin d'EDR, lorsque des objectifs de réhabilitation sont définis.

- Lors de l'élaboration d'un projet de réhabilitation. Le site contaminé est divisé en zones selon les filières d'évacuation ou de dépollution envisagées, les restrictions d'usage mises en place et les types de réaménagement prévus. En cas de dépollution avec excavation des terres, un schéma de tri sélectif des matériaux est planifié à partir d'une classification des matériaux comme sains ou contaminés et de critères de décision choisis pour prendre en compte les incertitudes. Nous appelons **sain** un sol dont les teneurs en polluants sont conformes à la filière de revalorisation envisagée, **contaminé** un sol dont les teneurs ne sont pas conformes à cette filière.

1.3 Pratiques habituelles

Les pratiques les plus courantes pour la localisation des zones à dépolluer et la quantification des volumes de sols à traiter sont les suivantes.

1. Des règles empiriques sont appliquées selon une logique établie par les praticiens pour prendre en compte l'origine de la contamination et le contexte lithologique du site. Le principe revient à étendre à des mailles ou à des zones entières les teneurs en polluant mesurées localement sur le site. Ce type d'approche permet au praticien de contrôler par lui même, mais forcément de manière subjective, l'extension de la contamination autour des points identifiés comme pollués. Les conséquences sont des délimitations biaisées ou incomplètes de la pollution, selon que la personne souhaite en minimiser ou maximiser l'importance, mais aussi difficilement justifiables et inévitablement liées au praticien.
2. Des logiciels de cartographie sont employés pour délimiter les zones polluées. Ces logiciels servent à calculer les teneurs aux nœuds d'une grille ou de tout autre maillage pour en déduire des lignes de contour (en 2D) ou des surfaces (en 3D). En chaque nœud, une valeur (unique) de teneur est calculée comme *moyenne pondérée* des données avoisinantes (combinaison linéaire)[1]. Rentrent dans cette catégorie toutes les méthodes d'interpolation (par ex. par distance inverse) mais aussi le krigeage (souvent en boîte noire) pour l'estimation directe de valeurs de teneur. En pratique, le choix d'une méthode ou des paramètres d'une méthode se fait encore dans un souci de contrôle de l'étendue des zones polluées sans justification rationnelle de ce choix. Effet de lissage[2] et forte influence des valeurs extrêmes sont les caractéristiques principales de ces méthodes.
3. Les zones contaminées et les volumes de sol correspondant sont délimités manuellement par de simples contours autour des points de données de diagnostic. Des trois pratiques, celle-ci paraît encore plus subjective et difficile à justifier.

Toutes ces pratiques conduisent à affecter une teneur en polluant à chaque localisation qui peut être un point, une maille ou un secteur. **Cette teneur est considérée implicitement comme exacte (vraie) en ignorant toute erreur ou incertitude d'estimation possible.** Elle est ainsi nécessairement inférieure ou supérieure à un seuil critique. Dans le premier cas, les sols sont considérés comme sains, dans le second cas ils sont pollués. Ce mode de classification est déterministe[3] et sans appel sauf au moment de dépolluer où les erreurs d'estimation jusque là négligées apparaissent et sont subies par les acteurs de la revalorisation des sites. Les conséquences possibles sont nombreuses et constituent les préoccupations premières des professionnels.

---

[1] Toutes les méthodes de cartographie consistent à donner un poids plus ou moins important aux données en fonction de leur distance par rapport au point de calcul. La différence d'une méthode à l'autre se trouve dans la manière dont les poids (pondérateurs) sont calculés.

[2] L'effet de lissage se traduit par des variations beaucoup plus douces des teneurs que celles observées en réalité. Il s'ensuit des cartes de teneur erronées et très peu réalistes de la pollution.

[3] On parle d'approche déterministe par opposition à une approche probabiliste qui quantifie et gère les incertitudes d'estimation.

- La découverte tardive de volumes de sols à dépolluer plus importants que prévus et entraînant une augmentation des coûts de dépollution et des délais de réhabilitation.
- Une surestimation des coûts de dépollution qui peut signifier l'abandon d'un projet de réhabilitation.
- Le mauvais dimensionnement des procédés de dépollution dont l'efficacité ou la rentabilité s'en trouve réduite.
- L'aboutissement à des contentieux juridiques entre le praticien, à l'origine des estimations et délimitations des zones à dépolluer, et l'organisme responsable qui finance la dépollution.

Au-delà des incertitudes, ces pratiques présentent plusieurs autres inconvénients ou limitations importantes qui ne font qu'accentuer les risques d'erreur dans l'estimation des volumes à dépolluer. Nous pouvons citer l'impossibilité de prendre en compte l'influence du support de l'information (voir § 2.2), les corrélations avec des données auxiliaires (par ex. lithologie, mesures géophysiques) ou encore la variabilité spatiale des teneurs. Ce dernier point mérite une attention particulière. En effet, une représentation réaliste des teneurs signifie nécessairement une reproduction fiable de la manière dont elles varient horizontalement, verticalement et parfois par zone ou par type de sols. Dans une approche géostatistique, cette variabilité spatiale s'exprime en terme de corrélation spatiale[4]. Elle est établie à partir des données et reproduite par les modèles en simulant les teneurs comme cela est décrit plus loin. Tel n'est pas le cas de ces pratiques habituelles. La notion de continuité spatiale y est présente mais elle est dictée par la méthode utilisée (pratique 2) ou simplement par le praticien (pratiques 1 et 3). Qui dit **mauvaise représentation de la variabilité spatiale des teneurs** dit nécessairement **mauvaise délimitation des zones polluées** qui peuvent se retrouver très fortement sur ou sous-estimées selon les valeurs et la disposition des données.

1.4 Intérêt d'une approche géostatistique

La géostatistique fournit le formalisme probabiliste nécessaire pour aborder les aspects d'incertitudes locales sur les teneurs (localisation des sols contaminés), d'incertitudes globales sur les volumes de sols pollués, de classification des sols comme sains ou contaminés (à dépolluer) et de quantification de l'erreur de classification. Les apports d'une approche géostatistique pour délimiter et quantifier les zones à dépolluer peuvent se résumer comme suit.

- **Génération de modèles réalistes de la pollution** (§3.3.1) : délimitation de zones polluées et calcul de volumes de sols nécessitent une représentation réaliste de la pollution en particulier en terme de variabilité spatiale des teneurs (approche par simulation conditionnelle par opposition à l'interpolation ou l'estimation directe de teneurs par krigeage).
- **Intégration de données de nature diverse** : teneurs en polluant mesurées en laboratoire, mesures de paramètres globalisants (ex. : somme des 16 concentrations individuelles des Hydrocarbures Aromatiques Polycycliques de la liste de l'US EPA, appelée communément indice HAP ; Hydrocarbures Totaux), analyses rapides sur site de moindre précision ou de support (mode d'échantillonnage) différent, indices organoleptiques ou encore informations lithologiques. Les méthodes géostatistiques permettent d'exploiter l'ensemble des informations disponibles pour améliorer la représentativité et la précision des modèles. Nous nous réfèrerons au thème N°3 « Comment intégrer toute l'information, de l'historique aux analyses sur site ? » pour plus d'information à ce sujet.
- **Prise en compte des conditions d'échantillonnage** (voir §3.6) : la géostatistique permet de modéliser le changement de support et l'effet d'information qui sont sources potentielles d'erreur de classification des sols en cours de dépollution.
- **Prise en compte des incertitudes pour la classification des sols** selon des critères bien définis (voir §3.4.4) : l'approche géostatistique peut s'adapter à toutes les situations de réhabilitation, notamment aux conditions de tri des sols en dépollution, au choix des filières de dépollution, à la mise en place de zones de servitudes ou au choix des types de réaménagement.

[4] La notion de corrélation spatiale formalise le fait que des teneurs mesurées en deux points distincts tendent à être d'autant plus semblables que la distance de séparation entre points est petite.

- **Prise en compte des incertitudes** pour l'évaluation des erreurs de classification et la prise de décision (voir §3.4.5) : vers un rendu des résultats sous une forme compréhensible comme éléments d'aide à la décision pour les responsables des dépollutions de sites.

1.5 Objectifs

Ce chapitre présente une **approche géostatistique en quatre étape**s pour traiter de la problématique de la délimitation et de la quantification de zones à dépolluer. L'approche proposée se veut standard dans la mesure où, d'une part, elle s'inspire de démarches similaires déjà appliquées et vérifiées sur plusieurs cas d'étude de sites pollués, d'autre part, elle ne fait intervenir que des méthodes géostatistiques bien connues et fréquemment employées.

Cette approche est adaptée aux milieux sol et sous-sol. Elle convient à l'étude de pollutions statiques dans le sens où les phénomènes de migration ou de dégradation des polluants n'ont pas d'effet significatif sur les teneurs à l'échelle de temps d'une étude, c'est-à-dire depuis les premières campagnes d'échantillonnage jusqu'au démarrage de travaux de dépollution. Si des phénomènes dynamiques d'écoulement, de transport ou de changement d'état des polluants devaient affecter leurs teneurs dans les sols, la même approche géostatistique pourrait s'appliquer mais avec des méthodes d'estimation plus avancées pour prendre en compte les tendances spatio-temporelles plus ou moins complexes observées ou attendues sur les teneurs. La seule condition est que les objectifs de dépollution restent les mêmes. Si tel n'est pas le cas, une approche plus spécifique serait à prévoir pour répondre à ces nouveaux objectifs. Ces aspects ne sont pas abordés ici mais pourraient faire l'objet de thèmes ultérieurs plus spécialisés.

Conformément à la terminologie couramment employée par les praticiens, nous parlerons dans la suite de ***sols*** pour désigner les matériaux solides étudiés qu'ils soient naturels ou rapportés. Nous parlerons également de ***volumes à dépolluer*** sachant qu'ils peuvent être facilement traduits en tonnages de sols à traiter ou en coûts de dépollution.

2 POSITION DU PROBLEME ET HYPOTHESES DE TRAVAIL

L'étude de zones à dépolluer s'articule autour de cinq notions fondamentales qui peuvent se résumer comme suit.

- Notion de *seuil critique* pour déterminer si des sols sont contaminés ou sains.
- Notion de *teneur de maille ou de bloc*, sous entendu de dépollution, pour évaluer l'état de contamination des sols à partir de données relatives à des échantillons de sols de beaucoup plus petites dimensions.
- Notion de *critères de classification* comme moyen de prendre en compte l'incertitude sur les teneurs de maille.
- Notion de *critères de décision* comme moyen de juger de la validité globale ou partielle de la classification des sols et d'en tirer des lignes de conduite opérationnelles.
- Notion de tri sélectif en dépollution où des teneurs de maille doivent pouvoir être déduites à partir d'un échantillonnage réduit ou d'échantillons composites prélevés dans les mailles.

Ce paragraphe pose le problème de la délimitation et de la quantification des zones à dépolluer en ces termes et précise les hypothèses de travail dans lesquelles la démarche géostatistique s'inscrit. Une formalisation y est proposée des notions précédentes qui sont souvent abordées de façon intuitive dans le métier de la dépollution. Le vocabulaire employé dans ce chapitre y est aussi introduit.

2.1 Seuils de teneurs

Nous supposons que des ***seuils critiques de teneurs*** ont été définis pour chaque polluant. Ils peuvent découler indifféremment :

- d'objectifs de réhabilitation définis pour un usage donné du site, suite à l'élaboration d'une EDR santé et éventuellement eau.
- de seuils d'acceptation en décharge contrôlée ou en filière de dépollution spécialisée.

Dans les deux cas, il nous faut évaluer l'extension de la contamination ainsi que les quantités de sols à dépolluer (volumes, tonnages). Cette évaluation passe par un **mode de détermination** des sols contaminés (à dépolluer) et de ceux qui sont sains. Pour un seuil critique donné, des sols seront jugés comme contaminés si leur teneur est supérieure au seuil. Ils seront sinon considérés comme sains. Si plusieurs seuils critiques sont associés à des niveaux de pollution différents, seront affectés à un niveau de pollution les sols dont la teneur dépasse le seuil critique de ce niveau mais pas le suivant.

Il est important de noter qu'un seuil critique n'a de signification complète que s'il est associé à une échelle (support) à laquelle les teneurs en polluant sont à évaluer. Des teneurs (moyennes) de maille sont généralement calculées sur une grille de dépollution (voir §2.2).

2.2 Teneurs de maille

Nous supposons de plus que la dépollution s'accompagne de l'***excavation des sols*** pour un traitement sur site ou hors site. Ce mode de dépollution fréquemment employé constitue l'objectif premier de la démarche géostatistique proposée[5].

Les sols à excaver sont établis à l'échelle d'un maillage du site que nous désignerons dans la suite par **maillage de dépollution**. Des teneurs moyennes doivent être estimées à l'échelle des mailles dont les volumes sont beaucoup plus grands que ceux des échantillons de sols analysés. Au-delà de la modélisation de la répartition spatiale des teneurs dans les sols, le passage de teneurs sur échantillons à des **teneurs de maille** doit également être géré. L'importance de ce changement d'échelle, appelé aussi **changement de support**, se trouve dans le fait que les teneurs ne prennent généralement pas les mêmes valeurs aux deux échelles.

D'un point de vue statistique, les teneurs sur échantillons de sols et les teneurs de maille présenteront des moyennes semblables. La dispersion des teneurs autour de la moyenne tendra cependant à être plus grande pour les échantillons que pour les mailles avec « des teneurs fortes plus fortes et des teneurs faibles plus faibles » (voir par exemple Rivoirard 1994 au § Bibliographie, p.116).

S'il est possible sous certaines conditions d'estimer ou de simuler directement des teneurs de mailles (voir §3.3.2), il a été choisi ici de séparer l'aspect modélisation spatiale de celui de calcul de teneurs de maille. Le principe est le suivant.

1. Les teneurs sont simulées aux nœuds d'une grille fine, comparée au maillage de dépollution, les valeurs simulées étant représentatives d'une échelle équivalente à celle des échantillons de sols analysés. Nous parlerons de **grille de simulation** pour faire référence à une grille fine.
2. En appliquant le maillage de dépollution sur la grille de simulation où ont été simulés les teneurs, des teneurs (moyennes) de mailles sont déduites. Ces teneurs de maille peuvent intervenir également dans le cadre d'un processus de tri sélectif des sols en dépollution (voir §2.5).

L'avantage de cette modélisation en deux étapes est double. D'une part, la simulation des teneurs ne dépend pas du choix d'un maillage de dépollution. D'autre part, un plus grand nombre de méthodes de simulation et des méthodes plus simples peuvent être utilisées pour simuler des teneurs aux nœuds d'une grille (simulations dites de points) plutôt que des teneurs de mailles, avec ou sans données auxiliaires.

[5] D'autres situations de dépollution, en particulier in situ, pourraient faire l'objet de thèmes ultérieurs avec des objectifs différents.

2.3 Critères de classification

Comme cela a été évoqué plus haut, un des apports importants de la géostatistique se trouve dans la quantification possible des incertitudes, qu'elles soient locales sur les teneurs en polluants dans les sols ou globales comme sur des volumes de matériaux à dépolluer.

L'aboutissement de la démarche proposée ici est d'arriver en fin d'étude à classer toutes les mailles d'un maillage de dépollution comme saines ou contaminées à partir d'un état de connaissance actuel de la pollution (ensemble de données et modèles statistiques établis). Sans pouvoir dans bien des cas prédire avec suffisamment de confiance si la teneur d'une maille est supérieure ou inférieure à un seuil critique, des **critères de classification** doivent être définis pour gérer cette incertitude et réussir à affecter une classe à chaque maille. De cette classe dépend alors le devenir des sols de la maille entre traitement, mise en décharge ou conservation sur place.

Comme nous le verrons plus loin, les critères de classification peuvent être **purement statistiques** (§3.4.4), avec comme objectif de **minimiser le risque (statistique) de mauvaise classification des mailles, ou financiers,** en prenant en compte à la fois les incertitudes locales et les coûts associés à chaque situation de classification **pour minimiser les coûts d'erreur** (§3.5). Ces critères s'expriment assez simplement lorsque la contamination concerne un seul polluant affecté d'un unique seuil critique. Leur définition peut cependant être plus complexe dans les cas suivants.

- Plusieurs polluants à seuil. Il est fréquent qu'un site industriel présente des anomalies de teneurs en plusieurs substances. Tel est le cas des anciennes usines à gaz où différents hydrocarbures aromatiques polycycliques peuvent être détectés ainsi que des cyanures et des métaux lourds.
- Plusieurs types de réaménagement possibles sur un même site. Pour un même polluant, les seuils de dépollution dépendent alors du type de réaménagement. Tel est le cas du site de l'ancien centre de production de gaz présenté au §4. Le seuil de dépollution pour le cyanure y a été fixé à 2 mg/kg ms (matière sèche) en zone résidentielle et à 100 mg/kg ms en zone industrielle.
- Plusieurs filières de dépollution/restrictions d'usage/revalorisation/stockage des sols. Les seuils d'acceptation peuvent être différents par filière pour un même polluant. Par exemple, sur le site d'une ancienne usine à gaz, le tri des sols dépendait de trois seuils de teneurs en benzo(a)pyrène (Abrév. BAP) :

Teneurs en mg/kg ms	Intervention
BAP < 0.25	Pas de traitement
0.25 < BAP < 5	Mise en décharge de classe 3
5 < BAP < 25	Mise en décharge de classe 1
BAP > 25	Désorption thermique hors site + décharge de classe 1 pour les terres traitées

Dans tous les cas, le choix de critères de classification doit idéalement se faire en concertation avec les Autorités et les responsables du site. Loin de compliquer la procédure, la définition claire et précise de tels critères permet de replacer les choix économiques et politiques entre les mains des décideurs. Il ne revient en effet pas aux techniques de se prononcer sur le niveau de risque (sanitaire ou financier) auquel les responsables d'un site ou les décideurs sont prêts à prendre. Le bon déroulement des projets n'en est finalement que favorisé.

2.4 Critères de décision

A l'issue de l'étape de classification des sols, une classe est associée à chaque maille. Cette classe représente la meilleure prévision qui peut être faite sur l'état de contamination de la maille compte tenu des incertitudes et des critères de classification retenus. Il n'en demeure pas moins que les incertitudes présentes avant la classification subsistent après. Si la classification repose sur des critères financiers, il peut être déduit également les coûts de mauvaise classification possible des sols à partir des incertitudes sur les teneurs. Ces coûts même minimisés peuvent représenter des montants importants qui justifieraient d'en savoir plus sur l'état de contamination réel des sols et donc de décider de la réalisation d'investigations complémentaires.

La dernière étape de l'approche géostatistique consiste ainsi à **fournir des résultats utiles pour juger de la validité globale ou partielle (par zones) de la classification, évaluer les conséquences d'une mauvaise classification des mailles et décider avec discernement de la stratégie de dépollution et des actions à mener.**

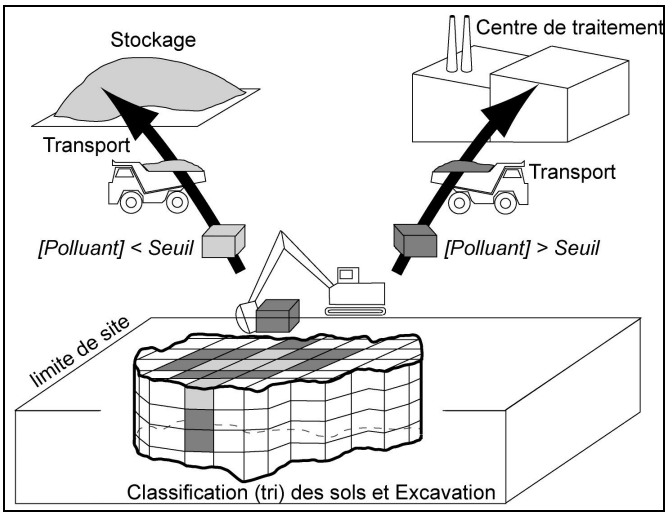
Nous parlerons pour cette étape d'**aide à la décision** et de **critères de décision**. Ces critères vont très au-delà de la géostatistique. Ils intègrent une vocation métier forte avec des décisions expertes sur les opérations de dépollution. A des situations bien définies doivent être associées des décisions opérationnelles qui pourront dépendre de l'expérience et des compétences des sociétés d'étude ou de dépollution impliquées dans un projet. Le domaine est vaste et nous nous limiterons aux quelques exemples simples suivants.

Type de situation	Préconisation
Identification de mailles très certainement saines.	Sols à laisser en place sans autre vérification.
Identification de mailles très certainement contaminées avec un niveau de contamination bien reconnu.	Traitement prévu des sols sans autre vérification.
Incertitude non négligeable sur des mailles classées comme contaminées ou saines.	Application d'un tri sélectif pour décider en dépollution du devenir de ces sols (voir §2.5).
Apparition de zones correspondant à des volumes de sols importants où la classification des mailles est fortement incertaine.	Mise en place de campagnes d'échantillonnage et d'analyse complémentaires de ces zones dans la mesure où les délais le permettent.

2.5 Tri sélectif des sols en dépollution

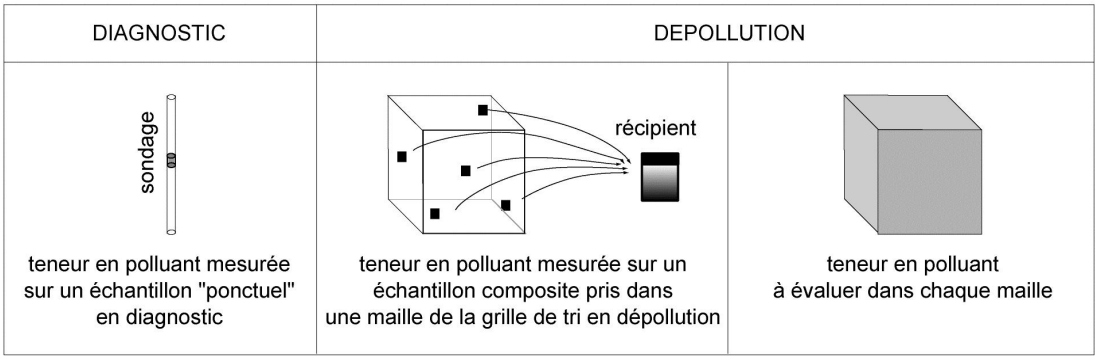
La dépollution est souvent accompagnée d'un tri des sols sur le maillage de dépollution. Des **échantillons composites** sont prélevés dans chaque maille pour être analysés sur site ou en laboratoire. Les résultats d'analyse sont comparés aux seuils critiques pour définir la filière d'évacuation adaptée aux sols de chaque maille (Figure II-1).

Nous supposons connaître la taille et éventuellement l'orientation des échantillons collectés dans les mailles pour pouvoir juger du ***support*** de l'information et en tenir compte (voir §2.2). Ces échantillons de sols devant être représentatifs des mailles, ***l'effet d'information*** peut être un autre phénomène à prendre en compte (Figure II-2). Il exprime l'écart possible entre la teneur moyenne vraie (mais inconnue) d'une maille et la teneur mesurée sur un échantillon représentatif de la maille. Cet écart peut être plus ou moins marqué selon la constitution de l'échantillon. Il tend à être plus fort lorsqu'un unique échantillon « ponctuel » de quelques cm^3 est pris au hasard dans la maille et plus faible lorsque plusieurs échantillons ponctuels sont prélevés systématiquement dans les mailles pour constituer un échantillon composite. Ces écarts peuvent être à l'origine d'erreurs de classification des mailles. Pour qu'ils puissent être estimés et corrigés, il est nécessaire de connaître le mode d'échantillonnage des mailles. Nous supposons cette information connue.



D'après Demougeot-Renard H. and De Fouquet C., 2004, "A geostatistical approach to assessing volumes of soil requiring remediation. Validation to lead-polluted soils underlying a former smelting works", Environmental Science and Technology, vol. 38, n°19, pp. 5120-5126.

Figure II-1 Schématisation des opérations de dépollution sur site et hors site avec tri des sols selon des critères spécifiques de classification et de décision.



D'après Demougeot-Renard and De Fouquet C., 2004, « A geostatistical approach to assessing volumes of soil requiring remediation. Validation to lead-polluted soils underlying a former smelting works", Environmental Science and Technology, vol. 38, n°19, pp. 5120-5126.

Figure II-2 Schématisation des effets de support et d'information. Effet de support : la taille des échantillons collectés en diagnostic est souvent très inférieure à celle des échantillons collectés en dépollution. Effet d'information : les échantillons collectés en dépollution sont des composites pris dans les mailles d'une grille régulière et sont supposés être représentatifs de ces mailles. Il existe cependant toujours un écart entre la teneur en polluant mesurée dans les composites et la teneur vraie moyenne des mailles.

3 DEMARCHE GEOSTATISTIQUE DETAILLEE

3.1 Méthodologie

Rappel sur le cadre de la modélisation

L'approche géostatistique s'inscrit dans le cadre suivant d'un projet de dépollution des sols d'un site.

- Campagnes d'investigation, EDR et études complémentaires éventuelles ont été réalisées et ont conduit à une décision de dépollution du site par excavation des sols contaminés.
- Maillage de dépollution du site : la taille des mailles détermine l'échelle à laquelle les teneurs en polluant doivent être évaluées à partir de données de teneur sur échantillons de sols (changement de support). Ce maillage peut être imposé, pour satisfaire à des contraintes de chantier ou autres, ou faire l'objet d'une analyse de sensibilité pour optimiser certains aspects de la dépollution[6].
- Classification des sols : il est attendu de l'approche de produire une classification de l'ensemble des mailles de la grille comme étant saines ou contaminées avec dans ce dernier cas une distinction possible de niveaux de contamination différents. Cette classification doit se faire selon un critère de classification à choisir (ou à définir) pour prendre en compte l'incertitude sur l'état de contamination réel des mailles compte tenu des données disponibles.
- Si la classification révèle des incertitudes importantes sur les zones à dépolluer, des éléments d'aide à la décision doivent pouvoir être produits pour décider d'une stratégie de dépollution adaptée (par ex. mise en place d'un tri sélectif).

L'approche géostatistique proposée pour délimiter les zones contaminées et estimer les volumes de sols à dépolluer comporte quatre grandes étapes.

1. **Analyse des données**
   - Détermination de caractéristiques statistiques et spatiales essentielles pour décrire la pollution des sols.
2. **Génération de simulations conditionnelles** (aussi appelées réalisations) des teneurs en polluants
   - Simulation de cartes réalistes de la répartition spatiale de la pollution dans les sols (représentation fine de la pollution telle que vue à l'échelle des échantillons analysés).
1. **Estimation de zones et de volumes contaminés** :
   - Calcul sur les réalisations des teneurs de maille pour un choix de maillage de dépollution.
   - Application de seuils critiques aux teneurs de maille calculées sur les simulations.
   - Statistiques sur les volumes de sols contaminés calculés sur les simulations.
   - Classification des mailles comme contaminées ou saines pour un critère choisi.
   - Localisation des zones classées comme contaminées et calcul du volume correspondant.
2. **Production d'éléments d'aide à la décision.**
   - Analyse de l'incertitude sur la classification des sols et éventuellement des coûts d'erreur.
   - Choix de critères de décision et application de ces critères à la classification des mailles.

[6] Vers un compromis entre des petites tailles de maille, qui permettent de délimiter au plus près les volumes contaminés mais avec beaucoup de mailles à traiter, et des grandes mailles qui moyennent davantage les teneurs avec pour conséquence des volumes à traiter qui peuvent être plus ou moins importants selon le cas.

Par souci de clarté et sans perdre en généralité, la présentation de l'approche se limite au cas d'un polluant unique associé à un seuil critique pour un type de réaménagement donné et avec une unique technique de dépollution. Les étapes en sont détaillées ci-dessous. Nous nous référerons à l'Index, p. 117, pour la terminologie géostatistique.

3.2 Analyse des données

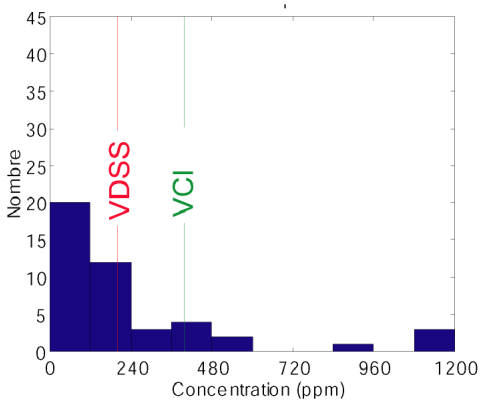
L'analyse exploratoire des données est une étape essentielle pour la compréhension des données et du phénomène de pollution de sols à modéliser. Ses principes généraux font l'objet du thème N°1 « Comment passer des analyses chimiques à la cartographie des teneurs en polluant ? », page 13. L'objectif est de valider les données disponibles, d'établir des statistiques fiables sur la teneur pour un support cohérent des données, d'analyser la corrélation spatiale et les corrélations possibles entre la teneur et des informations auxiliaires, d'en déduire des modèles statistiques représentatifs de la pollution des sols.

L'évaluation de zones et de volumes à dépolluer signifie aussi s'intéresser aux teneurs mesurées supérieures à un seuil critique. Nous pouvons nous interroger sur la proportion de ces teneurs ou encore sur les écarts entre le seuil critique et les valeurs mesurées. Dans le premier cas, la proportion de valeurs supérieures au seuil nous renseigne sur la proportion en volume de sols pollués sur le site mais pour une échelle fine d'observation en rapport avec la représentativité des données. Dans le second cas, la présence de données proches du seuil critique signifie une incertitude sur un dépassement ou non du seuil selon la précision des mesures ou la représentativité de l'échantillon de sol analysé. Lorsque le doute subsiste quant à savoir si une donnée est supérieure ou non à un seuil, la solution peut être d'éliminer la donnée, à défaut de prendre la décision conservatrice de la conserver comme étant supérieure au seuil (et donc favoriser la dépollution au bénéfice du doute). La localisation des points de dépassement des seuils sur une carte ou sur un bloc-diagramme du site permet également de vérifier et mieux interpréter la répartition de ces points et d'identifier les points isolés qui mériteraient davantage d'attention.

A titre d'exemple, le Tableau II-1 fournit un résumé statistique des teneurs en plomb mesurées sur un ancien site industriel et la Figure II-3 montre l'histogramme correspondant avec le positionnement des valeurs-guides réglementaires VDSS et VCI comme seuils critiques. Les proportions de valeurs supérieures aux seuils sont de 42% pour la VDSS et de 18% pour la VCI. En terme de volumes, près de 40% des sols seraient en situation de dépassement de la VDSS et correspondraient à un volume double comparé à la VCI.

Tableau II-1 Statistiques élémentaires sur les teneurs en plomb mesurées sur un ancien site industriel.

		Unités
Nombre d'observations	45	-
Médiane	130.00	*mg/kg ms*
Moyenne	263.42	*mg/kg ms*
Ecart-type	301.77	*mg/kg ms*
Coefficient de variation	1.15	-
Minimum	30.00	*mg/kg ms*
Maximum	1200.00	*mg/kg ms*



Taux de dépassement des valeurs-guides	
VDSS	42 %
VCI	18 %

Figure II-3 Histogramme des teneurs en plomb avec indication des proportions de valeurs mesurées supérieures aux valeurs-guides réglementaires VDSS et VCI usage sensible[7] ainsi que de l'ampleur de ces dépassements.

3.3 Génération de simulations conditionnelles

La génération de simulations n'est pas une fin en soi. Elle constitue une étape intermédiaire pour accéder à d'autres variables dont les incertitudes dépendent directement de celles de la variable simulée. Les simulations servent ici de base au calcul de volumes de sols pollués et à leur localisation.

3.3.1 Simulation versus estimation (krigeage)

La génération de simulations conditionnelles de teneurs se place comme une solution mieux adaptée que l'estimation par krigeage[8]. Sauf cas particuliers rares de données très abondantes ou de phénomènes fortement continus, l'estimation directe de teneurs par krigeage, cokrigeage ou par toute autre méthode d'interpolation produit en effet des cartes dont la particularité première est d'être lissées (voir aussi §1.3). Ce lissage se traduit à la fois par une continuité spatiale accentuée des teneurs et par une variabilité faible à beaucoup plus faible des valeurs estimées comparées aux

[7] VCI : Valeur de Constat d'Impact ; VDSS : Valeur de Définition de la Source Sol. La définition de ces deux valeurs utilisées en France est celle du guide méthodologique : « Gestion des sites (potentiellement) pollués » du Ministère de l'Environnement (Version 2, Annexe 5C révision du 9 décembre 2002).

[8] On désigne ici sous le terme général de krigeage, toute méthode de krigeage adaptée à un problème d'estimation linéaire. Ce terme exclut ici certaines méthodes, telles que le krigeage d'indicatrice ou le krigeage disjonctif, adaptées à un problème non linéaire de dépassement de seuil. Ces méthodes ne sont pas présentées dans cette fiche, car elles ne présentent ni la souplesse ni la simplicité de formalisme des simulations conditionnelles.

valeurs réelles[9]. L'application d'un seuil à des valeurs ainsi estimées fournit un estimateur biaisé (erroné) de la variable « teneur en polluant supérieure au seuil » et par conséquent des volumes de sols concernés.

Une simulation est une réalisation possible des teneurs en polluant sur le domaine d'étude. En plus de retrouver les valeurs aux points de mesure, elle reproduit, à l'échelle de la carte, l'histogramme (distribution de probabilité) et la structure spatiale (variogramme) des teneurs mesurées. A la différence d'une carte krigée, une simulation restitue donc la variabilité spatiale du phénomène étudié, de sorte que l'application d'un seuil à la simulation fournit un estimateur non biaisé des teneurs en polluant supérieures au seuil (Figure II-4).

Contrairement à l'estimation par krigeage qui, pour un choix de méthode et de paramètres associés, fournit une carte unique, la simulation conditionnelle peut produire un grand nombre de **réalisations** (cartes) qui sont autant d'images (réalistes) possibles de l'état de contamination du site (Figure II-5). Visuellement, toutes les réalisations se ressemblent mais dans le détail les teneurs varient plus ou moins d'une carte à l'autre en fonction des zones. Plus les variations sont importantes dans une zone ou en un point donné, plus grande est l'incertitude sur la teneur à cet endroit.

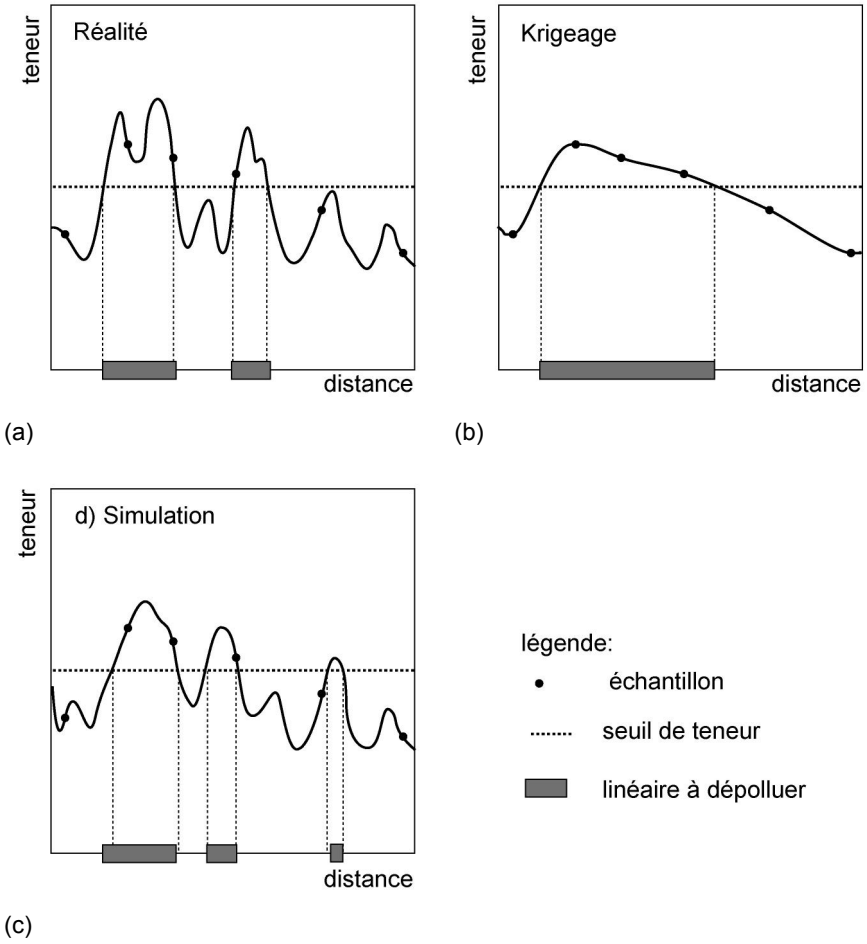


Figure II-4 Illustration de l'application d'un seuil à des valeurs réelles, à des valeurs estimées (par krigeage à moyenne inconnue, encore appelé krigeage ordinaire) et à des valeurs simulées le long d'un profil (1D). La simulation révèle ainsi des caractéristiques qui se rapprochent davantage de la réalité et rendent licite le calcul de zones (ici des linéaires) à dépolluer.

[9] On parle généralement de portées de corrélation plus grande pour ce qui est de la continuité spatiale et de variance plus faible pour ce qui est de la variabilité des valeurs.

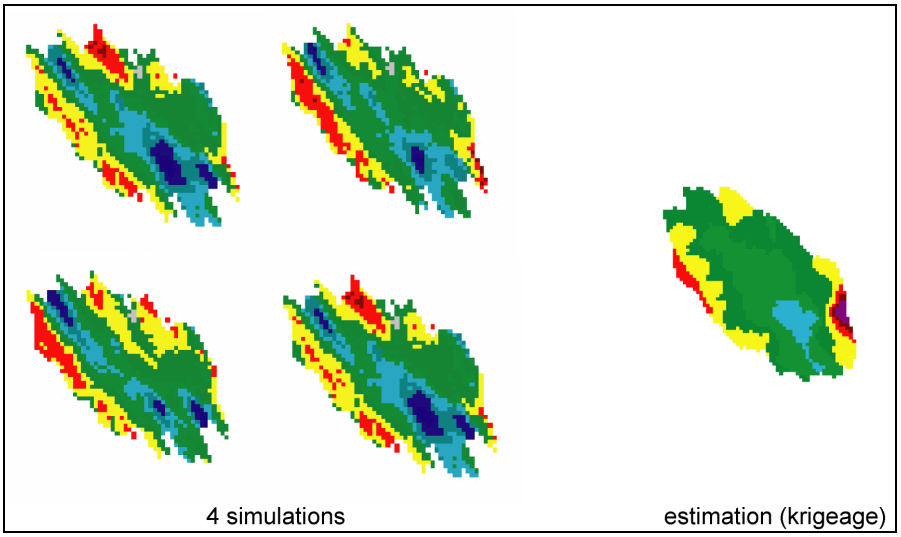


Figure II-5 Comparaison entre quatre simulations conditionnelles et l'estimation par krigeage du même phénomène à partir du même ensemble de données et du même modèle de variographie (corrélation spatiale). La même échelle de couleurs a été utilisée pour toutes les cartes.

3.3.2 Choix de la méthode de simulation

Rappel sur le cadre de la simulation

Les teneurs sont à simuler aux nœuds d'une grille (fine) qui recouvre le domaine d'étude. Les valeurs simulées sont représentatives d'un volume de sols (support) en rapport avec celui des échantillons de sols analysés. Ces derniers peuvent être *ponctuels* ou plus généralement composites (ou moyens), c'est-à-dire relatifs à une certaine épaisseur de sols le long d'un forage ou à l'intérieur d'une fosse.

Les méthodes les plus fréquemment utilisées pour simuler des variables continues et donc des teneurs en polluants sont (voir par exemple Chilès et Delfiner 1999, § Bibliographie, p.116) :

- Les simulations séquentielles : gaussiennes ou par indicatrices.
- La méthode des bandes tournantes.

La simulation séquentielle *consiste à générer successivement une valeur en chaque nœud de la grille de simulation, chaque nouvelle valeur simulée rentrant dans l'ensemble des données de conditionnement utilisées pour simuler le nœud suivant. Cette approche suppose de pouvoir estimer en chaque nœud la distribution de probabilité locale dite conditionnelle de la teneur en fonction des données avoisinantes et des nœuds déjà simulés*[10]*. Plusieurs méthodes existent pour estimer cette distribution. Elles donnent lieu à autant de méthodes de simulation différentes. Les plus courantes sont la simulation séquentielle gaussienne et la simulation séquentielle par indicatrices. La première se distingue pas sa simplicité et convient généralement bien à des teneurs. La seconde présente certains avantages pour intégrer des informations incomplètes sur l'état de contamination des sols. Une fois la distribution locale conditionnelle estimée, la valeur simulée est tirée aléatoirement dans cette distribution.*

Les étapes de la simulation séquentielle peuvent se résumer comme suit.

1. *Définir un chemin de visite aléatoire de tous les nœuds de la grille.*
2. *En chaque nœud, estimer la distribution locale conditionnelle par rapport aux données avoisinantes et aux valeurs précédemment simulées.*
3. *Tirer au hasard une valeur dans la distribution et l'affecter au nœud.*

[10] La distribution de probabilité locale conditionnelle décrit les valeurs que peut prendre la teneur à l'endroit du nœud, compte tenu des données de teneur qui l'entourent (mesures et valeurs déjà simulées), d'informations auxiliaires possibles et des corrélations qui peuvent exister entre toutes ces données (redondance de l'information) et avec la teneur au nœud simulé.

4. *Répéter les étapes 2 et 3 jusqu'à avoir parcouru tous les nœuds de la grille pour obtenir une réalisation.*
5. *Répéter les étapes 1 à 4 pour générer une nouvelle réalisation.*

Chaque méthode de simulation séquentielle présente des particularités que ce soit dans la préparation des données, la manière dont les distributions locales conditionnelles sont estimées ou encore les paramètres de la méthode. Nous pouvons mentionner les points suivants.

- *Simulation séquentielle gaussienne*
  - *La transformation gaussienne des données peut être nécessaire si, comme cela est généralement le cas, la teneur ne suit pas de loi Gaussienne. Les paramètres de simulation (par ex. le variogramme) doivent alors être relatifs aux données transformées et les valeurs simulées doivent subir une transformation inverse pour redonner des teneurs.*
  - *Les distributions locales conditionnelles étant supposées Gaussiennes, leur estimation se fait simplement en estimant la moyenne et la variance par krigeage simple.*
- *Simulation séquentielle par indicatrices*
  - *Une transformation par indicatrices des données est nécessaire. Elle consiste à se définir un certain nombre de seuils de coupure et à calculer pour chaque seuil une variable d'indicatrice qui prend la valeur 1 si la donnée est inférieure au seuil correspondant ou sinon 0. Les paramètres de simulation doivent alors être relatifs aux variables d'indicatrice.*
  - *Les distributions locales conditionnelles sont calculées en estimant par krigeage la probabilité que la teneur soit localement inférieure à chaque seuil de coupure.*

Les paramètres communs à toutes les méthodes de simulation séquentielles sont les suivants.

- *Distribution globale (sur l'ensemble du site) de la teneur, modèles de variogramme et éventuellement de covariogramme avec une variable auxiliaire : tirés de l'analyse exploratoire des données.*
- *Paramètres de recherche des données : ellipse de recherche qui définit l'étendu du voisinage autour d'un nœud (en rapport avec les portées de corrélation), le nombre maximal de données de teneur à prendre en compte pour l'estimation d'une distribution locale conditionnelle (généralement de 8 à 16).*

***L'algorithme des bandes tournantes** consiste à simplifier une simulation 2D (ou 3D) en plusieurs simulations 1D le long de lignes générées aléatoirement, puis à reconstruire la simulation 2D (3D) en moyennant les valeurs projetées des simulations 1D (Matheron 1973, Chilès and Delfiner 1999). Le seul paramètre requis pour assurer la cohérence des simulations (c'est-à-dire la reproduction de l'histogramme et du variogramme des données) est le nombre de bandes tournantes. Cet algorithme conduit à l'obtention de simulations non conditionnelles de la teneur en polluant : l'histogramme et le variogramme sont reproduits mais les données mesurées ne sont pas respectées. Le conditionnement des simulations aux données se fait alors par krigeage, et conduit aux simulations conditionnelles recherchées.*

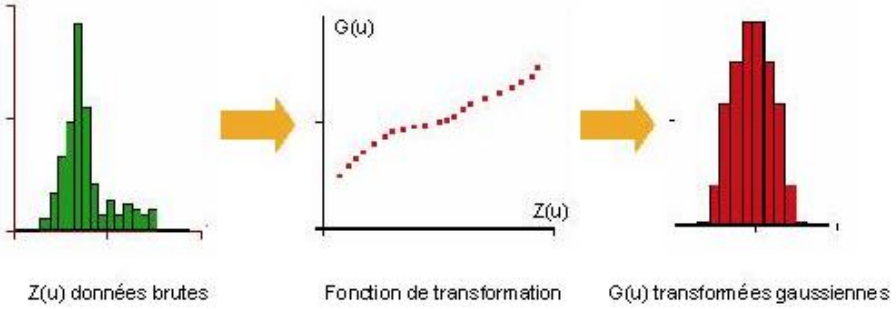
D'autres méthodes de simulations existent, mais sont moins fréquemment utilisées. Sans prétendre être exhaustif, citons (pour en savoir plus, se référer à la Bibliographie) :

- *La décomposition LU de la matrice de covariance (possible sur de très petites grilles).*
- *La simulation par champs de probabilité.*
- *La simulation par recuit simulé.*
- *Les méthodes spectrales.*

Par ailleurs, les simulations conditionnelles ne constituent pas les seules méthodes d'estimation adaptées au problème de dépassement de seuils. L'espérance conditionnelle, le modèle multigaussien, le krigeage d'indicatrices, ou bien encore le krigeage disjonctif sont adaptés à ce problème d'estimation non linéaire. Les méthodes de simulation ont cependant été retenues car elles offrent une flexibilité que n'ont pas ces méthodes, notamment lorsqu'il s'agit d'appliquer plusieurs seuils, plusieurs changements de support ou plusieurs fonctions de transfert.

Transformation gaussienne des variables

Certains des algorithmes de simulation décrit précédemment requièrent que la teneur en polluant suive une distribution gaussienne, ce qui n'est généralement pas le cas. Leur distribution est souvent dissymétrique en raison de leur positivité et de l'existence de quelques valeurs très fortes. Une utilisation adéquate de ces algorithmes passe par une transformation préalable des teneurs de sorte à obtenir une variable (transformée) gaussienne (par ex. transformation par anamorphose, gaussienne ou par logarithme). La figure suivante en illustre le principe.



Lorsque leur distribution n'est pas gaussienne, les données brutes sont transformées au moyen d'une fonction pour se placer dans les hypothèses d'utilisation des algorithmes de simulation des variables continues.

Les algorithmes de simulation de variables continues, telles que les teneurs en polluants, ont été développés dans le cadre de modèles de diffusion (Chilès et Delfiner 1999). Il est donc possible de vérifier que le phénomène à simuler entre dans le cadre du modèle sous-jacent à l'algorithme retenu. Beaucoup d'algorithmes de simulation de variables continues supposent non seulement que la variable suive une distribution gaussienne mais aussi qu'elle soit bigaussienne dans l'espace. Des tests existent pour vérifier l'hypothèse de bigaussiannité (voir par ex. Lajaunie 93)). Ces tests permettent de juger de la pertinence ou du niveau d'approximation du modèle. Les simulations de la variable initiale sont obtenues dans une deuxième étape par transformation inverse des valeurs simulées de la variable transformée.

3.3.3 Grille et nombre de simulations

Outre le choix de l'algorithme de simulation et la transformation possible de données, la grille de simulation et le nombre de simulations doivent être définis.

- ***Grille de simulation.*** Les teneurs sont simulées aux nœuds d'une grille dont l'orientation, la position de l'origine et les espacements entre nœuds sont à fixer au départ. La grille doit recouvrir le domaine d'étude et être suffisamment fine pour que des teneurs de maille puissent être calculées à partir des simulations[11], que ce soit pour établir une classification des mailles (voir §3.4.4) ou pour dimensionner un tri sélectif en dépollution (voir §3.6). En pratique, la taille de la grille de simulation doit rester raisonnable (généralement inférieure à quelques centaines de milliers de nœuds) pour éviter des temps de simulation trop longs. A titre d'information, il est courant de prendre des espacements entre nœuds de 2,5 à 5 m dans le plan horizontal et de 0,25 à 0,5 m dans la direction verticale.
- ***Nombre de simulations.*** Ce nombre doit être suffisant pour que les réalisations produites couvrent et reflètent l'espace de variabilité des teneurs à l'échelle du site. De ces réalisations dépend en effet la représentativité des incertitudes qui en sont déduites. Dans le même temps, l'ensemble des simulations doit pouvoir être généré dans des délais raisonnables. Un nombre de quelques centaines de réalisations est souvent considéré comme satisfaisant.

Le calcul de l'histogramme et du variogramme sur quelques simulations permet finalement de vérifier la cohérences des données entre elles et avec les modèles statistiques choisis.

3.4 Estimation des zones et des volumes contaminés

3.4.1 Calcul de teneurs de maille

Les volumes à dépolluer sont définis sur un maillage de dépollution dont la taille des mailles détermine le support de calcul des teneurs de maille. Les critères de choix du maillage de dépollution sortent du cadre de ce manuel. Nous supposons simplement que ce maillage est fixé, en accord avec le maître d'ouvrage et le prestataire de dépollution, et nous souhaitons calculer des teneurs de maille à partir des simulations conditionnelles (réalisations) obtenues à l'étape précédente.

[11] Les teneurs simulées étant représentatives de l'échelle des échantillons, c'est-à-dire de celle des données.

La méthode la plus simple pour modéliser ce ***changement de support*** est la suivante. Les teneurs ayant été simulées aux nœuds d'une grille de simulation fine, chaque valeur simulée représente une teneur en polluant sur un petit support en rapport avec les échantillons de sols analysés. En superposant le maillage de dépollution sur cette grille fine, il devient possible de calculer la teneur moyenne vraie de chaque maille comme la moyenne des valeurs simulées contenues dans la maille (voir §3.6, Figure II-8). En répétant l'opération sur chaque réalisation des teneurs, différentes cartes de teneurs de maille peuvent être produites.

D'autres solutions pour modéliser le changement de support existent, telles que :

- *Simulation directe des teneurs moyennes dans les mailles d'excavation.*
- *Utilisation de modèles de changement de support, tel que le modèle gaussien discret (Rivoirard 1994).*

3.4.2 Application du seuil critique aux teneurs de maille

Les sols contaminés (à dépolluer) sont formés des sols provenant des mailles de la grille de dépollution où la teneur moyenne en polluant est supérieure au seuil critique. Ces mailles peuvent être identifiées sur chacune des cartes de teneurs de maille tirées des simulations conditionnelles sur la grille fine. Les variations d'une carte à l'autre de la position et des volumes de sols à dépolluer traduisent alors l'incertitude sur l'étendue de la pollution (voir §3.4.3).

A titre de comparaison, il peut être aussi utile d'appliquer le seuil critique directement sur la grille (fine) de simulation où les teneurs simulées sont représentatives d'une échelle d'observation beaucoup plus petite en rapport avec les échantillons de sols analysés. En comparant les volumes de sols à dépolluer ainsi calculés à ceux établis sur le maillage de dépollution, il est possible d'apprécier l'impact de l'effet de support sur ces volumes.

3.4.3 Analyse statistique sur les teneurs de maille et sur les volumes contaminés

L'analyse statistique des teneurs de maille peut être effectuée par maille sur toutes les simulations ou bien encore par simulation en prenant en compte toutes les mailles. Ces deux modes d'analyse renseignent sur deux types d'incertitude dont la signification diffère.

- ***Analyse par maille et incertitude sur les teneurs (incertitude locale).*** Les simulations des teneurs fournissent, pour chaque maille de dépollution, la distribution (histogramme) des valeurs possibles de la teneur moyenne en polluant dans la maille (Figure II-6). Cette distribution dite locale suffit à décrire entièrement l'incertitude sur l'état de contamination d'une maille pour un niveau de connaissance donné du site. Dans une perspective de dépollution, cette incertitude locale spécifique à chaque maille doit être prise en compte pour décider des sols à dépolluer (étape de classification) mais aussi pour juger de la pertinence de la classification en terme de risque de mauvais classement des mailles (étape d'aide à la décision).

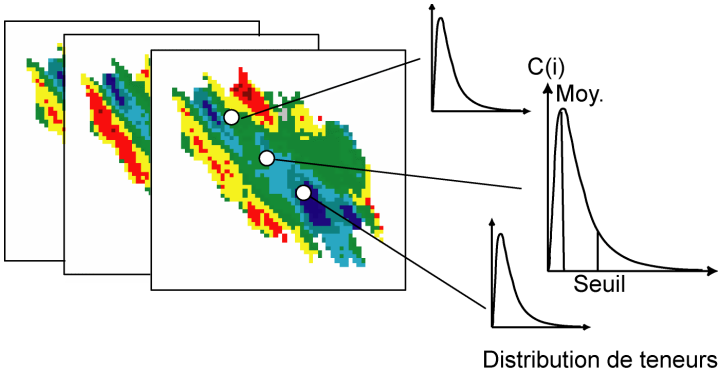


Figure II-6 Analyse statistique par maille et incertitude locale. Les simulations des teneurs sont analysées par maille. La dispersion de la distribution (histogramme) des teneurs simulées en chaque maille modélise l'incertitude locale (spécifique à chaque maille).

- ***Analyse par réalisation et incertitude sur les volumes (incertitude globale).*** La distribution (histogramme) des volumes de sols à dépolluer se déduit des volumes calculés sur les réalisations en appliquant le seuil critique aux teneurs de maille. La forme et l'étalement de la

distribution traduit alors l'incertitude sur ces volumes (Figure II-7). Cette incertitude est dite globale dans le sens où elle porte sur l'ensemble du domaine d'étude. Il est à noter que ces calculs ne permettent généralement pas une évaluation réaliste du volume qui sera finalement dépollué car ils ne prennent pas en compte les conditions de classification des sols et de gestion du chantier de dépollution.

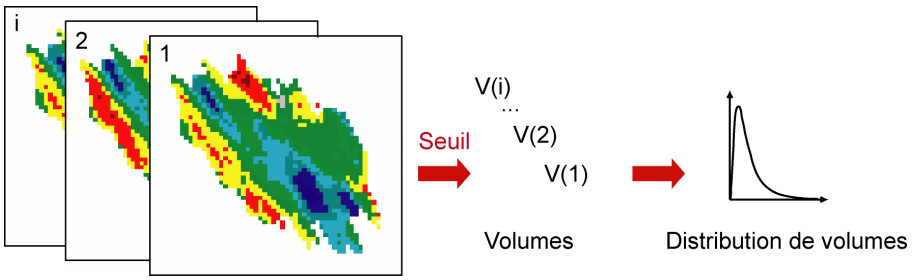


Figure II-7 Analyse statistique par simulation (réalisation) et incertitude globale. A chaque simulation, l'ensemble des mailles dont les teneurs sont supérieures au seuil constitue un volume global. La distribution (histogramme) des volumes ainsi calculés sur l'ensemble des simulations décrit l'incertitude sur le volume (réel) de sols à dépolluer.

3.4.4 Classification des mailles comme contaminées ou saines

L'application d'un seuil de teneur ne suffit pas à classer une maille comme saine ou contaminée. L'incertitude sur la teneur de maille, que les simulations conditionnelles retranscrivent (voir §3.4.3), se traduit en effet par une incertitude sur le dépassement ou non du seuil par la teneur de maille. **Les sols doivent donc être classés selon un critère qui prend en compte l'incertitude sur la teneur de maille**. Ce critère peut être purement statistique ou intégrer une fonction de coût d'erreur (ou fonction de perte) qui contribuera à pénaliser davantage certaines situations de mauvaise classification que d'autres. Le critère pourra aussi varier d'une zone à l'autre, en fonction du type de réaménagement prévu, ou d'un site à l'autre selon les spécificités propres à chacun. Quelques exemples de critères sont proposés dans la suite. Ils constituent généralement la formalisation de critères intuitifs déjà connus. En pratique, un dialogue avec les responsables du site est à favoriser pour arriver à une représentation correcte de leurs intentions à travers ces critères de classification.

Tous les critères de classification utilisent les distributions (locales) sur les teneurs de maille[12] (Figure II-6) pour attribuer à chaque maille une meilleure classe d'affectation dont la signification dépend du contenu du critère. Comme annoncé précédemment, deux types de critères peuvent être distingués.

- *Critères statistiques* : ils consistent à positionner un paramètre statistique particulier (ex. moyenne, médiane, autre quantile), relatif à la distribution (locale) de teneurs de maille, par rapport au seuil critique de dépollution.
- *Critères financiers* : ils consistent à minimiser le coût d'erreur de classification compte tenu de l'incertitude sur les teneurs de maille.

Seuls les critères statistiques sont utilisés actuellement dans les études. Trois critères de ce type sont décrits dans la suite.

Les critères financiers présentent quant à eux l'intérêt particulier de conduire à la classification dont les conséquences financières sont statistiquement les plus faibles, étant donnée l'incertitude sur les teneurs de maille. En outre, les coûts de classification et d'erreur devant être établis, il devient aussi possible de quantifier le coût de la réhabilitation et l'incertitude sur ce coût compte tenu du risque de mauvaise classification. Une introduction à ce type de critères se trouve au §3.5. **Leur utilisation dans les études nécessiterait à présent d'arriver à un consensus sur la définition des coûts**

[12] Rappelons que la distribution locale reflète l'incertitude sur la teneur vraie mais inconnue de la maille. Elle nous dit quelles valeurs peut prendre la teneur de maille et avec quelles probabilités compte tenu des données disponibles et des modèles statistiques établis pour décrire le phénomène.

d'erreur de classification et plus précisément des coûts de mauvaise classification des sols comme sains.

Critère statistique sur la teneur de maille moyenne

Les valeurs possibles de la teneur de maille sont moyennées et la maille est classée comme contaminée si la moyenne est supérieure au seuil critique[13]. En pratique, ce critère est à déconseiller car il est très sensible aux valeurs de teneur extrêmes et tend à classer à tort des mailles saines comme polluées ou inversement des mailles polluées comme saines.

Critère statistique de plus grande probabilité

Il est calculé pour chaque maille la probabilité que la teneur de maille soit supérieure au seuil critique. La probabilité est calculée comme le rapport du nombre de réalisations où la teneur de maille est supérieure au seuil sur le nombre total de réalisations. Si cette probabilité est supérieure à 0.5, la maille est classée comme contaminée car il y a plus de chances que la teneur de maille dépasse le seuil critique. La maille est sinon considérée comme saine.

Ce critère peut être généralisé à une probabilité de référence quelconque P_{ref}, la maille étant classée comme contaminée si la probabilité de dépassement du seuil de teneur critique est supérieure à P_{ref}, ce qui peut s'écrire mathématiquement Prob{*teneur* > *seuil critique*} > P_{ref}. Choisir P_{ref} inférieure à 0.5 revient alors à favoriser dans certains cas, tant que la probabilité de dépassement du seuil ne descend pas au-dessous de P_{ref}, la dépollution de mailles qui ont plus de chances d'être saines que contaminées. Exprimé autrement, le choix de P_{ref} inférieure à 0.5 signifie pénaliser la non dépollution de mailles qui pourraient être polluées.

Critère statistique de plus grande probabilité avec plusieurs niveaux de contamination

Le critère précédent peut être appliqué au cas où plusieurs niveaux de contamination associés à des seuils critiques différents doivent être reconnus. De la même façon que précédemment, il est calculé à partir des réalisations la probabilité de dépassement de chacun des seuils rangés par ordre croissant. Notons N le nombre de seuils critiques et P_i la probabilité de dépassement du $i^{ème}$ seuil critique. Si P_{ref} est inférieur à toutes les probabilités de dépassement, et donc $P_{ref} < P_1$, la maille est classée comme saine. Elle est sinon affectée à la classe de pollution correspondant à la plus grande probabilité P_i inférieure à P_{ref}, c'est-à-dire à la classe i telle que $P_i \leq P_{ref} < P_{i+1}$ si $i < N$.

A ce stade, toutes les mailles sont affectées d'une classe et l'étude peut s'arrêter pour laisser place à la dépollution. En pratique, se pose alors la question de la confiance qui peut être accordée à la classification. Cette confiance dépend directement des erreurs de classification qui peuvent être commises compte tenu de l'incertitude sur les teneurs de maille. Remettre en question à certains endroits la classification n'a de sens que s'il est possible de préconiser des actions à mener en fonction du risque d'erreur ou du coût de ces erreurs. La démarche géostatistique doit ainsi conduire à la production d'éléments d'aide à décision. Le paragraphe suivant en présente le principe dans ses grandes lignes.

3.4.5 Validation de la classification et critères de décision

Pour juger de la validité de la classification des sols, les incertitudes sur les teneurs de maille doivent être remises en face des classes affectées aux mailles. Selon les critères de classification employés, cette vérification peut prendre au moins deux formes.

- Evaluation statistique de la probabilité d'erreur. S'applique à toutes les situations.
- Evaluation du coût d'erreur de classification. S'applique aux situations où les critères de classification sont d'ordre financier (voir §3.5).

Dans tous les cas, les situations possibles doivent être clairement reconnues et être associées à des décisions opérationnelles dans le cadre de la dépollution du site. Les possibilités sont grandes et nous

[13] Un résultat semblable est obtenu en estimant directement par krigeage la teneur de maille moyenne.

nous limiterons dans la suite à discuter et fournir des pistes dans le cas d'une évaluation statistique de la classification.

L'approche la plus simple consiste à décider de dépolluer toutes les mailles classées comme contaminées, avec ou sans tri systématique. Cette approche présente cependant plusieurs limitations.

- Si un tri sélectif systématique est opéré, des mailles, dont nous pouvons savoir qu'elles sont très certainement contaminées, seront inutilement échantillonnées et évaluées à nouveau.
- Si aucun tri sélectif n'est prévu, des mailles, dont nous pouvons savoir qu'elles ont des chances non négligeables de ne pas être contaminées, seront traitées à tort.
- Dans les deux cas, des mailles classées comme saines, mais dont nous pouvons savoir qu'elles ont des chances significatives d'être contaminées, seront laissées en place.

Dans ces conditions, le principe de l'aide à la décision est de faire le bilan des incertitudes (ou des coûts d'erreur) sur la classification des mailles pour en tirer un ensemble de résultats utiles à l'organisation de la dépollution du site avec des préconisations sur les actions à mener.

Concrètement et à titre d'exemple, l'incertitude sur la classification peut conduire à identifier les **trois types de situation suivants.**

- Situation 1 : maille contaminée ou saine très certainement bien classée (faible risque d'erreur).
- Situation 2 : maille classée comme contaminée à confirmer (risque non négligeable d'erreur).
- Situation 3 : maille classée comme saine à confirmer (risque également non négligeable d'erreur)

A ces trois situations, **les préconisations suivantes peuvent être proposées comme critères de décision.**

- Situation 1 : dépollution systématique des mailles classées comme contaminées, non excavation des mailles classées comme saines.
- Situation 2 : tri sélectif pour confirmer ou non l'état de contamination et décider en dépollution du devenir de ces mailles.
- Situation 3 : nécessité d'une vérification complémentaire soit en intégrant ces mailles dans le processus de tri sélectif soit par une reconnaissance complémentaire ciblée.
- Situations 2 et 3 : si les délais le permettent et si les mailles incertaines constituent des zones de grands volumes et donc de coûts de traitement importants, des investigations complémentaires peuvent être décidées avant dépollution pour préciser dans tout ou partie de ces zones la classification des mailles.

Que le tri sélectif ou la reconnaissance complémentaire se fasse par mesures sur site ou en laboratoire, il est impératif que le mode d'évaluation de l'état de contamination des sols (en particulier les tailles de maille) soit comparable à celui utilisé dans les phases précédentes pour quantifier les volumes de sols à dépolluer.

3.4.6 Localisation des zones et évaluation des volumes contaminés

Par application de critères de décision tels que définis précédemment, la carte de classification des mailles est transformée en une carte de positionnement des zones de tri et de dépollution.

L'addition des volumes unitaires des mailles correspondant à chaque critère de décision pour chacune des simulations fournit de plus une distribution de volumes par type de réaménagement et par filière de dépollution. Selon le contexte de la réhabilitation, les valeurs retenues pour caractériser ce volume peuvent être une valeur centrale (médiane ou moyenne), l'écart-type de la distribution ou bien un intervalle interquartile ou autre.

3.5 Prise en compte de critères financiers dans la classification des sols

L'utilisation de critères financiers a pour but de minimiser le coût d'erreur de classification en affectant à chaque maille la classe qui conduira plus probablement aux conséquences financières les moins importantes compte tenu de l'incertitude sur les teneurs.

Pour définir ces critères, un coût doit être associé à chacune des situations de classification possible. Nous nous limitons ici à deux classes avec pour chaque maille la possibilité qu'elle soit saine ou polluée. Les situations de classification, au nombre de quatre, sont alors les suivantes.

- Maille polluée classée à raison comme polluée.
- Maille polluée classée à tort comme saine.
- Maille saine classée à raison comme saine.
- Maille saine classée à tort comme polluée.

La classe affectée à la maille est alors celle qui minimise le coût de classification ou d'erreur[14]. En pratique, les situations de mauvaise classification sont supposées entraîner des coûts plus élevés que celles de classification correcte. Dans ces conditions, seuls comptent les coûts de mauvaise classification.

A titre d'exemple, une définition des coûts est proposée ci-après pour chacune des situations de classification précédente. Le coût de chaque situation est désigné par la lettre C.

- Classification à raison d'une maille comme saine : $C = 0$ (aucun frais).
- Classification à raison d'une maille comme contaminée : $C = C_D$ (coût de dépollution).
- Classification à tort d'une maille comme contaminée : $C = C_D$ (coût de dépollution – inutile – de la maille).
- Classification à tort d'une maille comme saine : ce coût est très certainement le plus difficile à évaluer. Laisser en place des sols contaminés peut en effet avoir plusieurs conséquences.
  - Conséquences sanitaires et environnementales si le site est réaménagé sans que cette pollution ne soit jamais détectée et qu'elle finisse par atteindre la population ou l'écosystème. L'évaluation du coût d'une telle conséquence revient à des experts de ce domaine en prenant en compte le facteur temps.
  - Conséquences de chantier si les sols contaminés sont découverts au moment de la dépollution (par ex. mis en évidence par un tri sélectif). Une extension plus grande de la pollution signifie au mieux des délais et donc des coûts de chantier plus grands auxquels peuvent se rajouter des pénalités financières. Ce coût peut donc s'exprimer comme un facteur multiplicatif sur le coût de dépollution de base :

    $$C = C_D\,(1+P) \text{ où } P \geq 0 \text{ est un facteur de pénalité globale.}$$

  - Conséquences de retard dans le projet de réaménagement si la découverte tardive de ces sols, au mieux pendant les travaux, au pire après lors du contrôle de la dépollution résiduelle, entraîne un report dans le temps du réaménagement du site mais aussi un diagnostic complémentaire et la réalisation d'un nouveau chantier de dépollution. Les coûts financiers sont alors multiples :
    - Coûts de diagnostic et travaux complémentaires : C_T
    - Manque à gagner pendant la période de report si une activité industrielle ou commerciale est prévue sur le site : C_T
    - Intérêt financier que représente pour le vendeur le prix du site sur la période de report : C_F

[14] Ce critère revient à prendre un quantile particulier dans la distribution des valeurs possibles de la teneur. Ce quantile dépend alors des coûts associés aux différentes situations de classification.

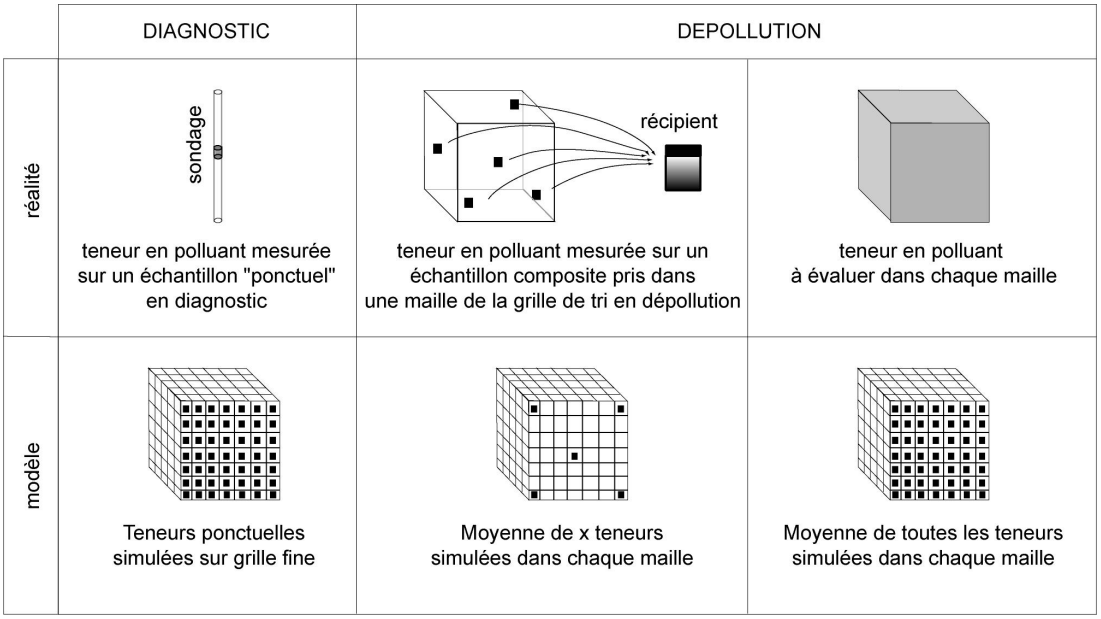
Une fois les coûts établis, il reste à déterminer comment les appliquer individuellement à chaque maille. La solution la plus simple est de répartir ce coût sur les mailles concernées qui peuvent être soit l'ensemble des mailles du site soit plus judicieusement celles dont la classification est incertaine. Connaissant la distribution locale de la teneur pour chaque maille, peuvent rentrer dans ce décompte les mailles pour lesquelles on ne peut pas établir avec une probabilité suffisamment forte (généralement au moins supérieure à 80 % de chances) qu'elles sont (certainement) saines ou contaminées. Si on note N_i le nombre de ces mailles, le coût s'exprime comme : $C = C_D + (C_T + C_R + C_F) / N_i$.

3.6 Dimensionnement d'un tri sélectif en dépollution

La mise en place d'un tri sélectif suppose de définir un mode d'échantillonnage des mailles dont l'état de contamination est incertain et à préciser davantage au moment de la dépollution. Le nombre et l'emplacement des échantillons dépendent des dimensions des mailles et de la variabilité spatiale des teneurs à l'intérieur de ces mailles. Les échantillons prélevés dans chaque maille sont regroupés pour constituer un échantillon composite qui est ensuite analysé. L'écart entre la teneur de maille estimée à partir de l'échantillon composite et la teneur vraie de la maille est connu sous le nom d'**effet d'information**. Cet écart pouvant entraîner un tri erroné des mailles, il n'est pas sans conséquences sur les volumes de sols qui sont dépollués et orientés à raison ou à tort vers les filières de traitement. Son importance peut être apprécié de la façon suivante (Figure II-8) en répétant si nécessaire la procédure sur différentes réalisations des teneurs dans les sols (simulations sur la grille fine).

- Calculer les teneurs de maille supposées vraies à partir des valeurs simulées sur la grille fine (principe de calcul décrit au §3.4.1).
- Simuler la prise d'un échantillon composite dans chaque maille en échantillonnant la grille fine et en déduire la teneur à partir des teneurs simulées sur la grille fine.
- Comparer les volumes ainsi calculés.

En répétant le processus pour différents nombres d'échantillons ou schémas d'échantillonnage des mailles, un dimensionnement convenable du processus de tri sélectif peut être obtenu.



D'après Demougeot-Renard and De Fouquet C., 2004, « A geostatistical approach to assessing volumes of soil requiring remediation. Validation to lead-polluted soils underlying a former smelting works", Environmental Science and Technology, vol. 38, n°19, pp. 5120-5126.

Figure II-8 Le changement de support entre données de diagnostic et données de dépollution est modélisé en calculant, dans chaque maille de la grille de dépollution, la moyenne des x valeurs simulées sur grille fine de même localisation que les échantillons prélevés réellement dans la maille pour former un composite. L'effet d'information est évalué en comparant cette valeur à la valeur moyenne de toutes les valeurs simulées sur grille fine dans la maille, constituant une valeur approchée de la teneur de maille vraie.

3.7 Evaluation financière du projet de dépollution

Sans entrer dans le détail, l'approche géostatistique permet aussi de traduire les cartes de positionnement des zones à dépolluer et les estimations stochastiques des volumes correspondant en des données financières sur le coût global d'un projet de dépollution. Une illustration en est donnée à la Figure II-9.

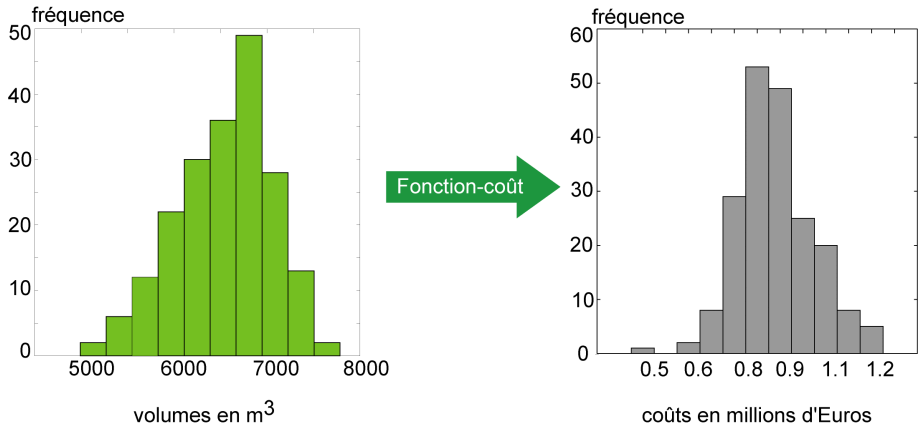


Figure II-9 Fonction de coût faisant la somme des coûts des différentes phase d'un chantier de dépollution (préparation du chantier, amené-repli du matériel, suivi et contrôle des travaux, tri des sols, excavation et transport sur site, criblage, transport hors site, traitement, stockage des sains, remise en état et comblement de la fosse) et appliquée à la distribution des volume de sol à dépolluer pour calculer les coûts de dépollution par lavage des sols d'une ancienne fonderie. On en déduit une distribution de coûts de dépollution (Tableau II-2), dont la moyenne est égale à 0.90 millions € et le coefficient de variation (rapport entre l'écart-type et la moyenne) est égal à 14 %.

Tableau II-2 Statistiques élémentaires des coûts de dépollution calculés pour chaque simulation.

		Unités
Nombre de simulations	200	-
Médiane	0.89	Millions €
Moyenne	0.90	Millions €
Ecart-type	0.13	Millions €
Coefficient de variation	0.14	-
Minimum	0.47	Millions €
Maximum	1.22	Millions €

4 ILLUSTRATION

Le site sélectionné présente une pollution multiple (6 polluants) sur un domaine de grande ampleur (43 ha, pollution jusqu'à 20 m de profondeur), sur lequel deux types de réaménagement sont envisagés : zone résidentielle et zone industrielle. Même s'il est exceptionnel par sa taille et la complexité de sa contamination, ce site est particulièrement intéressant car il permet de mettre en œuvre de nombreux aspects de la démarche géostatistique présentée précédemment. Notons que tout ou partie de la démarche a déjà été appliqué à d'autres sites dont les caractéristiques sont plus courantes, notamment le nombre de données disponibles (de 12 à 145 données par ha pour les cas d'étude traités par les membres experts de ***GeoSiPol***).

4.1 Contexte

Les campagnes de reconnaissance menées sur un ancien centre de production de gaz italien ont mis en évidence des teneurs élevées en quatre substances, supérieures aux valeurs réglementaires nationales : hydrocarbures aliphatiques renfermant plus de 12 atomes de Carbone ($HC_{>C12}$) et métaux lourds As, Cu et Pb. Le sous-sol est formé de remblais, graviers alluvionnaires, sables et argiles. Aucune nappe n'a été rencontrée jusqu'à 60 m de profondeur. Près de 900 échantillons de sols ont été analysés en laboratoire mais toutes les substances n'ont pas été mesurées sur tous les échantillons (Tableau II-3).

Les zones résidentielles et industrielles prévues dans le projet de réaménagement du site sont localisées sur le plan de la Figure II-10. Les objectifs de dépollution sont plus sévères dans les zones résidentielles que dans la zone industrielle. Les seuils de risque par type de réaménagement pour les quatre substances sont présentées au Tableau II-4. Dans toutes ces zones, les sols doivent être traités si l'un au moins des polluants présente une teneur supérieure au seuil de risque correspondant au type de réaménagement prévu pour la zone.

L'objectif de l'étude géostatistique est de localiser et d'évaluer les volumes de sols à traiter (dépolluer) dans chacun des deux types de zones, résidentiel et industriel.

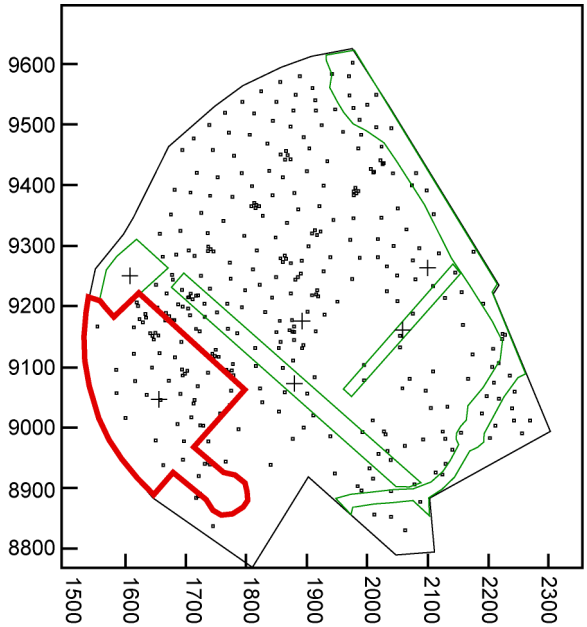


Figure II-10 Localisation de la zone industrielle (rouge), des axes routiers (vert) et des zones résidentielles (autres) sur le plan du site, ainsi que des points d'échantillonnage (points noirs).

Tableau II-3 Résumé statistique sur les teneurs en polluants mesurées sur le site de l'ancien centre de production de gaz (unité : mg/kg ms). Données corrigées du biais d'échantillonnage (dégroupées).

	Nombre	Minimum	Maximum	Moyenne	Ecart-type	Coefficient de variation
As	873	0.00	986.27	11.36	41.11	3.62
Cu	852	0.00	3201.00	15.75	73.68	4.68
Pb	874	0.00	7525.00	32.99	219.72	6.66
$HC_{>C12}$	225	0.00	6000.00	88.38	354.53	4.01

Tableau II-4 Seuils de risque (en mg/kg ms) sur les teneurs par polluant et par type de réaménagement.

	As	Cu	Pb	$HC_{>C12}$
Seuil en zone résidentielle ($S_{Résid}$)	20	120	100	50
Seuil en zone industrielle (S_{Ind})	50	600	1000	750

4.2 Formalisation des critères de classification et décision

Nous désignons par :

- C(i) la teneur en contaminant mesurée i (i=1, 2, …, 4).
- $S_{Résid}(i)$ le seuil de risque du contaminant i à partir duquel un traitement des sols est nécessaire pour un réaménagement en zone résidentielle, c'est-à-dire traitement si $C(i) \geq S_{Résid}(i)$.
- $S_{Ind}(i) > S_{Résid}(i)$ le seuil de risque pour un usage industriel, c'est-à-dire traitement si $C(i) \geq S_{Ind}(i)$.

4.2.1 Mailles de tri en dépollution

Les responsables du site prévoient d'excaver les sols par mailles de dimensions 20m x 20m x 0.5m. Les conditions de tri des sols par maille n'étant pas connues à ce stade, les calculs ont été effectués en modélisant directement les teneurs moyennes de maille.

4.2.2 Critère de classification dans les zones résidentielles

Dans les zones résidentielles, les sols doivent être traités si la teneur de l'un au moins des quatre polluants est supérieure à son seuil de risque. Les sols sont donc classés comme contaminés si :

$$C(1) \geq S_{Résid}(1) \text{ ou } C(2) \geq S_{Résid}(2) \ldots \text{ou } C(4) \geq S_{Résid}(4)$$

Compte tenu de l'incertitude sur les teneurs en polluants **en un endroit donné** du site, il est regardé la probabilité que ce critère soit vérifié. Nous appelons **probabilité globale de dépassement** des seuils de risque $S_{Résid}(i)$:

$$Prob_{Résid} = Prob\{ C(1) \geq S_{Résid}(1) \text{ ou } C(2) \geq S_{Résid}(2) \ldots \text{ou } C(4) \geq S_{Résid}(4) \}$$

Cette probabilité globale est calculée à partir des probabilités individuelles de dépassement des seuils des quatre substances, en supposant que les quatre variables « teneurs en polluant » sont indépendantes (hypothèse conservatrice) :

$$Prob_{Résid} = 1 – Prob\{C(1)< S_{Résid}(1)\}.Prob\{C(2)< S_{Résid}(2)\}…Prob\{C(4)< S_{Résid}(4)\}$$

$$Prob_{Résid} = 1 – Prob\{As<20\}.Prob\{Cu<120\}.Prob\{Pb<100\}.Prob\{HC_{>C12}<50\}$$

En l'absence de recommandation de la part des responsables du site, un critère statistique de plus grande probabilité (voir p. 68) a été proposé et finalement choisi comme critère de classification des sols comme à traiter ou non dans les zones résidentielles. Ce critère s'exprime comme suit.

- Si $Prob_{Résid} \geq 0.5$ (50% de chances de dépassement) : excavation et dépollution des sols (maille).
- Sinon : sols laissés en place sans aucun traitement.

4.2.3 Critère de classification dans la zone industrielle

En raison de contraintes de construction, la zone industrielle doit être excavée sur toute sa surface jusqu'à une profondeur de 12 m. Dans cette zone, sont à dépolluer les mailles où la teneur en au moins un des quatre polluants est supérieure à son seuil de risque.

Comme précédemment pour prendre en compte l'incertitude sur les teneurs, nous nous intéressons à la probabilité globale de dépassement des seuils de teneur $S_{Ind}(i)$, en supposant toujours que les quatre substances sont indépendantes. Si nous notons $Prob_{Ind}$ cette probabilité, elle s'écrit :

$$Prob_{Ind} = 1 – Prob\{C(1)< S_{Ind}(1)\}.Prob\{C(2)< S_{Ind}(2)\}…Prob\{C(4)< S_{Ind}(4)\}$$

$$Prob_{Ind} = 1 - Prob\{As<50\}.Prob\{Cu<600\}.Prob\{Pb<1000\}.Prob\{HC_{>C12}<750\}$$

De la même façon que pour les zones résidentielles, le critère de classification retenu est un critère statistique de plus grande probabilité qui devient ici :

- Si $Prob_{Ind} \geq 0.5$ (50% de chances de dépassement) : excavation et dépollution des sols (maille).
- Sinon : excavation mais non dépollution des sols (maille).

4.2.4 Critères de décision

La classification étant elle-même incertaine, l'étape suivante consiste à établir des critères de décision pour décider d'actions à mener en fonction du risque de mauvaise classification des mailles. En accord avec les responsables du site, il a été jugé acceptable de ne pas remettre en cause la classification d'une maille lorsque le risque d'erreur est inférieur à 20% de chances de se tromper. Lorsque ce risque est supérieur à 20%, un tri systématique des sols est préconisé, que la maille soit classée à dépolluer (à traiter) ou non.

Avec ces choix, les critères de décision peuvent s'énoncer comme suit.

Dans les zones résidentielles

- Si risque de mauvaise classification > 20% de chances de se tromper : tri systématique avec dépollution des sols si confirmation du dépassement du seuil de risque pour au moins un polluant ($C(i) \geq S_{Ind}(i)$).
- Sinon : dépollution des sols sans autre vérification si maille classée comme à traiter ; sols laissés en place sinon.

Dans la zone industrielle

- Si risque de mauvaise classification > 20% de chances de se tromper : excavation et tri systématique avec dépollution des sols si confirmation du dépassement du seuil de risque pour au moins un polluant ($C(i) \geq S_{Résid}(i)$).
- Sinon : excavation des mailles et dépollution de celles classées comme à traiter.

4.3 Résultats

4.3.1 Calculs géostatistiques

Les probabilités globales $Prob_{Résid}$ et $Prob_{Ind}$ nécessitent le calcul des probabilités individuelles de dépassement des seuils pour les quatre substances $HC_{>C12}$, As, Cu et Pb. Un ensemble de 100 simulations conditionnelles a été généré pour chaque substance. Ces simulations retrouvent les données de teneurs mesurées sur échantillons de sols et reproduisent les modèles statistiques tirés de ces données (histogramme, variogramme et tendances spatiales pour chaque substance).

La grille de simulation est une grille fine (mailles 2 m x 2 m x 0.5 m) qui permet de modéliser le changement de support entre les échantillons analysés en laboratoire et les mailles de dépollution. Chaque valeur de maille de la grille de dépollution (maille 20 m x 20 m x 0.5 m) est calculée comme la moyenne des valeurs simulées aux nœuds (de la grille fine) situés à l'intérieur de la maille. La hauteur des mailles de 0.5 m est identique pour les deux grilles : elle est conforme aux longueurs d'échantillons analysés au laboratoire et à l'épaisseur des couches excavées en dépollution. L'effet d'information a été négligé dans cette étude (voir §2.5).

Pour chacune des quatre substances potentiellement polluantes, 100 cartes de teneurs de maille ont été calculées à partir des 100 simulations conditionnelles des teneurs sur la grille fine. De ces cartes ont ensuite été déduites pour chaque maille les probabilités suivantes.

- Probabilité de dépassement du seuil de risque $S_{Résid}(i)$ ou $S_{Ind}(i)$ de chaque substance, comme le rapport entre le nombre de teneurs de maille calculées supérieures au seuil et le nombre total de valeurs calculées (c.-à-d. 100).
- Probabilité globale de dépassement des seuils de risque ($Prob_{Résid}$ ou $Prob_{Ind}$).

4.3.2 Classification des mailles de la grille de dépollution

Connaissant les probabilités globales de dépassement des seuils de risque, les critères de classification définis plus haut (voir §4.2.2 et 4.2.3) peuvent s'appliquer à la grille de dépollution. La Figure II-11 montre les cartes de classification obtenues dans les zones résidentielles et dans la zone industrielle entre 0 et 0.5 m de profondeur. La proportion de mailles à dépolluer y apparaît beaucoup plus importante dans les zones résidentielles où les seuils de risque sur les teneurs sont plus sévères (seuils plus faibles).

Par référence aux critères de classification choisis, ces cartes nous indiquent quelle « meilleure décision » serait à prendre s'il devait être décidé immédiatement du sort de chaque maille. Cette classification demande cependant à être validée en tenant compte du risque (probabilité) de mauvaise classification des mailles. Cela fait l'objet du paragraphe suivant.

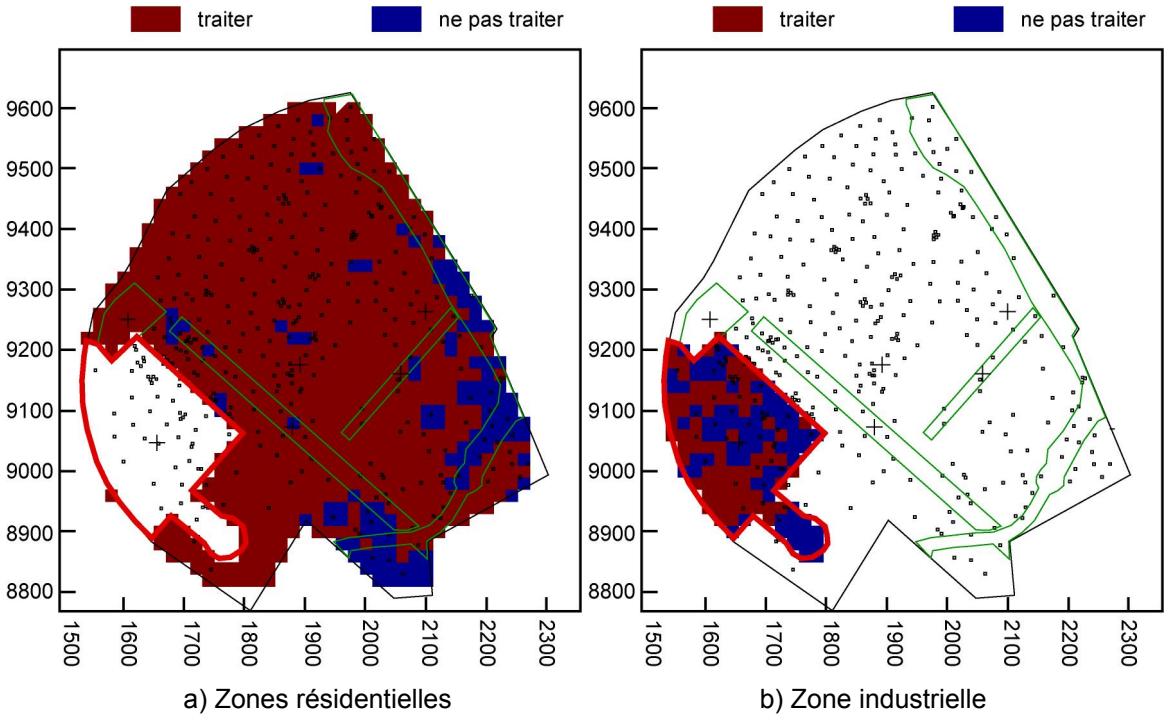


Figure II-11 Cartes 2D de classification des sols dans les différentes zones avec en rouge les mailles classées comme à traiter. Ces cartes portent sur les sols situés entre 0 et 0.5 m de profondeur sachant que la classification varie peu avec la profondeur (forte continuité verticale des teneurs).

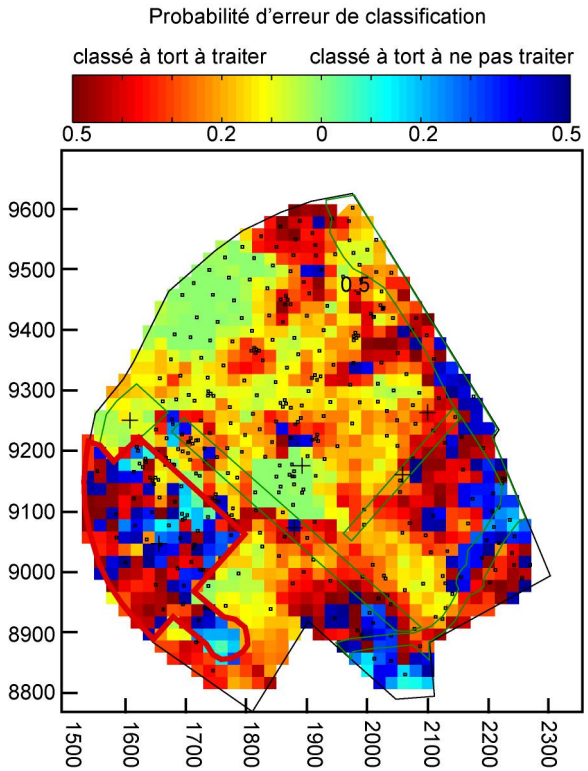


Figure II-12 Carte de la probabilité de mauvaise classification des mailles avec dans les tons de rouge à vert les mailles classées initialement comme à traiter et dans les tons de bleu celles à ne pas traiter.

4.3.3 Validation de la classification et application de critères de décision

La confiance qui peut être donnée aux cartes de classification des mailles dépend de la probabilité de mauvaise classification de ces mailles. Ces probabilités d'erreur sont illustrées à la Figure II-12 pour l'ensemble des zones en nous intéressant toujours aux sols situés entre 0 et 0.5 m de profondeur.

Ces probabilités sont utilisées ici pour appliquer les critères de décision définis plus haut (§4.2.4) et qui déterminent les actions à mener en fonction du risque (probabilité) de mauvaise classification des mailles que le détenteur du site est prêt à prendre. La Figure II-13 montre les volumes de sols qui seraient à traiter ou non en fonction du choix qui est fait sur la probabilité d'erreur de classification. Les volumes incertains sont ceux pour lesquels il est préconisé un tri systématique. Plus la probabilité d'erreur est faible, plus le volume de sols incertains est grand car ne sont acceptées comme à traiter ou non que les mailles qui présentent un faible risque de mauvaise classification. La ligne en rouge marque le niveau de risque finalement retenu pour ce site. Il correspond aux volumes de sols suivants.

Zone industrielle

Sols reconnus avec confiance comme à ne pas traiter	130 000 m^3
Sols reconnus avec confiance comme à traiter	18 000 m^3
Sols nécessitant un tri systématique en dépollution	580 000 m^3

Zones résidentielles

Sols reconnus avec confiance comme à ne pas traiter	170 000 m^3
Sols reconnus avec confiance comme à traiter	2 060 000 m^3
Sols nécessitant un tri systématique en dépollution	5 700 000 m^3

L'emplacement de ces volumes est illustré à la Figure II-14 qui montre les actions à mener pour les sols situés entre 0 et 0.5 m de profondeur.

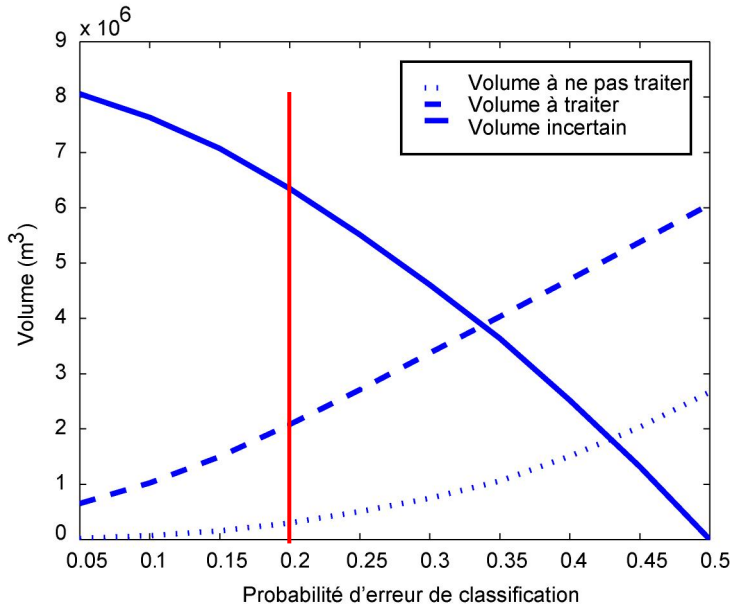


Figure II-13 Volumes de sols à ne pas traiter, à traiter et incertain en fonction de la probabilité d'erreur de classification. La ligne en rouge indique le risque (probabilité) de mauvaise classification que le responsable du site serait prêt à prendre pour décider de traiter ou non des mailles.

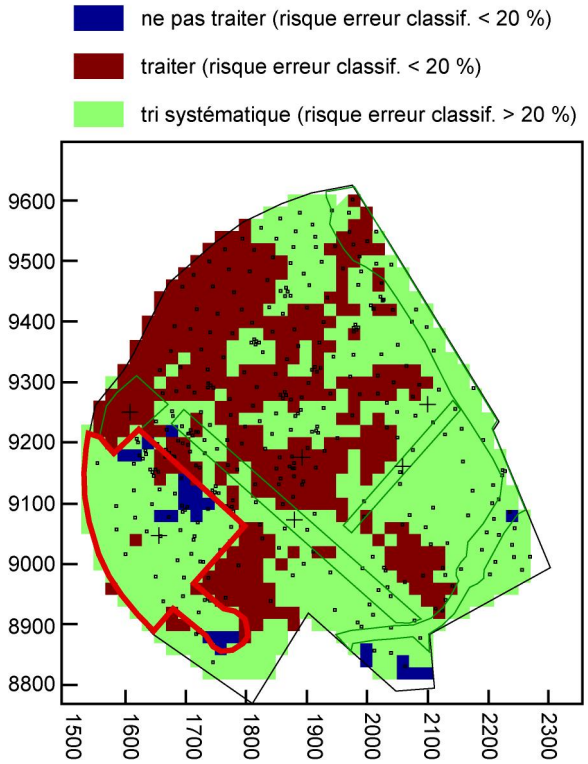


Figure II-14 Carte de décision avec les actions à mener pour les sols situés entre 0 et 0.5 m de profondeur compte tenu des critères de décision retenus.

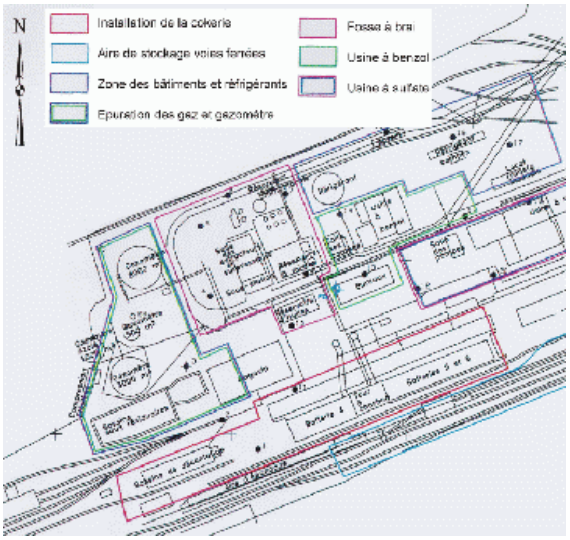
III THÈME N°3 : COMMENT INTEGRER TOUTE L'INFORMATION, DE L'HISTORIQUE AUX ANALYSES SUR SITE ?

Comment intégrer toute l'information, de l'historique aux analyses sur site ?

Fiche technique N°3

NOUS SOMMES TOUS CONFRONTES A CE PROBLEME

A chaque étape de la caractérisation d'un site potentiellement contaminé (diagnostic initial et Evaluation Simplifiée des Risques, diagnostic approfondi et Evaluation Détaillée des Risques, au cours d'un chantier de dépollution), **différents types d'information** sont disponibles sur l'état de contamination d'un site :



- **Informations historiques** quant à l'origine de la contamination, permettant une mise en évidence partielle ou exhaustive des zones a priori contaminées.
- **Lithologie** (coupes de sondage, données géophysiques renseignant sur les types de sols).
- **Information organoleptiques** : présence / absence d'indicateurs de contamination (visuels ou autres).
- **Analyses de terrain** : analyses par fluorescence X, kits chimiques, indicateurs colorimétriques, analyses de gaz FID et PID (détecteurs à ionisation de flamme et photo-ionisation).
- **Indices globalisants** analysés en laboratoire : somme des 16 HAP (Hydrocarbures Aromatiques Polycycliques), indice HC (Hydrocarbures).

Toutes ces informations complètent les résultats d'analyses en laboratoire dont le nombre reste toujours limité.

Se pose alors la question de « comment intégrer de manière rigoureuse toute l'information disponible dans la cartographie des teneurs en polluants et affiner ainsi la quantification des volumes contaminés ? ».

LES PRATIQUES ACTUELLES

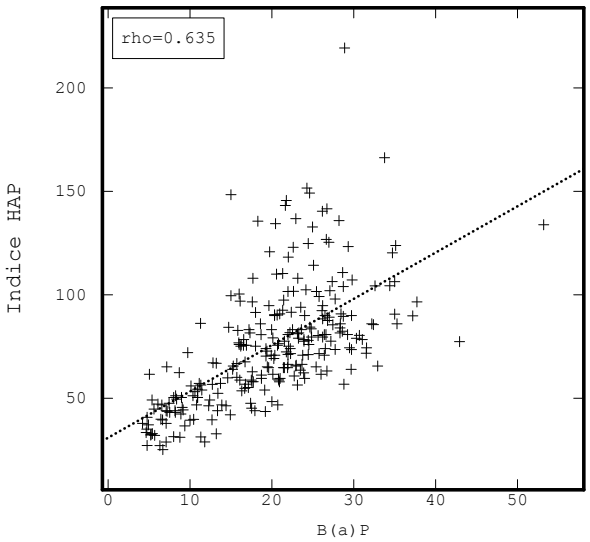
Les informations autres que les analyses en laboratoire présentent des avantages de coût, de temps de réalisation mais aussi de couverture plus grande des sites. **Même si ces informations indirectes et précieuses sont collectées sur la plupart des sites pollués, elles sont souvent « laissées de côté » ou bien prises en compte de manière peu rigoureuse** :

- A priori non justifié ni forcément justifiable de la relation possible entre l'état de contamination des sols et ces informations.
- Utilisation exclusive de ces informations pour décider des zones à échantillonner.
- Absence d'exploitation de l'apport complémentaire de ces informations à des fins de cartographie ou d'estimation.

CE QUE NOUS APPORTE LA GEOSTATISTIQUE

Par rapport à ces pratiques, l'approche géostatistique **permet d'intégrer de différentes manières les données indirectes sur les teneurs dans les sols** : corrélations, tendances, conformité à des indicateurs de pollution, influences lithologiques, etc. En fonction de l'apport d'information de ces données sur la pollution, un poids plus ou moins élevé leur est attribué dans l'estimation des teneurs.

L'information indirecte peut être de **nature qualitative**, telle que la lithologie, **semi-quantitative**, telle que les résultats d'analyses sur site, ou bien encore **quantitative**, telles que les teneurs en matière organique du sol ou les teneurs en un autre polluant.



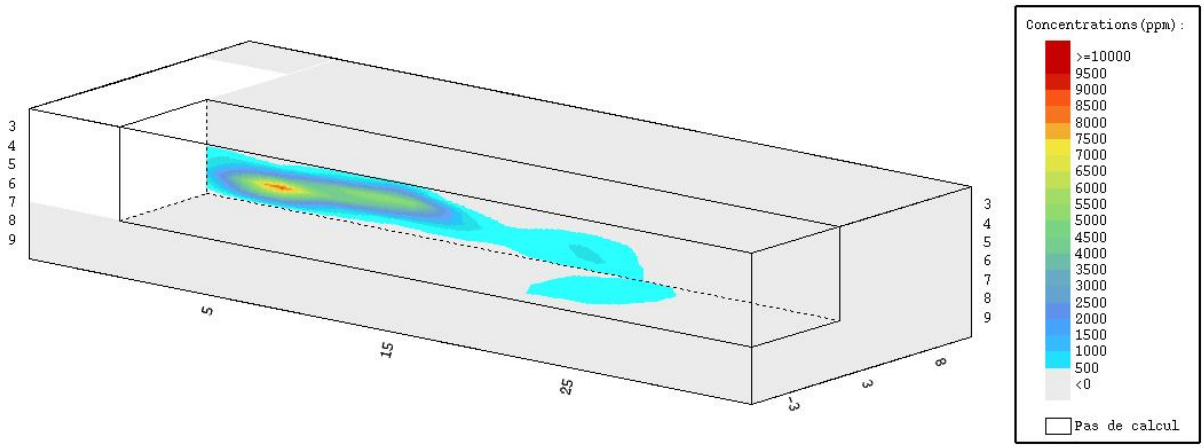
L'intérêt est donc de pouvoir introduire dans la modélisation de teneurs en polluants toute information indirecte corrélée ou en rapport avec les teneurs, quelle que soit la nature et la précision de cette information.

Au-delà du gain de précision sur l'estimation de teneurs, l'intégration d'informations indirectes signifie des modèles de pollution cohérents avec un ensemble de données riches et complémentaires, et par conséquent une meilleure délimitation des volumes contaminés.

Pour en savoir plus…

Concernant ce thème, le manuel méthodologique ***GeoSiPol*** présente une approche géostatistique standard pour exploiter les informations indirectes dans la cartographie des teneurs en polluants.

L'approche est illustrée sur le cas d'une contamination aux hydrocarbures, où la cartographie des teneurs en polluants a été réalisée à partir d'analyses en hydrocarbures totaux issues de sondages (variable d'intérêt), et d'informations organoleptiques et lithologiques (informations indirectes), moins précises mais peu coûteuses et par conséquent beaucoup plus nombreuses.

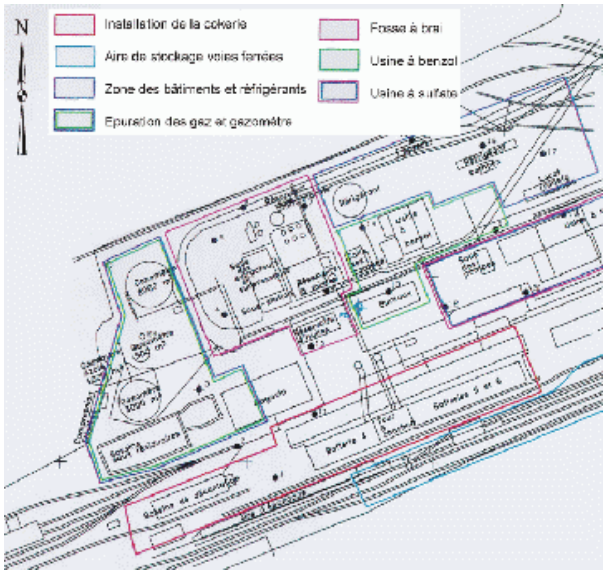


Comment intégrer toute l'information, de l'historique aux analyses sur site ?

1 PRESENTATION DU THEME

1.1 Introduction

A chaque étape de la caractérisation d'un site potentiellement contaminé (diagnostic initial et Evaluation Simplifiée des Risques, diagnostic approfondi et Evaluation Détaillée des Risques, au cours d'un chantier de dépollution), **différents types d'information** sont disponibles sur l'état de contamination d'un site :



- **Informations historiques** quant à l'origine de la contamination, permettant une mise en évidence partielle ou exhaustive des zones a priori contaminées.
- **Lithologie** (coupes de sondage, données géophysiques renseignant sur les types de sols).
- **Information organoleptiques** : présence / absence d'indicateurs de contamination (visuels ou autres).
- **Analyses de terrain** : analyses par fluorescence X, kits chimiques, indicateurs colorimétriques, analyses de gaz FID et PID (détecteurs à ionisation de flamme et photo-ionisation), indice HAP (Hydrocarbures Aromatiques Polycycliques) obtenu par spectrophotométrie UV, etc.
- **Indices globalisants** analysés en laboratoire : somme des 16 HAP (Hydrocarbures Aromatiques Polycycliques), indice HC (HydroCarbures).

Toutes ces informations complètent les résultats d'analyses en laboratoire dont le nombre reste toujours limité.

Se pose alors la question de « comment intégrer de manière rigoureuse toute l'information disponible dans la cartographie des teneurs en polluants et affiner ainsi la quantification des volumes contaminés ? ».

1.2 Cadre métier

Au niveau de l'**ESR (Evaluation Simplifiée des Risques)** et du **diagnostic initial**, l'analyse historique, les paramètres organoleptiques, des techniques analytiques de terrain ou des

investigations in situ (analyses de gaz, géophysique), peu coûteuses et par conséquent permettant une meilleure couverture du site, peuvent s'avérer intéressantes afin d'identifier les sources de pollution et les risques éventuels que le site présente.

Mesurer l'extension de la pollution dans les milieux de transfert relève ensuite de l'**EDR (Evaluation Détaillée des Risques)** et du **diagnostic approfondi**. Pour répondre à cet objectif, il est généralement nécessaire de compléter les investigations déjà réalisées.

Le guide EDR de Gestion des sites pollués (version 2, Juin 2000, p. 23) présente explicitement la valeur ajoutée de certaines informations secondaires ; par exemple, les techniques analytiques de terrain permettent de (citations du guide indiquées entre guillemets) :

- « Réaliser des identifications rapides de familles de polluants, voire des analyses semi-quantitatives pour certaines d'entre elles ; ces résultats, obtenus en temps « réel », sont très utiles lors des phases de dégrossissage.»
- « Cibler des zones de prélèvement en vue d'analyses spécifiques ou plus ponctuelles.»
- « Dresser des documents cartographiques synthétiques. »

Ces informations permettent de réduire les coûts financiers (analytiques, délais), facteur limitant classique lors d'investigations complémentaires. Elles sont ainsi exploitées en vue d'orienter des prélèvements ponctuels ou de dresser des cartes préliminaires, voire dans certains cas en cours de chantier de dépollution. Même si la bonne pratique de ces techniques recommande un calibrage avec des analyses classiques en laboratoire, l'importance de cette validation n'est pas explicitement mentionnée. **Plus important, ces informations ne sont pas explicitement intégrées dans la modélisation de la pollution du site, réalisée à partir d'analyses classiques en laboratoire,** ces dernières étant considérées comme plus exactes[15] et moins sujettes à des interférences (autres polluants ou constituants des matrices solides, teneur en eau, etc.) ; pourtant, l'impact sur le résultat de l'analyse des nombreuses étapes allant du prélèvement de l'échantillon à l'analyse en laboratoire est reconnu.

L'objectif visé par l'intégration géostatistique de l'ensemble des informations disponibles est donc multiple :

- Prendre en compte de façon optimale l'ensemble de ces informations - quelle que soit leur nature et leur exactitude - : information permettant une plus grande densité de points (ex : analyses semi-quantitatives) voire une couverture totale du site (ex : campagne géophysique), etc.
- Réduire les coûts d'investigation tout en améliorant la qualité des estimations de teneurs en échantillonnant de façon plus extensive des informations moins coûteuses que des analyses en laboratoire, une fois avérée la corrélation avec ces dernières.
- Réduire par conséquent les incertitudes liées au niveau de pollution et à son extension spatiale. De ce point de vue, bien que l'analyse des incertitudes soit à juste titre jugée « essentielle et indispensable » dans la méthodologie nationale, l'approche à mettre en oeuvre pour caractériser ces incertitudes est relativement floue, ne facilitant pas la tâche du praticien. Ainsi, le guide méthodologique indique par exemple :
  - Concernant les données qualitatives, la détermination d'une incertitude « n'est pas chose facile. Il est alors de la responsabilité de l'évaluateur d'apprécier le plus précisément possible les gammes de variations des différents paramètres. »
  - Concernant les données quantitatives, l'analyse des incertitudes est jugée plus facile, en renvoyant à la théorie de l'échantillonnage, à l'incertitude sur les appareils de mesure, etc. L'évaluateur « s'attachera donc, autant que faire se peut, à mettre en œuvre les moyens (mathématiques ou autres) qui permettront d'évaluer avec précision le degré d'incertitude associé à une valeur donnée. »

  Qu'entend-on exactement par paramètre ? Quels sont les moyens « mathématiques ou autres » ? Comment définit-on « évaluer avec précision » ?

[15] Par exactitude, on entend à la fois « justesse » (adéquation entre valeur vraie et valeur expérimentale) et « fidélité » (reproductibilité de l'analyse, faible variabilité des teneurs mesurées sur un même échantillon). Par abus de langage, le terme « précis » sera parfois utilisé dans le même sens qu'« exact ».

L'approche géostatistique intégrera les différentes informations disponibles, qu'elles soient qualitatives ou quantitatives, dans une analyse de l'incertitude globale liée à la contamination.

1.3 Position du problème et hypothèses de travail

Afin de formuler clairement le problème métier concerné par ce thème, les termes suivants doivent tout d'abord être définis : variable d'intérêt, information indirecte, variable auxiliaire.

1.3.1 Variable d'intérêt

Usuellement, la **variable d'intérêt** est constituée par les analyses classiques en laboratoire du polluant d'intérêt ; en effet, ce sont sur ces analyses, considérées comme les plus exactes (justes et fidèles), que repose notamment l'étude du dépassement de valeurs guides. Nous parlerons dans la suite directement du **polluant d'intérêt**, lorsque cela ne prête pas à confusion (lorsque par exemple plusieurs types d'analyses sont disponibles pour un même polluant, il faudra être prudent et indiquer quel type d'analyses, privilégié par rapport aux autres, constitue la variable d'intérêt)

Il peut arriver que ces analyses soient, pour un même polluant, issues de campagnes d'investigation différentes (espacées dans le temps et mises en œuvre selon des protocoles d'échantillonnage et d'analyse éventuellement différents) ; ce problème est abordé dans le thème N°1. Néanmoins, il peut arriver qu'il soit délicat de combiner les données issues de deux campagnes successives (notamment en raison de différences de support de prélèvement de l'échantillon). Dans ce cas, on pourra considérer comme variable d'intérêt celle issue de la campagne conduisant aux données de meilleure qualité[16] et intégrer les données des autres campagnes en les considérant comme une information indirecte, en exploitant l'éventuelle corrélation inter-campagnes.

Une pollution implique fréquemment plusieurs polluants d'intérêt. Deux cas sont alors possibles :

- *Les polluants ne présentent pas de corrélation statistique* (exemple : polluants issus d'activités différentes sur le site ; ou encore, en particulier dans le cas de sites anciens : polluants qui, bien qu'issus d'une même activité, ont des propriétés physico-chimiques et biologiques différentes influant sur leur évolution à long terme).

  L'estimation des teneurs doit alors être menée de façon indépendante polluant par polluant, la connaissance des teneurs en un polluant n'apportant pas d'information sur les teneurs d'un autre polluant.

- *Certains polluants présentent une corrélation statistique significative* (exemple : différents types d'hydrocarbures issus d'une même activité).

  Il est alors possible d'exploiter ces corrélations en estimant chaque polluant à partir des mesures de l'ensemble des polluants disponibles qui lui sont corrélés. Exemple : dans le cas de deux polluants corrélés A et B, on considérera A comme variable d'intérêt et B comme information indirecte, puis inversement.

  Cette approche n'est cependant intéressante que dans le cas d'un **échantillonnage hétérotopique**, c'est-à-dire lorsque les analyses disponibles pour les différents polluants ne proviennent pas toutes des mêmes points de prélèvements (exemple : certains polluants mieux reconnus, ou bien non analysés en certains points). En effet, si deux polluants sont mesurés uniquement aux même points, ils n'apportent l'un à l'autre aucune information supplémentaire intéressante par rapport à la connaissance du polluant lui-même.

Seule exception : dans le cas d'un échantillonnage isotopique (ou homotopique, i.e. les polluants sont analysés aux mêmes points), si on observe une meilleure structuration spatiale d'un polluant par rapport à d'autres, alors ce polluant améliorera l'estimation de ces autres polluants. Exemple : supposons deux polluants corrélés A et B, connus aux mêmes points. Si, contrairement à B, A pose de fortes difficultés analytiques, sa structure spatiale (variogramme) présentera une variabilité non négligeable à toute petite distance, non pas due à l'hétérogénéité du sol mais par exemple à la variabilité issue de la mauvaise reproductibilité des analyses. Dans ce cas, la connaissance plus exacte du polluant B améliorera la qualité de l'estimation du polluant A.

[16] Le critère permettant de juger de la qualité respective des campagnes peut dépendre de l'objectif visé.

1.3.2 Informations indirectes et variable auxiliaire

On appellera **information indirecte** tout type d'information disponible pouvant affiner la connaissance de la variable d'intérêt (le polluant), tel que :

- Analyses sur site (techniques analytiques de terrain) : analyses par fluorescence X, kits chimiques, indicateurs colorimétriques, analyses de gaz PID, FID, etc. ;
- Indices globalisants analysés en laboratoire : indice HC, somme des 16HAP, etc. ;
- Variables qualitatives et organoleptiques : présence/absence d'éléments indicateurs de contamination (indication visuelle, traces de goudron dans le cas d'hydrocarbures) ;
- Mesures géophysiques ;
- Autres polluants.

Il est entendu que ce qui différencie les analyses en laboratoire de certaines de ces informations, notamment les techniques analytiques de terrain, ne peut dans certains cas n'être que le mode de constitution de l'échantillon et son support. L'objectif de ce chapitre n'est en aucun cas de se prononcer de façon générale sur l'exactitude comparée de ces différents types d'informations, mais bien de voir dans quelle mesure et selon quelle méthodologie il est possible de les combiner au mieux.

Ces informations indirectes peuvent être regroupées en plusieurs catégories : quantitatives continues (i.e. connues de façon exhaustive, par exemple des mesures géophysiques), quantitatives discrètes (i.e. échantillonnées avec une densité/répartition de points différente, par exemple des analyses in situ ou sur site, analyses d'autres polluants) ou qualitatives (indices organoleptiques, par ex. présence/absence d'odeur).

L'analyse des relations entre ces informations indirectes et le polluant d'intérêt mettra en évidence quelles informations doivent être privilégiées et comment, dans certains cas, ces informations indirectes peuvent être combinées entre elles. Ainsi, on appellera **variable auxiliaire** la variable qui synthétise au mieux l'ensemble des informations indirectes et qui sera finalement explicitement utilisée pour l'estimation de la variable d'intérêt. Cette variable auxiliaire peut être :

- simplement une information indirecte de départ (exemple : un polluant corrélé au polluant d'intérêt),
- le résultat d'un pré-traitement d'une information indirecte de départ (transformation pour linéariser ou améliorer une corrélation, nouveau codage des différentes classes d'une information qualitative),
- le résultat de la combinaison de plusieurs informations indirectes.

Il peut dans certains cas être intéressant de considérer explicitement plusieurs variables auxiliaires (voir § 2.3.1).

Le problème posé, reformulé grâce aux termes définis ci-dessus, peut donc s'énoncer comme suit : **« Comment améliorer l'estimation d'un polluant d'intérêt en intégrant la connaissance d'une ou plusieurs variables auxiliaires qui lui sont corrélées, ces variables constituant une synthèse intéressante de l'ensemble des informations indirectes disponibles ? »**

1.4 Objectifs

Ce chapitre est axée sur l'application de techniques géostatistiques **multivariables** pour intégrer des informations indirectes afin d'améliorer les conditions d'estimation ou de simulation de teneurs en polluants dans les sols (variables d'intérêt). Nous ne traitons pas ici le cas de la pollution des eaux souterraines, qui peuvent également gagner à prendre en compte des informations indirectes (issues par exemple de modèles déterministes d'écoulement et de transport d'éléments) pour affiner la cartographie de polluants suivis en quelques piézomètres.

En outre, pour les raisons mentionnées au paragraphe 1.3.1, nous nous limitons sans perdre de généralité au cas d'une seule variable d'intérêt, dont la reconnaissance est complétée par une ou

plusieurs informations indirectes quantitatives, semi-quantitatives ou qualitatives. En effet, dans le cas d'une pollution par plusieurs éléments, chacun d'entre eux sera tour à tour considéré comme le polluant d'intérêt (la variable d'intérêt), les autres étant intégrés comme informations indirectes dans sa modélisation (en cas de corrélation).

Juger de la pertinence de certaines informations indirectes n'est pas le propos du manuel. Nous supposons que si elles sont mises en œuvre, leur recours est fondé et qu'elles sont adaptées au problème posé. Nous nous intéresserons par contre à la validation de l'information apportée par ces informations indirectes pour améliorer la connaissance du polluant d'intérêt.

Précisons que notre but n'est pas ici de remplacer le prélèvement d'échantillons de sols et l'analyse des données de teneur en polluant par quelques informations indirectes moins coûteuses, mais d'identifier et quantifier l'apport de ces informations afin de les exploiter au mieux. L'intérêt est de ne pas perdre l'information que ces informations indirectes contiennent au stade de la modélisation (cartographie ou calcul de volumes). Cela peut aussi aider à réduire le nombre d'analyses en laboratoire, donc à optimiser l'échantillonnage avec mise en balance des coûts avec la précision attendue de l'estimation (voir thème N°4). En effet, la mise en évidence d'une bonne corrélation entre polluant et information indirecte peut orienter des échantillonnages complémentaires qui privilégieraient une reconnaissance par des informations indirectes, moins coûteuses et donc permettant à coût fixé une meilleure couverture de la zone d'étude.

Finalement, il a été mentionné que l'analyse historique peut également être considérée dans certains cas comme une information indirecte (à condition que cette considération ne soit pas faite a priori mais bien validée par le terrain). En effet, dans la mesure où elle peut apporter une connaissance fiable a priori de plusieurs zones d'intensité de pollution différente, elle constitue une information sur la pollution. Classiquement, la connaissance de plusieurs zones d'intensité de pollution différente suggère une modélisation des teneurs zone par zone. Néanmoins, cette modélisation requiert davantage de données ; certaines hypothèses de similitude de la structure spatiale peuvent néanmoins être posées entre les différentes zones (voir thème N°1).

1.5 Pratiques habituelles

Comme cela a été évoqué en introduction, les informations indirectes (historiques, lithologiques, géophysiques ou autres) sont souvent utilisées pour conduire les campagnes d'échantillonnage et décider de l'emplacement des sondages. On constate cependant un manque fréquent de fondement méthodologique dans l'exploitation et l'intégration de ces informations :

- Limitation de l'étude à des zones supposées contaminées à la vue de données historiques, rare remise en question de ces zones.
- Pas d'analyse systématique des éventuelles corrélations entre mesures in situ et en laboratoire.
- Corrélations supposées a priori mais non vérifiées. Par exemple, certaines relations linéaires entre indices globalisant et teneurs en polluant individuel (exemple[17] : indice HAP = [benzo(a)pyrène]*10) sont tenues pour acquises mais non validées expérimentalement.
- Plus généralement exploitation insuffisante des informations déjà acquises pour reconnaître les zones où l'état de contamination des sols reste encore incertain.
- Non prise en compte des problèmes de support et d'hétérogénéité des prélèvements, que ce soit en vue d'analyses classiques de laboratoire lors d'une investigation de site ou pour la réalisation d'analyses sur site, qui sont notamment de plus en plus utilisées pour le tri des terres en cours de dépollution.

[17] Voir « Bouchez M., D. Blanchet, F. Haeseler et JP. Vandecasteele (1996), Les hydrocarbures aromatiques polycycliques dans l'environnement: propriétés, origine, devenir. Revue de l'IFP, 51 (3),pp. 407-419. » ou « Oosterbaan-Eritzpokhoff J. (2000), Utilisation des méthodes de l'analyse exploratoire de données environnementales pour l'établissement de descripteurs macroscopiques de la dynamique des termes-sources polluants dans la géosphère : application à la caractérisation de la pollution par des hydrocarbures aromatiques polycycliques (HAP) de sites industriels anciens ayant utilisé la pyrolyse de la houille (usines à gaz et cokeries). Thèse de Docteur en Hydrologie et Hydrogéologie Quantitatives, Ecole des Mines de Paris. »

1.6 Intérêt d'une approche géostatistique

Par rapport à ces pratiques, l'approche géostatistique présente l'intérêt de reposer essentiellement sur l'exploitation des corrélations observées entre polluant d'intérêt et informations indirectes. La validation expérimentale de cette corrélation sera une condition *sine qua non* de prise en compte de l'information indirecte. Évidemment, la connaissance exhaustive sur une zone donnée d'une information sans lien avéré avec la pollution suivie n'apportera aucune information sur celle-ci. En fonction de la qualité de la corrélation entre informations indirectes et variable d'intérêt, un poids plus ou moins élevé sera ainsi donné à l'information indirecte. Meilleure sera la corrélation, plus important sera l'apport de l'information indirecte.

L'intérêt est donc d'introduire dans la modélisation des teneurs en polluant toute information indirecte corrélée avec les teneurs, que cette information soit qualitative, semi-quantitative ou quantitative.

Le bénéfice attendu est de pouvoir capturer, à l'échelle du site, des tendances spatiales que ces données indirectes plus nombreuses ou de plus grande continuité spatiale permettront de suivre plus facilement.

1.7 Méthodologie

Dans ce chapitre sont détaillés des aspects importants de l'intégration d'informations indirectes dans une approche géostatistique. A la différence de certains thèmes (par exemple le thème N°2), les étapes présentées ne sont pas forcément toutes nécessaires et dépendent à la fois de l'objectif visé et des données disponibles. Les informations indirectes peuvent être utilisées pour l'optimisation de l'échantillonnage, pour la cartographie de teneurs (estimation locale), ou bien pour une estimation globale (volumes à dépolluer) basée sur des simulations.

Comme lors d'une approche géostatistique monovariable (voir thème N°1), l'analyse des données est la première étape, fondamentale, de l'approche géostatistique pour l'intégration d'informations indirectes ; c'est en particulier cette analyse qui permet de valider l'existence ou non de corrélations entre le polluant d'intérêt et les informations indirectes pressenties.

Ensuite, l'approche géostatistique diffère en fonction des informations disponibles et du problème posé. Nous distinguerons notamment les points suivants :

1. Estimation des teneurs avec prise en compte d'information indirecte.
2. Cas de plusieurs informations indirectes.
3. Application à la délimitation et quantification des zones à dépolluer.

2 DEMARCHE GEOSTATISTIQUE DETAILLEE

Nous nous plaçons dans le cadre suivant : un polluant d'intérêt connu en quelques points, que nous cherchons à estimer à l'aide de données de teneurs en polluant dans les sols et d'un seul type d'information indirecte (excepté pour le § 2.3), connue à la fois en des points de mesure de la teneur et en d'autres non renseignés.

2.1 Considérations pratiques : analyse des données disponibles

2.1.1 Echantillonnage de l'information indirecte et corrélation avec le polluant

L'objectif étant d'intégrer une information indirecte dans l'estimation des teneurs en polluant, l'essentiel du travail géostatistique consiste à préciser dans quelles conditions cette intégration peut se faire et d'évaluer le poids et le crédit qui peuvent être donnés à l'information indirecte ?

L'analyse expérimentale de la relation entre teneur en polluant et information indirecte est donc fondamentale. Cette analyse passe tout d'abord par l'examen du **nuage de corrélation** entre les deux

variables, dont le calcul requiert que les deux variables soient mesurées en des points communs. Par conséquent, si l'on désire exploiter une information indirecte que l'on pressent corrélée au polluant d'intérêt, il est primordial que cette information indirecte soit en partie échantillonnée aux mêmes points que le polluant d'intérêt, afin qu'il soit possible de valider expérimentalement la corrélation entre les deux variables.

La Figure III-1 présente un exemple de corrélation entre deux variables : le polluant d'intérêt est le Zinc, supposé dans cet exemple moins bien reconnu (16 analyses) que le Plomb (36 analyses). Les teneurs en Plomb sont considérées comme une information indirecte potentielle pour l'estimation des teneurs en Zinc. L'implantation superposée des données des deux polluants, avec une représentation symbolique proportionnelle à la teneur en polluant, suggère une bonne relation entre les deux polluants : les fortes teneurs en Zinc correspondent aux fortes teneurs en Plomb, et il semble en être de même pour les faibles teneurs.

Le nuage de corrélation précise cela, en représentant sur un même diagramme chaque couple de teneurs Pb – Zn analysées au même point : la teneur en Plomb en abscisse, celle en Zinc en ordonnée. Ce nuage confirme la bonne relation entre les deux éléments, corrélation qui est en partie résumée par la droite de régression linéaire et le **coefficient de corrélation**.

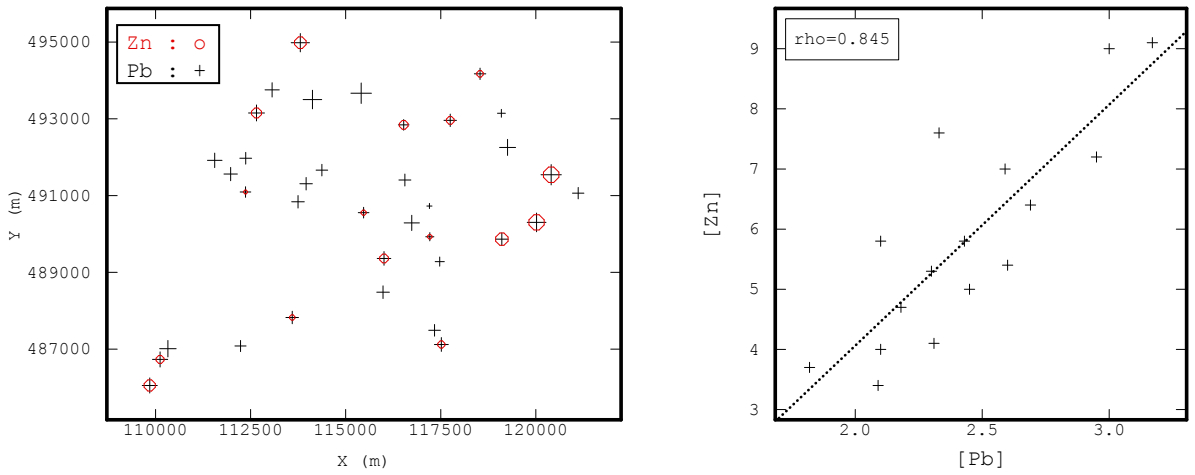


Figure III-1 Relation entre teneurs en polluant ([Zn], en mg/kg ms) et information indirecte ([Pb], en mg/kg ms) : implantation des données (à gauche) et nuage de corrélation, droite de régression linéaire et coefficient de corrélation (à droite).

Les conditions d'évaluation de la corrélation entre les deux variables ci-dessus sont favorables, dans la mesure où l'information apportée par Pb sur Zn concerne aussi bien les faibles teneurs que les fortes teneurs et permet donc un bon calage de la corrélation. En effet, le nuage de corrélation, évalué à partir des mesures communes aux deux variables, va ensuite être supposé **représentatif** du lien entre les deux variables sur l'ensemble du site. Ainsi, si l'on considère un polluant connu en quelques points et une information indirecte échantillonnée de façon plus dense, l'observation expérimentale (à partir des points communs entre les deux variables) d'une forte corrélation entre les deux variables poussera à conclure qu'aux points où seule l'information indirecte est connue, la variable d'intérêt doit suivre le comportement de l'information indirecte.

Il est donc important que les différentes classes de teneurs du polluant d'intérêt soient si possible échantillonnées par l'information indirecte, afin d'éviter de baser l'ensemble des calculs sur une corrélation entre les deux variables qui n'aurait été calibrée que sur les faibles ou les fortes teneurs du polluant, et extrapolée aux autres classes de teneurs.

La Figure III-2 présente quelques exemples classiques de corrélation entre deux variables X et Y. En supposant que X est la variable d'intérêt et Y l'information indirecte, les différents cas peuvent être décrits de la façon suivante :

- Forte corrélation positive (en haut à gauche) entre X et Y : cas très favorable d'une très bonne correspondance entre les valeurs des deux variables, qui prennent simultanément des valeurs fortes, moyennes ou faibles. La connaissance de l'information indirecte Y en des points où X n'est pas connu fournit, en exploitant la régression, une bonne information sur la valeur attendue pour X.

- Forte corrélation négative (en bas à gauche) entre X et Y : ce cas est tout aussi favorable que le précédent, sachant que le sens de la corrélation sera bien évidemment pris en compte dans la modélisation.
- Corrélation positive faible entre X et Y (en haut à droite) : même s'il existe une relation entre les deux variables, celle-ci n'est pas forte. Par conséquent, on accordera un poids beaucoup plus faible à l'information indirecte lors de sa prise en compte dans l'estimation de X.
- En cas d'absence de corrélation entre X et Y (en bas à droite), la connaissance même excellente de l'information indirecte Y n'est d'aucune utilité en vue de l'estimation de la variable d'intérêt X.

Seule la qualité de la corrélation entre les deux variables, établie à partir de l'observation du nuage de corrélation, est donc importante ; que les variables soient positivement ou négativement corrélées n'aura pas d'incidence, ces deux cas étant similaires en termes de qualité d'information apportée par l'information indirecte.

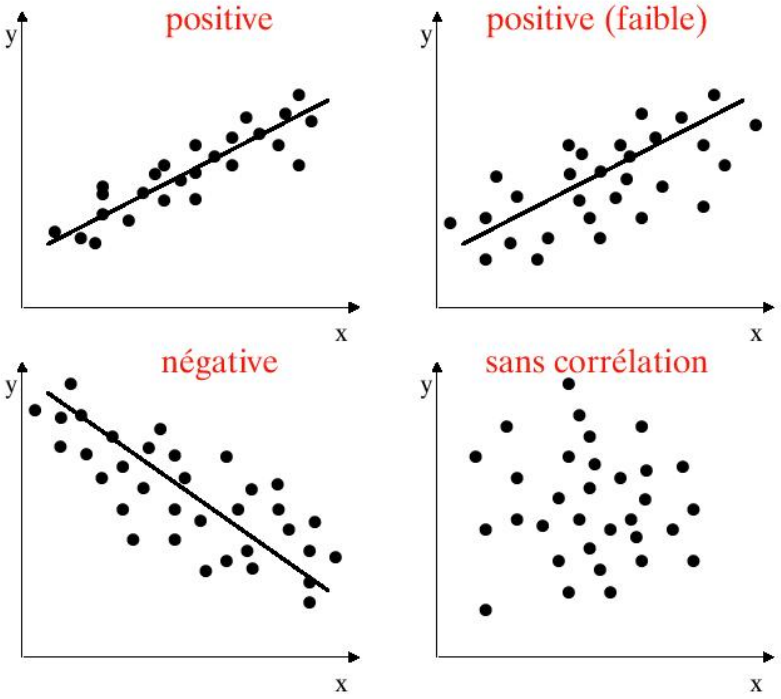


Figure III-2 Nuages de corrélations classiques entre deux variables X et Y.

Finalement, certaines relations entre variables peuvent ne pas être linéaires mais plus complexes ; il peut alors être intéressant de transformer l'information indirecte (ex : transformation logarithmique ou exponentielle, linéaire par morceaux) de façon à linéariser la corrélation, les techniques statistiques étant mieux opérationnelles en cas de corrélation linéaire. La Figure III-3 illustre cela : on constate une relation très nette entre X (la teneur en polluant) et Y (l'information indirecte), mais qui n'est pas linéaire, rendant illicite le calcul d'une droite de régression ou d'un coefficient de corrélation. Par contre, en appliquant une transformation logarithmique à X, on constate que la relation entre ce logarithme et Y est devenue linéaire et de très bonne qualité. Le travail gagnera dans pareil cas à être réalisé sur la transformée logarithmique de X, tout en assurant in fine un retour en concentrations brutes.

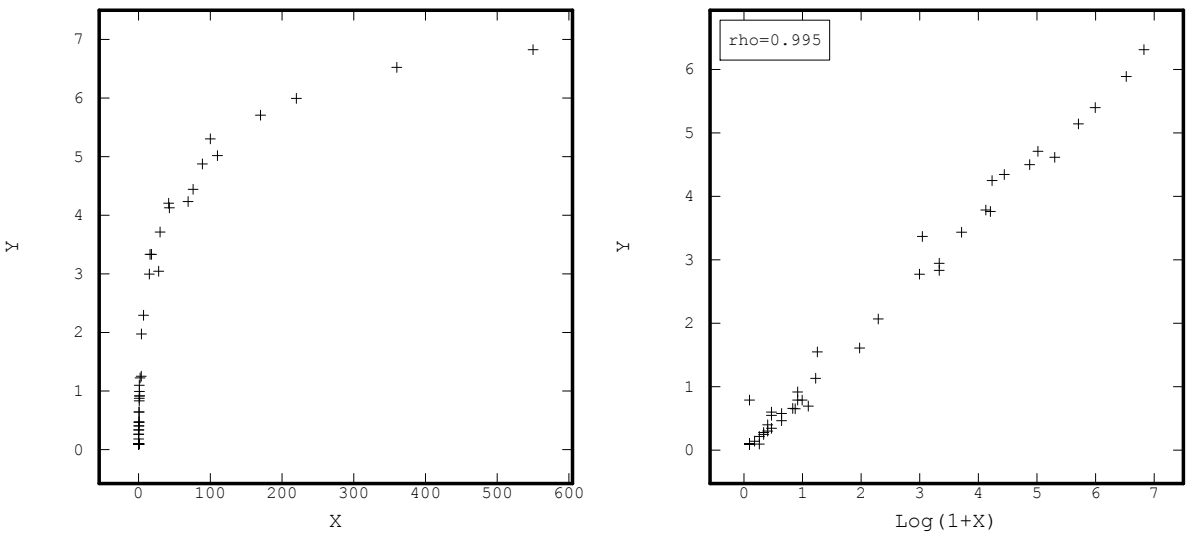


Figure III-3 Nuages de corrélations entre une variable d'intérêt X et Y (à gauche) ou entre le logarithme de X et Y (à droite).

Nous nous limitons ici à une corrélation linéaire entre variables. Les données indirectes quantitatives peuvent dans certains cas présenter des relations plus complexes avec la variable d'intérêt, ou apporter une information partielle (incomplète) sur les teneurs ; c'est par exemple le cas lorsqu'elles ne renseignent que sur des teneurs plutôt faibles ou fortes. Une solution possible pour prendre en compte de telles informations indirectes est de traduire les données indirectes quantitatives en une information qualitative comme cela est indiqué dans le paragraphe suivant. D'autres plus avancées sortent du cadre de ce manuel.

2.1.2 Cas d'une information indirecte qualitative

Usuellement, une information indirecte qualitative est codée numériquement. Par exemple, la présence ou l'absence d'odeur associée à une pollution sera codée 0 pour « absence », 1 pour « faible odeur » et 2 pour « forte odeur ». Un tel codage, arbitraire, pose problème pour intégrer cette information indirecte dans l'estimation des teneurs. En effet, il n'est pas du tout évident que ce codage permette de mettre en avant une corrélation avec les teneurs en polluant. Cela est encore plus vrai en présence d'un paramètre qualitatif de type lithologie si l'on code arbitrairement les différents horizons rencontrés par 1, 2, 3, etc.

Plusieurs statistiques sont utilisées classiquement pour examiner les relations entre la teneur en polluant et une information indirecte qualitative, supposée désormais codée en N classes ou catégories : moyenne, médiane et histogramme des teneurs par classe.

La meilleure corrélation entre une variable numérique Z et une information qualitative P prenant plusieurs valeurs est obtenue en substituant à P la moyenne de Z par classe de l'information qualitative. Cette transformation permet ainsi de convertir l'information P en une variable auxiliaire de même « dimension » que Z, et maximisant leur corrélation, sans nécessiter aucun codage arbitraire (voir Saporta, 1990). Cela sera illustré au § 3.2.2 pour une information indirecte de type lithologie.

D'autres méthodes permettent également d'intégrer des informations indirectes qualitatives (approche par indicatrices, par bayesian maximum entropy, etc.).

2.2 Estimation de teneurs avec prise en compte de l'information indirecte

Dans le cas où l'analyse des données a permis de valider l'existence d'une corrélation entre le polluant *Z* et la variable auxiliaire *Aux* (information indirecte de départ, ou transformée), il s'agit à présent d'intégrer effectivement la connaissance de la variable auxiliaire dans l'estimation des teneurs en polluant.

En généralisant le principe de l'estimation par krigeage présenté dans le thème N°1, l'estimation $z^*(x_0)$ du polluant au point x_0 par **cokrigeage** en utilisant la variable auxiliaire *Aux* est obtenue par combinaison linéaire des n teneurs aux points x_i et des p valeurs de la variable auxiliaire aux points x_j :

$$z^*(x_0) = \sum_{i=1}^{n} \lambda_{1i}\, z(x_i) + \sum_{j=1}^{p} \lambda_{2j}\, aux(x_j) \ .$$

L'objectif du cokrigeage est donc de déterminer les poids de cette combinaison linéaire.

Il est important de garder à l'esprit que cette estimation reste évidemment une ***estimation non biaisée*** *de la teneur en polluant ; le système de cokrigeage inclut en effet des conditions de non biais permettant d'assurer que l'estimation de la teneur n'est pas « perturbée » par des valeurs de variable auxiliaire qui, pouvant être dans une échelle différente, ne peuvent être combinées avec les teneurs de façon quelconque. Tout comme le krigeage, le cokrigeage permet également l'estimation de la teneur en polluant sur des blocs de taille donnée.*

La résolution du système de cokrigeage requiert la connaissance de la structure spatiale de chacune des variables, approchée par le calcul de leurs variogrammes expérimentaux (variogrammes simples). Généralement, plus le niveau de corrélation entre les variables est élevé et plus les variogrammes simples ont tendance à être similaires (proportionnels). Ces variogrammes ne sont cependant pas suffisants pour réaliser l'estimation par cokrigeage. Ils ne renseignent pas en effet sur la relation qui unit les deux variables et qui nécessairement varie aussi avec la distance entre points de mesure. Cette relation est décrite par un **variogramme croisé** qui résume la corrélation entre les deux variables en fonction de la distance (voir Figure III-4) ; par exemple, deux variables peuvent présenter une faible corrélation à petite distance (sources de pollution différente, mécanismes de transformation chimique ou autre entre éléments) mais une bonne corrélation à grande échelle (liée à la présence commune des polluants sur certaines zones du site).

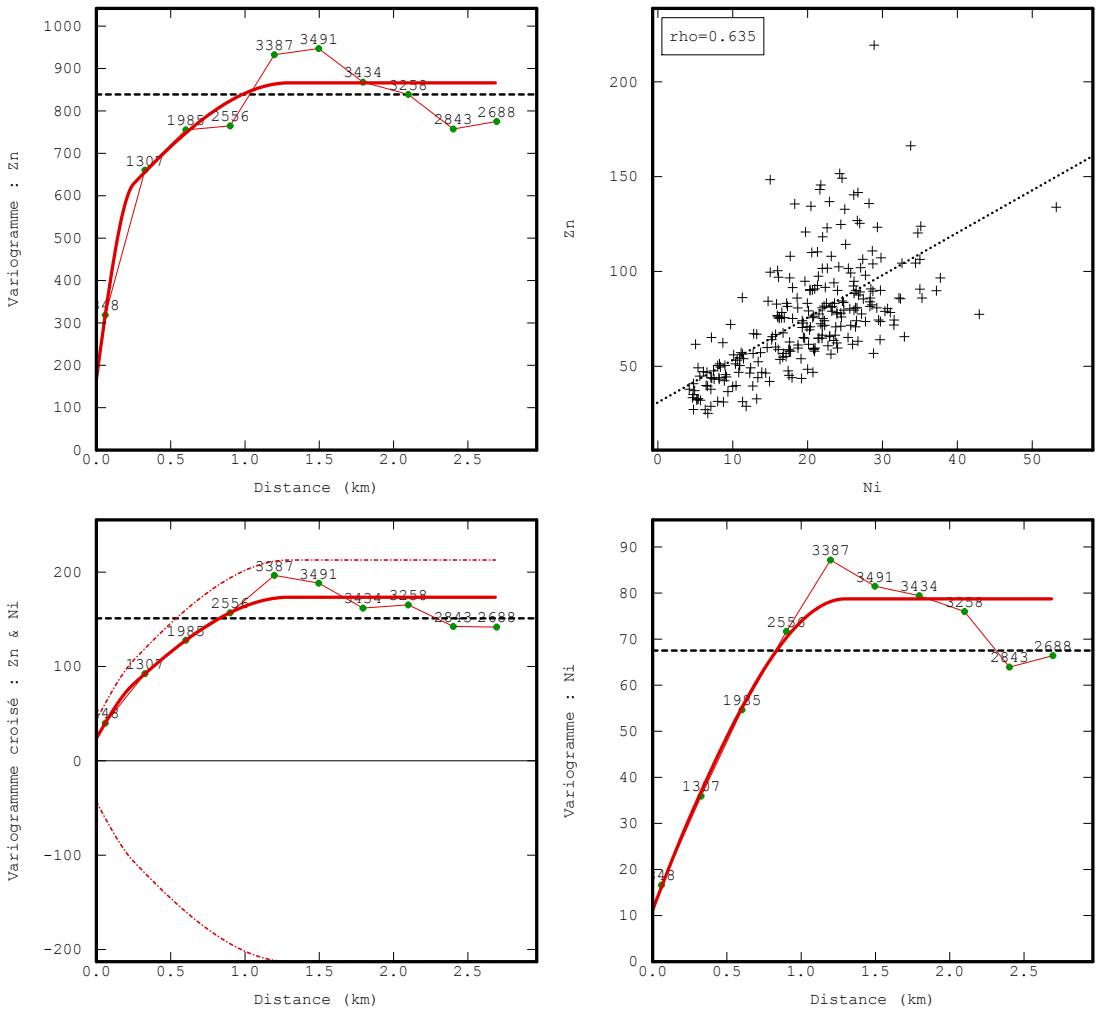


Figure III-4 Variogrammes simples et croisé entre deux polluants, nuage de corrélation entre les polluants (teneurs exprimées en mg/kg ms) avec indication de la droite de régression.

Une fois le calcul expérimental des variogrammes simples et croisé effectué, un modèle bivariable de variogramme est ajusté aux structures expérimentales. Contrairement au cas monovariable, il est ici nécessaire que ce modèle ajuste en même temps les trois variogrammes expérimentaux. Ce modèle de cokrigeage permet alors l'estimation des variables par cokrigeage.

Caractérisant le degré de corrélation pour une distance donnée, le variogramme croisé peut prendre des valeurs négatives (en cas de corrélation négative entre les deux variables). Une approche classique pour ajuster un modèle de variogramme à plusieurs variables consiste à se placer dans le cadre d'un ***modèle linéaire de corégionalisation****, c'est-à-dire à ajuster les différents variogrammes par une combinaison linéaire des mêmes structures de base (par exemple, un effet de pépite plus une structure sphérique), l'importance relative de chaque structure de base pouvant varier selon les variogrammes, à condition de vérifier certaines relations d'autorisation.*

Quelques remarques :

- Certaines informations indirectes continues peuvent être disponibles pour l'ensemble du site (exemple : campagne géophysique). Dans ce cas, la mise en œuvre pratique du cokrigeage classique devient délicate en raison du trop grand nombre de données impliquées. On aura alors recours au **cokrigeage colocalisé**, simplification du cokrigeage consistant à réduire la taille du système de cokrigeage en ne prenant pas en compte toutes les données disponibles pour la variable auxiliaire mais uniquement celle présente au point d'estimation.
- Homogénéité des données de teneur en polluant : plutôt que de chercher à combiner des données issues de campagnes différentes et obtenues selon des protocoles potentiellement hétérogènes (exemples : prélèvement par carottage ou sur des parois de fosses), il est possible d'utiliser les résultats de campagnes précédentes sous forme de variable auxiliaire pour améliorer l'estimation des données de teneur issues de la campagne jugée la plus informative (en terme par exemple de justesse des mesures ou de nombre d'analyses), à condition de disposer de quelques points communs entre les deux campagnes.
- Dans certains cas, l'information collectée lors de l'investigation est essentiellement obtenue par une technique d'analyse sur site (par exemple fluorescence X), complétée par quelques analyses en laboratoire. Ces dernières, en nombre insuffisant, ne permettent pas une analyse bivariable dans de bonnes conditions. Il peut alors être préférable d'opter pour une approche géostatistique (cartographie) directe sur les données issues des analyses sur site.
- Il est possible qu'un polluant d'intérêt présente une relation avec une information auxiliaire, mais que cette relation présente un décalage spatial, suite par exemple à une migration depuis une source de pollution. Cette corrélation différée peut être appréhendée par des outils avancés d'analyse de la structure spatiale (covariances croisées par exemple).

2.3 Cas de plusieurs informations indirectes

En général, plus d'une information indirecte est disponible ; le §3 présentera en illustration l'estimation d'une pollution en hydrocarbures à l'aide d'analyses classiques en laboratoire et d'informations lithologiques et organoleptiques plus nombreuses.

Nous présentons ci-dessous différentes possibilités d'intégrer plusieurs informations indirectes dans la modélisation du polluant d'intérêt.

2.3.1 Cokrigeage basé sur l'ensemble des informations indirectes

Il n'existe aucune limite théorique sur le nombre de variables auxiliaires pouvant être intégrées dans une approche de type cokrigeage. Néanmoins, ces variables auxiliaires :

- sont usuellement partiellement redondantes,
- ne présentent pas toutes les mêmes qualités de corrélation avec la variable d'intérêt.

En outre, considérer un modèle intégrant de nombreuses variables pose rapidement des problèmes d'inférence et de calage.

Cette approche sera néanmoins intéressante dans le cas où les quelques variables auxiliaires disponibles apportent des informations différentes et complémentaires par rapport à la variable d'intérêt ; ainsi, un polluant auxiliaire peut expliquer le comportement à petite échelle de la variable d'intérêt, tandis qu'un autre polluant ou la lithologie présentera au contraire une corrélation à grande échelle.

2.3.2 Pré-traitement et combinaison des informations indirectes

Afin de résoudre les problèmes posés par la mise en œuvre d'un modèle impliquant plus de 2 ou 3 variables, plusieurs approches sont possibles :

- Hiérarchisation des variables en fonction de la corrélation avec les teneurs et sélection des variables les plus corrélées. Il est en outre possible de combiner ces informations (voir l'illustration du §3).
- Utilisation de techniques statistiques multivariables classiques, synthétisant en une ou deux variables les informations indirectes de départ : régressions linéaires multiples, analyse en composantes principales. En particulier, dans le cas de plusieurs indices qualitatifs / organoleptiques, l'analyse factorielle des correspondances permet de synthétiser l'information de différentes variables qualitatives. Les indices sont décomposés en facteurs point à point non corrélés, dont on conserve ceux qui restituent une part importante de l'information. Si ces facteurs sont corrélés aux teneurs, ils sont utilisables comme variables auxiliaires pour le cokrigeage des teneurs (voir Jeannée & de Fouquet, 2003).

2.4 Application à la délimitation et la quantification des zones à dépolluer

L'approche géostatistique présentée dans le thème N°2 peut être généralisée afin d'exploiter l'existence d'informations indirectes corrélées aux teneurs de la variable d'intérêt.

Des techniques de co-simulation peuvent être mises en œuvre dans ce but, en utilisant la méthodologie des simulations étendue au cas de deux variables : transformation gaussienne, analyse des variogrammes simples et croisés des deux variables, co-simulations.

Les techniques de co-simulation exploitent la décomposition du modèle de variogramme bivariable issue du modèle linéaire de corégionalisation : chacun des facteurs de base du modèle est simulé, puis les simulations finales sont reconstruites à partir de ces facteurs.

Le formalisme des indicatrices permet également de répondre au problème de la simulation de teneurs en polluant en prenant en compte des informations indirectes qualitatives.

3 ILLUSTRATION

3.1 Contexte

L'illustration présentée concerne la dépollution d'un site arrêté Charbonnages de France en vue de sa réhabilitation. Les travaux de terrain ont révélé l'existence d'une dalle de béton très épaisse (métrique) sous l'ancienne mazouterie ainsi que des fuites de fuel vers les remblais situés au nord du bâtiment. A l'issue des premiers carottages, une analyse qualitative a permis la délimitation grossière de la zone contaminée (présentant des valeurs supérieures à 500 mg/kg ms en hydrocarbures totaux (dénotés HT dans la suite), seuil jugé sécuritaire d'un point de vue sanitaire, la VDSS étant égale à 2500 mg/kg ms). Un échantillonnage complémentaire 3D de la zone concernée (23 carottages, voir **Figure III-5**) a été réalisé, avec description de la lithologie et relevé systématique d'informations organoleptiques (information olfactive et visuelle). En outre, les prélèvements ciblés de 32 échantillons le long des carottages ont fait l'objet d'une analyse des teneurs en hydrocarbures totaux.

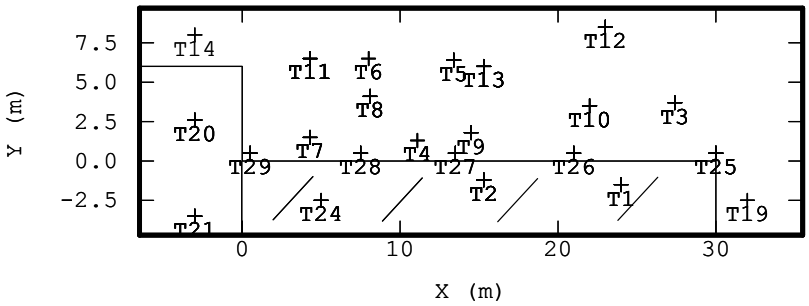


Figure III-5 Implantation des carottages – Vue du dessus. Localisation du bâtiment chaufferie contenant l'ancienne mazouterie (au sud) et de la salle des machines à gauche (à l'ouest).

L'objectif de cette illustration géostatistique est de montrer l'apport de paramètres organoleptiques et lithologiques pour la représentation tridimensionnelle de la contamination. Comme cela sera illustré, l'estimation à partir des teneurs en HT seules conduit à une surestimation irréaliste de la contamination associée à une forte incertitude. En outre, le protocole d'échantillonnage validé par un expert indépendant ne préconisait pas de prise d'échantillon pour analyse dans la partie supérieure des forages où il n'apparaît pas de pollution significative. La cartographie de la contamination obtenue permet alors de valider les hypothèses formulées quant à l'origine de la contamination : fuite des hydrocarbures depuis la mazouterie vers les remblais naturels adjacents au bâtiment.

3.2 Analyse des données

3.2.1 Analyses en hydrocarbures totaux

L'histogramme des 32 teneurs en HT disponibles est donné à la **Figure III-6**. Ce nombre d'échantillons, peu élevé en particulier pour un problème de modélisation tridimensionnel, conduit à une forte incertitude au niveau de la caractérisation de la variable d'intérêt (distribution, structure spatiale).

La **Figure III-7** montre que les trois plus fortes teneurs (supérieures à 5000 mg/kg ms) se situent entre 4.8 et 6 m de profondeur, l'essentiel des autres teneurs élevées se situant entre 5.8 et 7m.

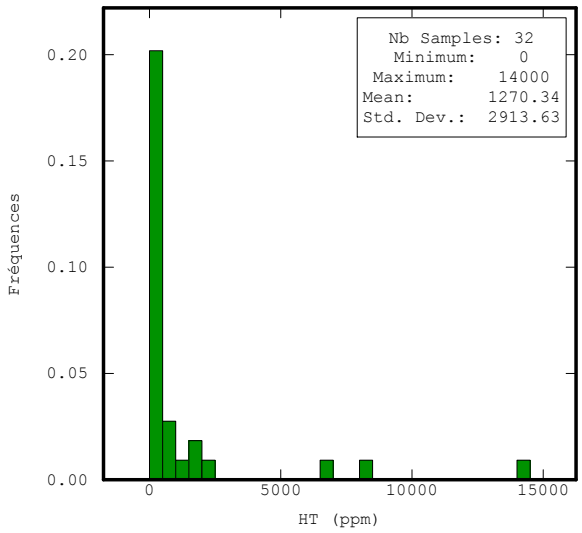


Figure III-6 Histogramme des teneurs en hydrocarbures totaux.

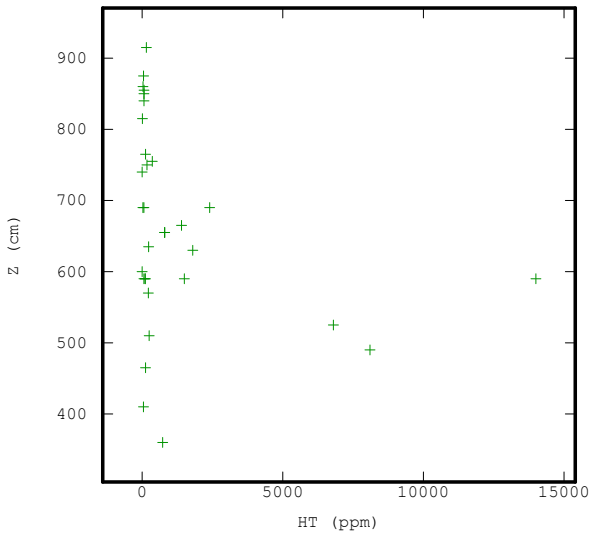


Figure III-7 Nuage de corrélation entre les teneurs en hydrocarbures totaux HT (en mg/kg ms) et la profondeur moyenne de l'échantillon correspondant (Z, en cm).

3.2.2 Corrélation avec les informations lithologiques et organoleptiques

Outre l'information fournie par les 32 analyses en HT, plusieurs types d'informations indirectes ont été relevées lors des investigations. De coût négligeable et par conséquent plus nombreuses, ces informations gagneront à être intégrées dans la modélisation en cas de corrélation avec les teneurs en HT. Elles constitueront une variable auxiliaire, de poids moindre que la variable d'intérêt (teneur en HT) mais qui pourra révéler des tendances et en affinera l'estimation en l'absence d'analyse en HT.

Lithologie

Les informations lithologiques, au nombre de 93, sont regroupées en 4 catégories :

- marne jaune (MJ), en 16 points de prélèvement : les analyses en HT disponibles pour cet horizon ont une teneur moyenne de 50 mg/kg ms ;
- argile graveleuse (AG) en 29 points, les analyses disponibles en HT ayant une teneur moyenne de 150 mg/kg ms ;
- argile (A) en 14 points ; bien qu'aucun échantillon n'ait fait objet d'une analyse en HT, la position de cet horizon en exclut la contamination, étant situé au-dessus de la couche d'argile graveleuse et de la source de contamination en hydrocarbures (conforté par les observations organoleptiques) ; une teneur de 50 mg/kg ms (valeur sécuritaire en termes sanitaires) a été associée à ces points.
- remblais (R) en 42 points, les analyses en HT ayant une teneur moyenne de 625 mg/kg ms.

Pour les trois horizons ayant fait l'objet d'analyses en HT, l'information lithologique est remplacée par la moyenne en HT correspondant à l'horizon. Cette variable transformée de l'information lithologique, appelée « moyenne par classe de lithologie », présente le double avantage d'être exprimée dans la même échelle que les teneurs en HT et de maximiser la corrélation avec ces dernières (par rapport à un codage arbitraire en catégories 1, 2 et 3 par exemple, voir §2.3.2).

Cette relation rend pertinente l'utilisation de l'information lithologique, connue de façon plus dense, pour affiner l'estimation des teneurs en hydrocarbures totaux. Néanmoins, les seules analyses en HT disponibles dans les remblais proviennent de la tache d'hydrocarbures, aucune analyse n'ayant été réalisée à plus faible profondeur. Par conséquent, utiliser la connaissance des remblais pour guider l'estimation conduirait à surestimer les teneurs en hydrocarbures totaux pour l'ensemble des remblais, y compris ceux à faible profondeur. Cette limitation de l'utilisation exclusive de l'information lithologique sera résolue par l'intégration des indices organoleptiques.

Indices organoleptiques

Outre la description lithologique, chaque zone de carottage, échantillonnée ou non, a fait l'objet d'une observation visuelle et olfactive d'hydrocarbures. Ces deux indices fournissent une information similaire par rapport aux teneurs en hydrocarbures : les analyses réalisées sur les prélèvements sans indication olfactive ou visuelle de contamination ont une teneur moyenne en HT égale à 50 mg/kg ms ; de même, en cas d'indication olfactive de pollution, la teneur moyenne en HT issue des analyses disponibles est égale à 2500 mg/kg ms.

Combinaison des informations indirectes

Les deux informations indirectes présentées (lithologique et olfactive), riches en terme d'information sur la teneur supposée en HT en des points non analysés, gagnent à être combinées.

L'information olfactive est privilégiée par rapport à l'information lithologique ; cette dernière n'est utilisée que pour affiner, dans le cas où il y a indication olfactive de contamination, la pollution attendue en fonction de la lithologie.

Cela a pour effet de ramener à 50 mg/kg ms les moyennes par classe de lithologie élevées dues aux remblais ou aux argiles graveleuses en cas d'absence d'indication olfactive de contamination.

3.3 Structure spatiale bivariable

Dans le cas de distributions très dissymétriques, le recours à une transformation de type logarithmique est fréquent. L'analyse de la structure spatiale et l'estimation sont menées sur la transformée Y=log([HT]+1) de la concentration en hydrocarbures totaux. Une transformation inverse est ensuite appliquée au résultat de l'estimation.

L'estimation de teneurs en polluant ne nécessite aucune hypothèse sur la distribution de la variable d'intérêt (voir thème N°1). Néanmoins, l'étude de variables ayant une distribution très dissymétrique présente souvent des problèmes de caractérisation de la structure spatiale et d'estimation. Plusieurs types de transformation sont envisageables : logarithmique, gaussienne. L'objectif de cette illustration étant uniquement la cartographie tridimensionnelle de la contamination, il n'est pas nécessaire de se ramener à une variable précisément gaussienne (cela serait le cas si des calculs de volumes, nécessitant la mise en œuvre de techniques non linéaires, étaient présentés)

Comme nous désirons intégrer dans la modélisation des teneurs en HT la variable auxiliaire construite à partir des informations indirectes combinées, il nous faut calculer non seulement la structure spatiale de la variable d'intérêt, mais également celle de la variable auxiliaire ainsi que la structure spatiale croisée entre les deux variables, qui prend en compte leur niveau de corrélation. La Figure III-8 présente les variogrammes simples et croisé ainsi obtenus, ainsi que l'ajustement de modèle réalisé. C'est la connaissance de ces trois variogrammes qui permet de déterminer les poids de la combinaison linéaire conduisant à l'estimation par cokrigeage, en chaque bloc, de la transformée logarithmique de la teneur en HT à partir des mesures de teneur en HT et des points d'observation de la variable auxiliaire.

Le modèle de variogramme bivariable est constitué de deux structures :

- un effet de pépite, lié à l'hétérogénéité des teneurs dans les remblais,
- une structure de type sphérique, liée à l'existence d'une structure spatiale pour la contamination ; cette structure présente une anisotropie, la structure spatiale n'ayant pas la même portée dans la direction verticale qu'horizontalement.

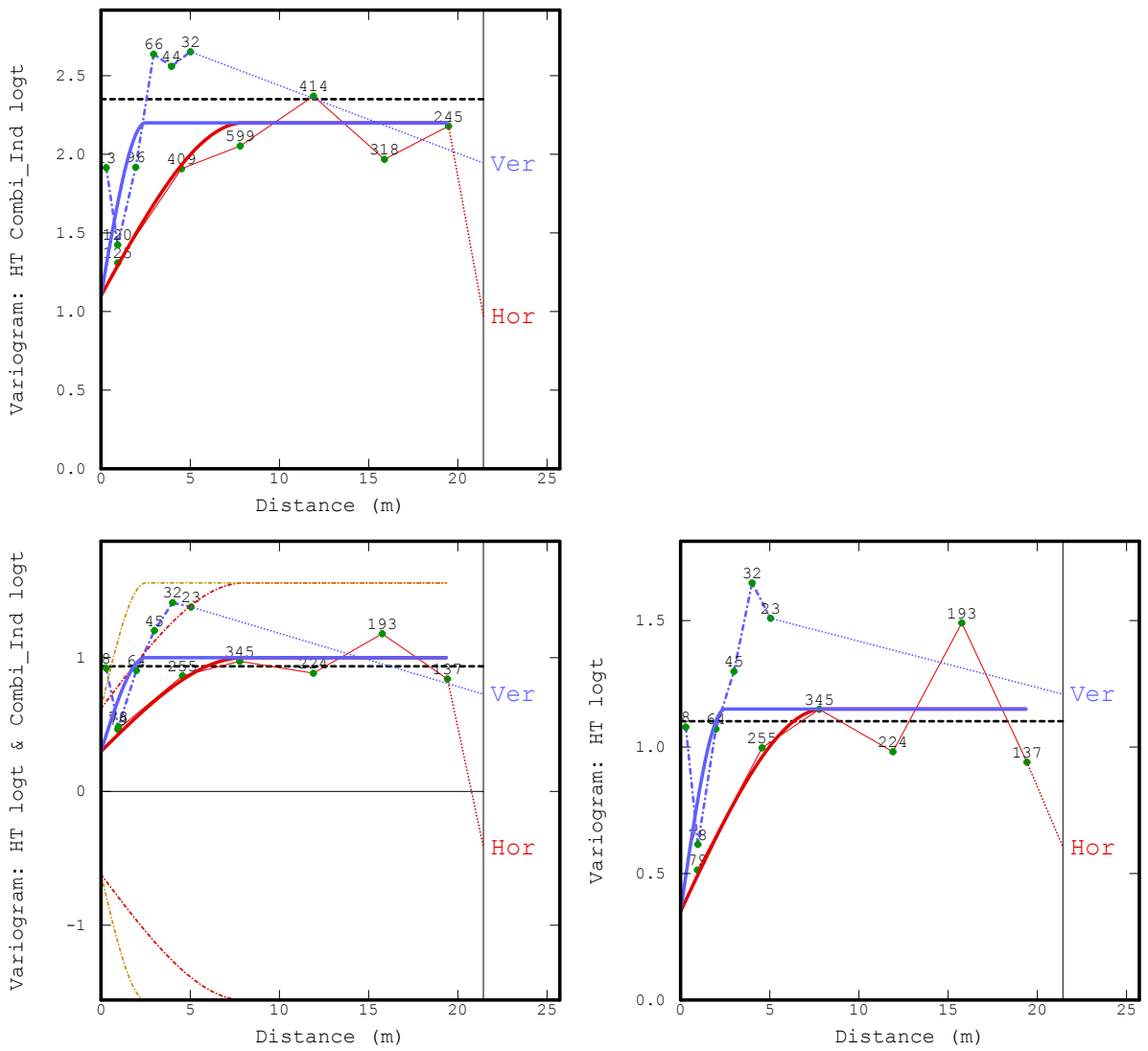


Figure III-8 Variogrammes expérimentaux (trait fin) et modèles ajustés (trait épais) pour la transformée logarithmique de la teneur en HT et de la combinaison des informations qualitatives. Direction horizontale en rouge, verticale en bleu, nombre de paires indiqué pour chaque pas de calcul, variance expérimentale (tirets noirs).

Une analyse géostatistique menée pour chaque type de lithologie présenterait l'avantage de permettre un travail sur des populations plus homogènes. Cette analyse nécessiterait cependant une bonne connaissance des limites entre les différentes géologies rencontrées, ce qui n'est pas le cas ici.

3.4 Cartographie des teneurs en hydrocarbures totaux

La cartographie est réalisée :

- d'une part, à partir des seules analyses HT disponibles,
- d'autre part, en intégrant sous forme de variable auxiliaire les informations indirectes.

La Figure III-9 illustre clairement les différences entre krigeage et cokrigeage, en particulier dès que l'on se situe en extrapolation par rapport aux analyses en hydrocarbures totaux. L'intégration des informations indirectes corrélées à la teneur permet alors d'en guider l'estimation et de réduire l'incertitude qui lui est associée.

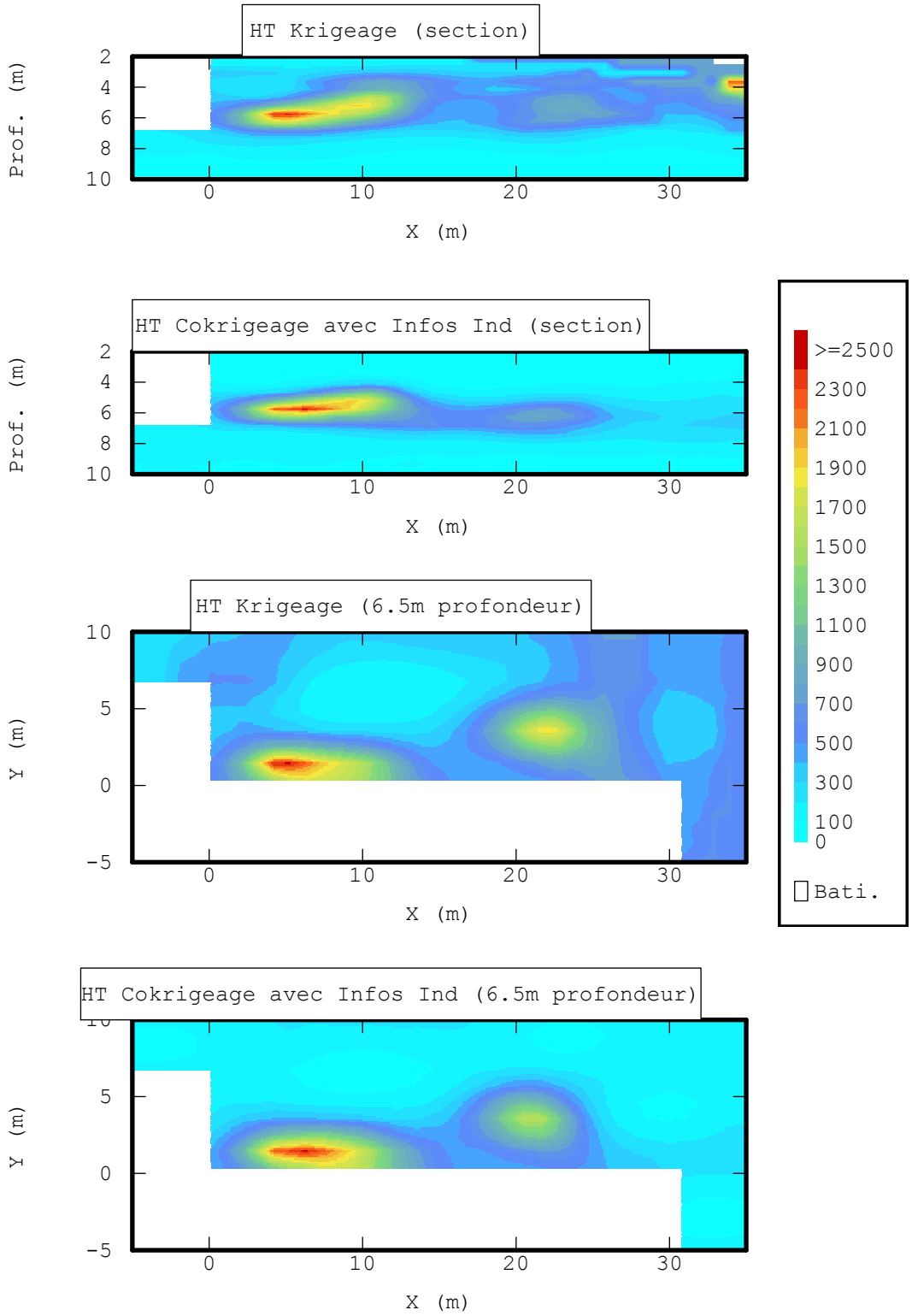


Figure III-9 Vues en section et du dessus de la répartition spatiale des teneurs en hydrocarbures totaux : comparaison entre krigeage et cokrigeage avec les informations indirectes. Les zones non estimées correspondent aux bâtiments.

Grâce à l'intégration des informations indirectes, la cartographie permet de localiser les zones de fortes teneurs, tout en ayant conscience de l'erreur d'estimation possible, notamment en bordure de la zone et dans les zones sous-échantillonnées, en extrapolation.

La Figure III-10, illustration du résultat obtenu par cokrigeage, apporte les informations suivantes sur l'état de contamination des sols :

- la tache en amont (X ~ 5m), à partir de 4.5 m de profondeur, s'interprète logiquement comme la fuite d'hydrocarbures depuis la mazouterie,
- la tache à droite (X ~ 13m), un peu plus profonde et moins concentrée que la tache en amont, s'interprète comme le résultat de l'écoulement gravitaire depuis la tache amont vers l'Est, les terrains accusant un pendage à cet endroit.

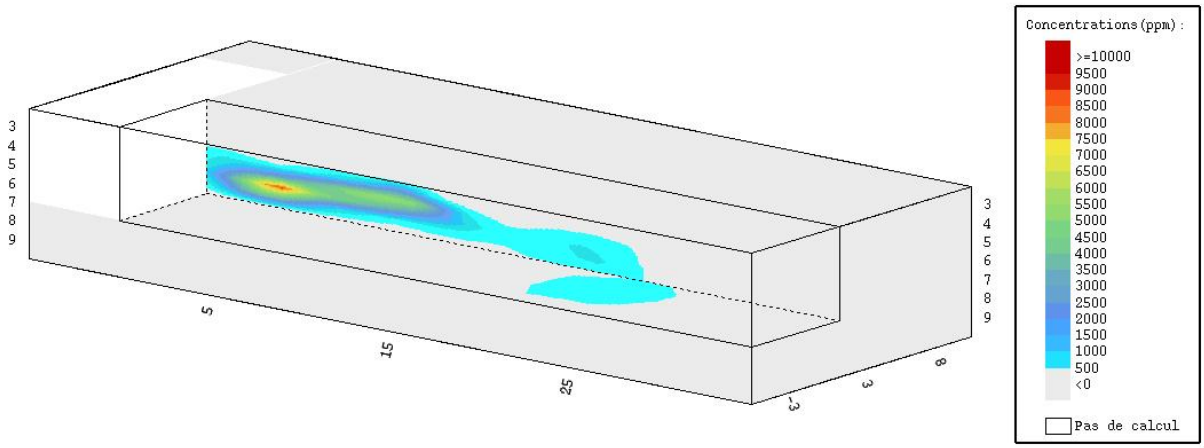


Figure III-10 Visualisation 3D des principales taches d'hydrocarbures, au Nord du bâtiment et en dessous de la dalle.

3.5 Conclusions

L'objectif de l'illustration consistait à représenter l'extension de la zone contaminée en hydrocarbures au nord du bâtiment chaufferie, à partir d'analyses en hydrocarbures totaux réalisées sur des échantillons de sols provenant de 23 carottages (de 1 à 3 prélèvements par carottage) et d'informations organoleptiques et lithologiques, peu coûteuses et par conséquent plus nombreuses.

L'analyse exploratoire des données a mis en évidence une corrélation entre la teneur en HT et les informations qualitatives dont l'exploitation par une approche géostatistique pouvait se justifier. Les variogrammes montrent une structure spatiale nette du phénomène. L'intégration des informations auxiliaires enrichit à moindre coût et de façon significative la modélisation des teneurs en hydrocarbures totaux.

La géostatistique permet ainsi d'établir, grâce à l'intégration de l'ensemble des informations disponibles, une cartographie tridimensionnelle cohérente avec l'origine de la pollution : fuite de fuel depuis la mazouterie puis écoulement vers l'est le long des couches pentées, cet écoulement étant limité en profondeur par la présence d'une couche de marnes jaunes imperméables.

IV THÈME N°4 : QUEL EST L'APPORT DE LA GÉOSTATISTIQUE À L'ÉCHANTILLONNAGE DES SITES POLLUÉS ?

Quel est l'apport de la géostatistique à l'échantillonnage des sites pollués ?

Fiche technique N°4

NOUS SOMMES TOUS CONFRONTES A CE PROBLEME

Un site pollué peut être **échantillonné** à différentes étapes de sa caractérisation ou réhabilitation, **pour répondre à des objectifs variés** :

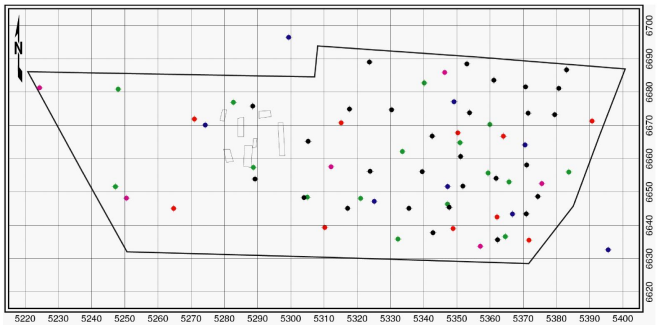
- Mise en évidence et identification d'une pollution des sols au cours d'une ESR.
- Description détaillée de sources-vecteurs-cibles au cours d'une EDR.
- Localisation et estimation de quantités de matériaux à dépolluer.
- Dimensionnement d'un projet de dépollution.
- Tri des matériaux en cours de chantier de dépollution.
- Contrôle de l'état de pollution résiduelle après réhabilitation.
- Suivi de la dégradation des polluants en cas de dépollution in situ ou en tas.
- Suivi de l'atténuation naturelle d'une pollution.

Dans tous les cas, se pose le problème de la **stratégie d'échantillonnage** en terme de **« comment atteindre un niveau de connaissance suffisant, par rapport à des objectifs d'étude, avec un minimum de données positionnées au mieux ? »**. Parler de stratégie d'échantillonnage signifie alors aussi **prendre en compte des contraintes financières ou de délais**, décider de tailles d'échantillon adaptées aux objectifs d'étude, pouvoir programmer des phases successives d'échantillonnage pour mieux lever les incertitudes ou encore pouvoir intégrer l'ensemble des données existantes pour décider du nombre et de l'emplacement d'échantillons complémentaires.

LES PRATIQUES ACTUELLES

Les stratégies d'échantillonnage sont issues le plus souvent d'une réflexion des praticiens faisant appel à leur bon sens, à leur connaissance d'un site ainsi qu'à leur expérience. Les critères de choix demeurent cependant empiriques et dictés par des motivations autres que de réduction des incertitudes ou des erreurs par rapport à des objectifs d'étude : limitation des coûts d'analyse, reconnaissance préférentielle de zones sources ou jugées « lieux d'anomalies », sélection d'échantillons visuellement suspects, etc.

Ces pratiques ne sont pas toujours satisfaisantes comme l'ont montré les retours d'expérience sur des sites pollués déjà traités. Des sols contaminés ou sains sont souvent découverts à des endroits où ils n'étaient pas attendus faute d'avoir cherché à comprendre et quantifier les incertitudes.



CE QUE NOUS APPORTE LA GEOSTATISTIQUE

En pratique, il est attendu d'une stratégie d'échantillonnage d'établir le nombre et la localisation des échantillons à analyser. Cette détermination passe par la définition de **critères d'optimisation** nécessairement liés à des objectifs d'étude. Il peut s'agir par exemple **de minimiser :**

- **l'incertitude sur les teneurs** en polluant pour en évaluer le stock,
- **l'incertitude sur les volumes** à dépolluer,
- le risque de laisser en place des sols contaminés.

Selon le critère retenu, **le principe consiste à quantifier une incertitude ou un risque à partir d'un ensemble de données disponibles pour en déduire un plan d'échantillonnage optimal pour des analyses complémentaires**.

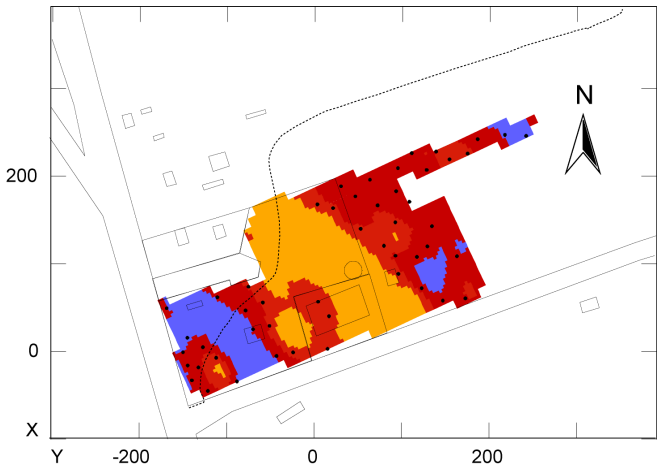
Etant établi que l'échantillonnage vise à réduire des incertitudes spatiales, **la géostatistique fournit les outils et les résultats nécessaires à une rationalisation rigoureuse et systématique de programmes d'échantillonnage.** Quantification des incertitudes d'estimation, prise en compte de corrélations spatiales ou entre variables, intégration de données hétérogènes ou de natures diverses, correction de biais d'échantillonnage sont autant d'apports de la géostatistique pour appréhender cette problématique.

Pour en savoir plus…

Concernant ce thème, le manuel méthodologique ***GeoSiPol*** présente une approche géostatistique standard pour rationaliser le nombre et la localisation d'échantillons en fonction des données disponibles et des objectifs d'étude.

Elle se focalise sur **l'élaboration d'une campagne d'échantillonnage complémentaire** pour des analyses classiques en laboratoire. Les aspects suivants y sont abordés.

- Types de données disponibles lors des diagnostics, établissement de recommandations sur les pratiques de l'échantillonnage.
- **Exploitation de la totalité des informations disponibles pour évaluer les incertitudes.**
- Choix du positionnement des échantillons en fonction des incertitudes.
- **Eléments de réponse aux questions que soulève fréquemment l'échantillonnage.**



L'approche est appliquée au cas d'une **ancienne décharge sauvage** lieu d'une contamination des sols par des **hydrocarbures aromatiques polycycliques** et des **métaux lourds**. Une campagne d'échantillonnage complémentaire destinée à améliorer la délimitation de la zone à dépolluer est définie sur la base d'une étude géostatistique.

Quel est l'apport de la géostatistique à l'échantillonnage des sites pollués?

1 PRESENTATION DU THEME

1.1 Introduction

Un site potentiellement pollué est amené à être échantillonné à différentes étapes de sa caractérisation ou réhabilitation, pour répondre à des objectifs variés :

- Mise en évidence et identification d'une pollution des sols au cours d'une ESR[18].
- Description détaillée des sources-vecteurs-cibles au cours d'une EDR[19].
- Localisation et estimation de quantités de matériaux à dépolluer.
- Dimensionnement d'un projet de dépollution.
- Tri des matériaux en cours de chantier de dépollution.
- Contrôle de l'état de pollution résiduelle après réhabilitation.
- Suivi de la dégradation des polluants en cas de dépollution in situ ou en tas.
- Suivi de l'atténuation naturelle d'une pollution.

Dans tous les cas, se pose le problème de la **stratégie d'échantillonnage** en terme de **« comment atteindre un niveau de connaissance suffisant, par rapport à des objectifs d'étude, avec un minimum de données positionnées au mieux ? »**. Parler de stratégie d'échantillonnage signifie alors aussi prendre en compte des contraintes financières ou de délais, décider de tailles d'échantillon adaptées aux objectifs d'étude, pouvoir programmer des phases successives d'échantillonnage pour mieux lever les incertitudes ou encore pouvoir intégrer l'ensemble des données existantes pour décider du nombre et de l'emplacement d'échantillons complémentaires.

A l'heure actuelle, les stratégies d'échantillonnage résultent le plus souvent d'une réflexion des praticiens basée sur leur expérience, leur connaissance du site ainsi que leur bon sens. De telles stratégies « au jugement » ne permettent pas d'éviter l'« oubli» de certaines sources de contamination, notamment lorsque les informations documentaires sur la contamination sont très lacunaires. Elles n'empêchent pas non plus de sous-estimer (ou plus rarement de surestimer) l'ampleur d'une contamination. Le besoin apparaît ainsi de procédures d'aide à la conception de plans d'échantillonnage qui réduisent au mieux les incertitudes d'estimation par rapport à des objectifs d'étude tout en respectant des impératifs budgétaires. L'objectif de ce chapitre est de montrer quels peuvent être les apports de la géostatistique à la **rationalisation de l'échantillonnag**e des sites pollués. Après avoir répondu aux **questions les plus fréquentes** des professionnels sur l'échantillonnage, une démarche simple mettant en œuvre des méthodes géostatistiques est proposée pour aider à **adapter le nombre et la localisation** de points d'échantillonnage aux objectifs d'étude.

18 Evaluation Simplifiée des Risques

19 Evaluation Détaillée des Risques

1.2 Cadre réglementaire

Le guide méthodologique national français[20] sur les sites pollués fournit quelques recommandations très générales sur les stratégies d'échantillonnage ainsi que sur les conditions d'échantillonnage des sols.

Le guide rappelle qu'aucune démarche systématique ne peut être proposée, chaque programme d'échantillonnage devant être adapté aux caractéristiques du site et aux objectifs d'étude. Il est par ailleurs reconnu la difficulté de définir une stratégie d'échantillonnage lorsque l'information sur la contamination est très réduite. Tel est généralement le cas des ESR dans le cadre desquelles un programme d'échantillonnage doit être établi avec un nombre réduit de points. Une localisation des points « au jugement », c'est-à-dire orientée par la connaissance de l'historique de l'activité industrielle, de l'inspection visuelle du site et d'éventuelles campagnes de reconnaissance déjà réalisées, est recommandée, sans pour autant écarter les autres alternatives.

La méthodologie française ne fournit pas de recommandations précises sur l'échantillonnage aux autres étapes de la caractérisation et de la réhabilitation. Le guide méthodologique rappelle les différentes approches possibles de localisation des points et renvoie aux recommandations élaborées par l'US EPA en 1991, et reprises dans le Tableau IV-1 ci-dessous, pour adapter les stratégies aux objectifs d'étude.

Tableau IV-1 Evaluation des stratégies d'échantillonnage en fonction de l'objectif (US EPA, 1991).

Stratégie / Objectif	Aléatoire	Systématique (Aux nœuds d'une grille)	Systém. Aléatoire (Aléatoire par maille)	Recherche spécifique (ex. points source, limite de zone contaminée)	Aléatoire stratifié (Aléatoire par secteur)	Profils (Coupes transver-sales)	Jugement
Evaluation du risque	4	2[A]	3	3	3	2	1
Identification des sources	4	2[A]	3	2	2	3	1
Délimitation de l'extension	3	1[B]	1	1	3	1	4
Perspectives de décontamination	3	2	2	4	1	2	3
Contrôle de la décontamination	1[C]	1[B]	1	1	3	1[D]	4

Stratégie :

1 : la plus adaptée

2 : possible

3 : peu adaptée

4 : la moins adaptée

Commentaires :

A : à appliquer avec une technique de criblage analytique de terrain

B : adaptée seulement si des tendances sont connues

C : validation statistique si l'ensemble du site est couvert

D : possible si le site est supposé propre (plusieurs techniques)

Au final, nous constatons que le guide méthodologique, en restant volontairement très évasif sur les stratégies d'échantillonnage à adopter, laisse la liberté à chacun de s'adapter aux caractéristiques et objectifs d'étude spécifiques à chaque site mais, dans le même temps, rend difficile la tâche des praticiens.

1.3 Pratiques habituelles

Dans la pratique, on retrouve une définition « au jugement » des plans d'échantillonnage. Les critères de choix apparaissent souvent empiriques et dictés par d'autres motivations que la réduction des

[20] Guide méthodologique « Gestion des sites (potentiellement) pollués », Ministère de l'Environnement, Version 2, Annexe 6, révision du 9 décembre 2002.

incertitudes ou des erreurs d'estimation par rapport à des objectifs d'étude tels que l'évaluation des risques sanitaires ou la classification des sols par filière de traitement/revalorisation. Le nombre de points d'échantillonnage découle généralement de l'enveloppe budgétaire que le responsable du site accorde aux investigations. Les zones sources ou jugées « lieux d'anomalies » sont souvent échantillonnées de façon préférentielle.

Ces pratiques ne sont pas toujours satisfaisantes comme l'ont montré les retours d'expérience sur des sites pollués déjà traités. Des sols contaminés ou sains (pour une définition de sain et contaminé, voir thème N°2, §II1.2, p.51) sont souvent découverts à des endroits où ils n'étaient pas attendus faute d'avoir cherché à comprendre et quantifier les incertitudes.

1.4 Intérêt d'une approche géostatistique

Etant établi que l'échantillonnage vise à réduire des incertitudes spatiales, la géostatistique fournit les outils et les résultats nécessaires à une **rationalisation rigoureuse et systématique des programmes d'échantillonnage**. **La quantification des incertitudes d'estimation, la prise en compte de corrélations spatiales ou entre variables, l'intégration de données hétérogènes ou de natures diverses et la correction de biais d'échantillonnage sont autant d'apports de la géostatistique pour appréhender cette problématique.**

La géostatistique permet aussi de **valider ou à l'inverse de mettre à défaut des pratiques d'échantillonnage** qui ne sont pas sans conséquences sur la qualité des estimations. Ainsi, certaines sous-populations, observées sur des données de teneurs et associées à des statistiques particulières, peuvent être interprétées comme la trace d'échantillons de tailles différentes. Il est impératif de pouvoir identifier et distinguer de telles sous-populations pour estimer plus précisément et plus justement les teneurs dans les sols. Il en est de même de la constitution des échantillons et de l'analyse chimique qui sont aussi sources de disparité sur les mesures de teneurs. **Reconnaître la part de variabilité des teneurs due à l'échantillonnage ou à la mesure de celle présente « naturellement » dans les sols contribue à mieux concevoir un plan d'échantillonnage et par suite à mieux estimer les teneurs.**

1.5 Objectifs

Les étapes de diagnostic et d'évaluation des risques d'un site potentiellement contaminé sont soumises à de fortes contraintes financières. Notre but n'est donc pas d'exiger des conditions d'échantillonnage idéales qui nous permettraient d'utiliser une démarche géostatistique. Au contraire, il s'agit d'**utiliser les capacités de la géostatistique pour extraire le maximum d'information des données collectées dans le cadre d'un effort d'investigation réaliste**.

L'objectif principal est de présenter une démarche géostatistique simple qui puisse aider à **rationaliser un plan d'échantillonnage en termes de nombre et de localisation des points**. Nous nous plaçons dans la situation où l'échantillonnage est réalisé par étapes. La démarche géostatistique intervient pour établir un **plan d'échantillonnage complémentaire** à partir d'informations collectées lors de précédentes campagnes de reconnaissance.

La rationalisation de l'échantillonnage passe par la définition de **critères d'optimisation** nécessairement **liés aux objectifs d'étude**. Il peut s'agir par exemple **de minimiser :**

- **l'incertitude sur les teneurs** en polluant pour en évaluer le stock,
- **l'incertitude sur les volumes** à dépolluer,
- **le risque** de laisser en place des sols contaminés.

Selon le critère retenu, la démarche consiste alors à quantifier une incertitude ou un risque à partir d'un ensemble de données disponibles pour en déduire un plan d'échantillonnage optimal pour des analyses complémentaires.

Des recommandations générales sur les conditions d'échantillonnage qui permettent d'augmenter la précision d'estimation, à nombre égal de points d'échantillonnage, sont de plus fournies au § 2.

Le chapitre est limité au cas de l'échantillonnage du sol et du sous-sol. Nous n'abordons donc pas la question de l'échantillonnage des eaux souterraines, des eaux de surface ou de l'air interstitiel du sol. De même, ce chapitre ne traite pas d'autres questions soulevées par l'échantillonnage, par exemple pour savoir « quelles analyses effectuer sur les échantillons » ou « comment constituer un échantillon ».

2 QUESTIONS LES PLUS FRÉQUENTES SUR L'ÉCHANTILLONNAGE

Nous ne prétendons pas répondre de manière exhaustive à l'ensemble des questions que se posent les praticiens sur l'échantillonnage. Mais nous souhaitons fournir certains éléments de réponse à des questions revenant constamment et sur lesquelles les praticiens se sentent démunis.

2.1 Combien d'échantillons pour une approche géostatistique ?

La question qui nous est posée le plus systématiquement concerne le nombre de données, en terme de « combien faut-il disposer de données pour pouvoir mettre en œuvre une démarche géostatistique ». Nous ne pouvons répondre de manière précise à cette question, car le nombre de données doit être adapté au phénomène étudié et aux objectifs d'étude, mais nous pouvons tout de même fournir certaines indications. **Il est clair que nous ne pouvons prétendre à une étude géostatistique (ou statistique) si nous ne disposons que d'un nombre de données très réduit, disons inférieur à 10**. A l'inverse, cependant, une idée fausse et largement répandue est de penser qu'une démarche géostatistique ne peut pas être mise en œuvre si nous ne disposons pas de plusieurs centaines d'échantillons. Entre ces deux extrêmes, le nombre de données « nécessaires » à une étude géostatistique dépend de plusieurs facteurs.

- **La qualité des données disponibles**. De « bonnes » pratiques d'échantillonnage sont essentielles pour que la totalité des données disponibles puisse être valorisée. Par exemple, il a été démontré sur un ancien site d'usine à gaz que seules 36 données sur 202 collectées (soit 18%) étaient exploitables ensemble, sans nécessiter d'hypothèses ni d'approximations. Un lien a pu être fait entre les hétérogénéités observées sur les 202 données et les pratiques d'échantillonnage.
- **La taille du domaine d'étude, sa dimension (1D, 2D ou 3D) et l'échelle d'étude du phénomène**. Nous ne définirons pas le même plan d'échantillonnage pour estimer les teneurs en polluant dans des sols au niveau régional (« bruit de fond ») ou pour estimer l'état de contamination des sols d'une friche industrielle de 1 ha (pollution locale).
- **La structuration du phénomène spatio-temporel étudié.** Dans le cas extrême où le phénomène est purement aléatoire à l'échelle considérée (effet de pépite pur, voir définition au thème N°1, § I2.3.7, page 29), augmenter le nombre de données collectées n'améliorera pas la précision d'estimation. Dans le cas extrême opposé où le phénomène est très structuré (grande continuité), un objectif fixé peut être atteint avec un nombre réduit de données. Entre ces deux cas, une démarche géostatistique de rationalisation de l'échantillonnage telle que décrite au § 3 a tout son sens.
- **Les caractéristiques des données disponibles.** Rappelons que par **données** nous désignons non seulement les informations sur les variables d'intérêt, que sont les teneurs en polluants mesurées au laboratoire ou sur site, mais aussi d'autres informations qualitatives (ex. critères organoleptiques), semi-quantitatives ou quantitatives (ex. mesures géophysiques) qui peuvent renseigner indirectement sur l'état de contamination des sols. Lorsque des corrélations sont mises en évidence entre les teneurs et des informations indirectes, ces dernières peuvent être exploitées pour réduire encore le nombre de données de teneurs qui sont généralement coûteuses. Nous renvoyons au thème N° 3 sur ce sujet.

2.2 Quelles conditions de prélèvement respecter ?

Nous rappelons certaines recommandations qui facilitent l'exploitation, la comparaison et plus généralement la valorisation des données dans le cadre ou non d'une démarche géostatistique.

L'échantillonnage doit respecter les deux règles essentielles suivantes.

- Les prélèvements doivent être effectués **systématiquement** en respectant les mêmes procédures et les mêmes règles de constitution et de sélection des échantillons d'un sondage à l'autre. C'est à cette condition que les données pourront être exploitées pour reconnaître au mieux des états de contamination différents des sols d'un point à l'autre du site ou pour évaluer les teneurs à des endroits non échantillonnés. Dans une approche géostatistique, de la cohérence des données dépendra la fiabilité des statistiques et des corrélations qui pourront être établies sur ou avec les teneurs mesurées.
- Chaque prélèvement doit être **renseigné le plus précisément possible**. Pour chaque échantillon, un protocole de prélèvement doit indiquer de manière précise les conditions de prélèvement, la nature des terrains rencontrés, etc. L'objectif est de limiter la perte d'information entre l'opérateur sur le terrain et la personne chargée de l'interprétation des résultats. La documentation doit inclure notamment le protocole opératoire appliqué à tous les échantillons, le protocole de prélèvement par échantillon, la carte avec la localisation (précise) des forages et des échantillons prélevés et si possible la photographie des échantillons prélevés (totalité de l'échantillon, zoom sur texture particulière) et les coupes de forage.

Il est recommandé par ailleurs d'appliquer les règles de base suivantes.

- Prendre des mesures pour **éviter les contaminations croisées** : nettoyage des outils de prélèvement entre chaque échantillon, nettoyage des tubes de forage à l'eau propre entre chaque forage, pas de fluide de forage, diamètre de forage suffisant, prélèvement d'échantillon après élimination des terrains du pourtour de la carotte, ceux-ci étant susceptibles d'avoir été mélangés les uns aux autres sur la verticale.
- Prélever une quantité suffisante de matériaux par prélèvement.
- Prévoir un double à chaque prélèvement, à conserver à 4°C à l'obscurité. Le double peut être obtenu par un quartage de l'échantillon, technique de séparation la plus basique, ou bien par une technique plus élaborée telle que celle faisant appel à un diviseur à flux tombant.
- Prévoir un conditionnement (flaconnage) adapté à la nature des terres et de la contamination.
- Ne pas enlever de matériaux. Ne pas cribler. Si des matériaux doivent être enlevés, l'indiquer sur le protocole de prélèvement.

2.3 Comment échantillonner sur la verticale?

Il existe de nombreuses variantes pour échantillonner le sol et le sous-sol le long d'un sondage (le terme « sondage » désignant ici indifféremment sondage à la tarière, sondage carotté ou bien encore fosse à la pelle mécanique). Les échantillons peuvent être « ponctuels » ou « composites », collectés à des profondeurs régulières par passes constantes ou prélevés par lithologie. Verticalement, le mode d'échantillonnage adopté dépendra de l'objectif de la campagne d'échantillonnage. Il sera différent selon qu'il s'agit d'identifier les niveaux contaminés d'un site, de caractériser la pollution par lithologie ou bien encore d'évaluer un volume de sols à traiter. Notons que le problème du mode d'échantillonnage est très général et concerne aussi l'extension latérale des échantillons. Nous renvoyons au thème N°1, §I2.2.4, page 21 où la question du **suppor**t de l'information est abordée.

2.4 Faut-il faire des analyses précises, mais chères et donc peu nombreuses, ou des analyses moins précises, mais peu coûteuses et donc plus nombreuses ?

Les informations sur l'état de contamination d'un site se composent d'une part d'analyses au laboratoire ou de mesures rapides sur site sur échantillons de sols, d'autre part de données indirectes

telles que l'historique de l'activité industrielle, des plans d'implantation des installations industrielles, des coupes lithologiques, ou bien encore des mesures géophysiques. Ces données indirectes peuvent d'une façon ou d'une autre être rapprochées aux mesures directes et ainsi renseigner sur les teneurs en polluant dans les sols. Les premières sont précises mais coûteuses et donc généralement peu nombreuses. Les deuxièmes moins précises mais aussi moins onéreuses peuvent facilement être récoltées en de nombreux endroits et donc offrir un meilleur recouvrement du site. Il n'est actuellement pas possible d'envisager un diagnostic ou une dépollution sans analyse en laboratoire, les valeurs-limites établies par arrêtés préfectoraux de diagnostic ou de dépollution s'appliquant généralement à des mesures chimiques de ce type. Les données indirectes peuvent cependant être valorisées pour réduire le nombre d'analyses chimiques effectuées sur un site ou pour améliorer la précision d'estimation. Nous renvoyons au thème N°3 qui présente les conditions de valorisation de ce type de données.

2.5 Comment procéder pour que des données collectées dans le cadre des évaluations de risques puissent être utilisées aussi pour une quantification des volumes à traiter ? Un échantillonnage spécifique doit-il être envisagé ?

Les données collectées dans le cadre des évaluations de risques sanitaires sont destinées à fournir l'ensemble des informations nécessaires à la caractérisation des sources, des vecteurs et des cibles de la contamination. Leur utilisation courante pour délimiter les zones à dépolluer et en évaluer les volumes résulte d'un souci d'économie, les projets de revalorisation étant généralement soumis à de fortes contraintes financières qui tendent à interdire l'acquisition de nouvelles données.

Il n'y a pas d'objection à utiliser les données d'évaluation des risques pour positionner et évaluer des volumes de sols à traiter. La valorisation de ces données peut simplement se révéler plus délicate et les résultats d'étude moins précis que si les données avaient été collectées dans le but de délimiter et quantifier des volumes. Le choix d'un mode d'échantillonnage (support), d'un nombre d'échantillons ou de leurs emplacements peut ainsi favoriser ou non la précision des estimations.

Pour garantir la cohérence des données, et donc la qualité des résultats qui en sont tirées avec ou sans géostatistique, nous conseillons de respecter de façon systématique et dès le diagnostic initial les règles de prélèvement des échantillons présentées au § 2.2.

Si les données disponibles issues de l'évaluation des risques s'avèrent être finalement insuffisantes pour positionner les zones à dépolluer (grandes incertitudes), l'approche géostatistique décrite dans la suite peut être suivie pour proposer un plan d'échantillonnage complémentaire adapté à l'objectif de dépollution des sols.

2.6 Doit-on envisager un échantillonnage préférentiel de certaines parties du site ?

La question de la nécessité ou non de mener une reconnaissance plus serrée de certaines parties du site peut se poser à différents stades de l'évaluation des risques. « Faut-il resserrer la maille d'échantillonnage au niveau des sources de contamination », « faut-il densifier les points de prélèvement dans les zones où les teneurs mesurées sont proches des seuils critiques » sont des questions qui se posent fréquemment. Il n'est pas possible ici de fournir de règle générale à ce sujet, les réponses dépendant à la fois des caractéristiques du site et des objectifs d'étude. Il apparaît clairement que des resserrements locaux peuvent apporter des compléments d'information dans certaines situations mais pas de façon systématique. Quelques resserrements locaux peuvent être utiles à la définition du comportement du variogramme[21] (variabilité spatiale) à petite distance, information essentielle à une modélisation géostatistique qui permet d'étudier le degré de continuité et de régularité spatiale de la contamination. Notons cependant que ces resserrements peuvent signifier des biais d'échantillonnage, certaines *populations* étant plus échantillonnées que d'autres. Ces biais ne sont pas sans effets sur les statistiques calculées sur les données et doivent être corrigées par une procédure appropriée (dégroupement).

[21] Se référer au § I2.3.5, p.27, pour une définition du variogramme.

2.7 Suffit-il d'échantillonner à la maille définie par la portée du variogramme ?

La portée du variogramme caractérise la distance jusqu'à laquelle les teneurs mesurées en deux points sont corrélées. Aussi recommande-t-on parfois de prendre pour maille d'échantillonnage une maille égale à la portée. Si cette recommandation paraît tout à fait conforme à l'intuition, il faut pourtant lui tordre le cou. On s'en rend aisément compte dans les cas extrêmes : si la portée est très grande, par exemple beaucoup plus grande que le site étudié, un seul sondage suffirait ; à l'inverse, si les teneurs sont très erratiques, ce qui se traduit par une portée très courte, il faudrait un nombre extrêmement élevé de sondages alors qu'en fait on est ramené en ce cas aux statistiques classiques qui n'exigent pas tant de données.

En fait les formules géostatistiques permettent de dire quelle précision pourra être atteinte avec telle maille d'échantillonnage. Si on se fixe la précision à atteindre, on obtient la maille d'échantillonnage en inversant ces formules. La réponse dépend de l'ensemble du variogramme et pas seulement de sa portée. On conçoit par exemple qu'à portée identique il faudra plus de sondages si l'amplitude des variations, mesurée par le palier, est grande.

3 APPROCHE GEOSTATISTIQUE DETAILLEE

3.1 Méthodologie

L'approche géostatistique proposée pour rationaliser l'échantillonnage comporte **cinq étapes** principales. Elles peuvent être appliquées dans le cadre d'une reconnaissance en phases pour définir une phase d'échantillonnage complémentaire, sur la base des informations déjà collectées :

- Analyse exploratoire des données disponibles.
- Définition de critères de rationalisation de l'échantillonnage (ex. minimisation de l'incertitude sur les teneurs ou les volumes).
- Estimation des incertitudes intervenant dans la définition des critères d'optimisation de l'échantillonnage. Approche par estimation ou par simulation selon le cas.
- Calcul d'un plan d'échantillonnage complémentaire rationalisé.
- Production de résultats pour l'aide à la décision.

3.2 Analyse exploratoire des données

Nous renvoyons au thème N°1 qui présente la ligne de conduite à tenir pour mener une analyse exploratoire des données. Ce chapitre se limite à l'analyse d'une seule variable d'intérêt, constituée le plus fréquemment de la teneur en un polluant. Nous renvoyons au thème N°3 pour ce qui est des spécificités de l'analyse exploratoire dans le cas où des informations indirectes sont à exploiter. Nous conseillons de porter un intérêt particulier aux paragraphes concernant le **support** des échantillons (voir §I2.2.4, page 21), c'est-à-dire leur volume et leur géométrie.

3.3 Définition de critères de rationalisation de l'échantillonnage

Les critères de rationalisation de l'échantillonnage dépendent des objectifs d'étude visés par la reconnaissance. Nous présentons ici quelques exemples de critères, la liste n'étant pas exhaustive.

3.3.1 Minimiser l'incertitude sur les teneurs

Un échantillonnage complémentaire peut servir à évaluer le stock de polluants (répartition spatiale de quantités de polluants) qui interviendra comme terme source dans un modèle d'écoulement et de transport ou dans un modèle de risque sanitaire. Il s'agit alors de **minimiser l'incertitude sur les teneurs en polluants** à l'intérieur du domaine, point par point (incertitude locale).

3.3.2 Minimiser l'incertitude sur l'évaluation des zones à dépolluer

L'objet des études est rarement d'estimer avec précision les teneurs dans les sols mais plutôt de juger si localement les terrains sont sains (risque sanitaire négligeable) ou contaminés (à traiter) pour un usage futur du site, c'est-à-dire de pouvoir établir si les teneurs en polluants sont supérieures ou inférieures à des seuils donnés. Il est alors attendu d'un échantillonnage complémentaire de **minimiser l'incertitude sur la délimitation des zones contaminées** (à dépolluer) et sur l'**évaluation des volumes** de matériaux correspondants (incertitude globale).

3.3.3 Trouver un compromis entre réduction d'incertitude et augmentation des coûts

Parce qu'elle nécessite souvent un plus grand nombre de données, la réduction de l'incertitude sur une classification des sols comme sains ou contaminés s'accompagne d'une augmentation des coûts de diagnostic. La rationalisation d'un échantillonnage complémentaire peut consister alors à définir un **« meilleur compromis » entre l'augmentation des coûts de diagnostic et la réduction de l'incertitude sur les volumes à dépolluer qui doit se traduire par une diminution des coûts de dépollution**. Tant que le prix à payer pour mieux reconnaître les zones à dépolluer reste inférieur aux gains réalisés sur la dépollution (sols à dépolluer mieux ciblés, tri sélectif évité, dimensionnement du dispositif de dépollution adapté, etc.), un échantillonnage complémentaire se justifie.

Selon la complexité des critères de rationalisation, le problème sera ou non formalisé (formulation mathématique). Le cas d'étude simple présenté au § 4 n'a pas nécessité de formulation mathématique explicite du problème d'échantillonnage. Dans certains situations plus complexes, il est cependant nécessaire d'exprimer une **fonction objectif** (ou de perte) qui quantifie l'impact financier ou autre de l'incertitude sur l'état de contamination des sols par rapport à des objectifs d'étude. Le nombre d'échantillons complémentaires et leurs positions doivent alors être déterminés de manière à minimiser la fonction objectif.

3.4 Estimation des incertitudes

Les incertitudes sont estimées par des méthodes différentes selon les objectifs de rationalisation de l'échantillonnage. Lorsqu'il s'agit de minimiser l'incertitude sur les teneurs en polluant en vue d'évaluer un stock de polluants (voir §3.3.1), les incertitudes peuvent être estimées par une méthode d'estimation (krigeage linéaire ou non-linéaire après transformation des données), telle que présentée dans le thème N°1. Lorsqu'il s'agit de minimiser l'incertitude sur la classification des sols comme sains ou contaminés (voir §3.3.2), les risques de mauvais classement sont souvent calculés à partir de simulations conditionnelles des teneurs en polluants dans les sols, comme décrit dans le thème N°2. La prise en compte d'aspects financiers dans la rationalisation de la reconnaissance (voir §3.3.3) s'appuie également sur un ensemble de simulations conditionnelles.

3.5 Calcul d'un plan d'échantillonnage rationalisé (localisation, nombre)

Les incertitudes quantifiées par estimation ou par simulations conditionnelles permettent d'évaluer les critères de rationalisation de l'échantillonnage (ou une fonction objectif) et d'en déduire une « meilleure solution » pour un échantillonnage complémentaire. Idéalement, nous souhaitons pouvoir déterminer le nombre minimum d'échantillons complémentaires et leurs emplacements pour réduire au mieux des incertitudes ou une fonction objectif. Dans bien des cas, cependant, le nombre d'échantillons est limité et l'exercice consiste à positionner au mieux les échantillons afin de minimiser

les incertitudes ou la fonction objectif. Le cas d'étude présenté au § 4 illustre parfaitement cette situation.

Qu'il s'agisse de minimiser des incertitudes ou une fonction objectif, l'optimisation pose certaines difficultés dans la mesure où les emplacements optimaux pour un schéma à n échantillons ne le sont généralement pas pour un schéma à n' > n échantillons (par ex., pour réduire l'incertitude locale sur les teneurs en polluants, nous placerons un échantillon unique là où l'incertitude est la plus grande, alors que l'emplacement optimal de 2 échantillons sera décalé par rapport à ce point de plus grande incertitude selon la configuration des données existantes et la corrélation spatiale des données). La résolution du problème nécessite donc une approche autre que séquentielle dans laquelle le nombre de données est augmenté progressivement à chaque étape, les emplacements optimaux des données à l'étape i étant conservés à l'étape i+1.

3.6 Production de résultats pour l'aide à la décision

Les résultats fournis sont au minimum des cartes précisant le nombre et la localisation des points d'échantillonnage complémentaire établis pour atteindre les objectifs de réduction des incertitudes.

D'autres résultats peuvent être produits dans certains cas. Nous pouvons par exemple évaluer l'impact de l'échantillonnage complémentaire sur l'étendue des zones incertaines dans une perspective de dépollution (Figure IV-1) ou la réduction de l'incertitude sur les coûts de dépollution (Figure IV-2) en fonction du nombre d'échantillons complémentaires.

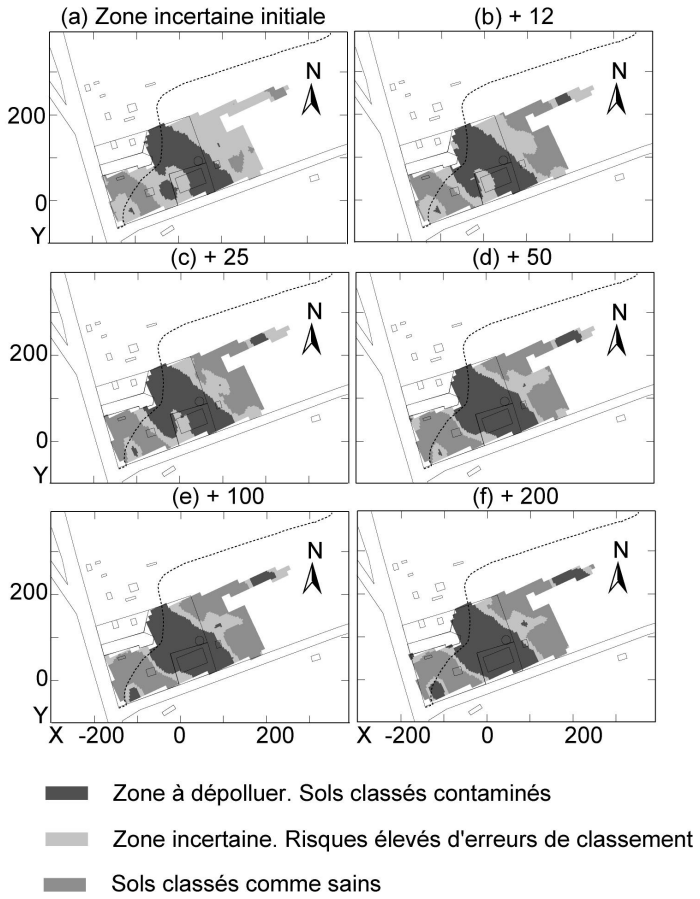


Figure IV-1 Estimation des zones très certainement à dépolluer, très certainement saines et incertaines en fonction du nombre de points d'échantillonnage complémentaire : (a) carte sans échantillons complémentaires, (b) à (f) cartes avec un nombre croissant d'échantillons complémentaires (12, 25, 50, 100 et 200 échantillons). La zone incertaine, représentée en gris très clair sur les cartes, est d'autant plus réduite que le nombre d'échantillons rajoutés est grand. Au-delà de 100 échantillons, la zone incertaine ne diminue cependant pratiquement plus.

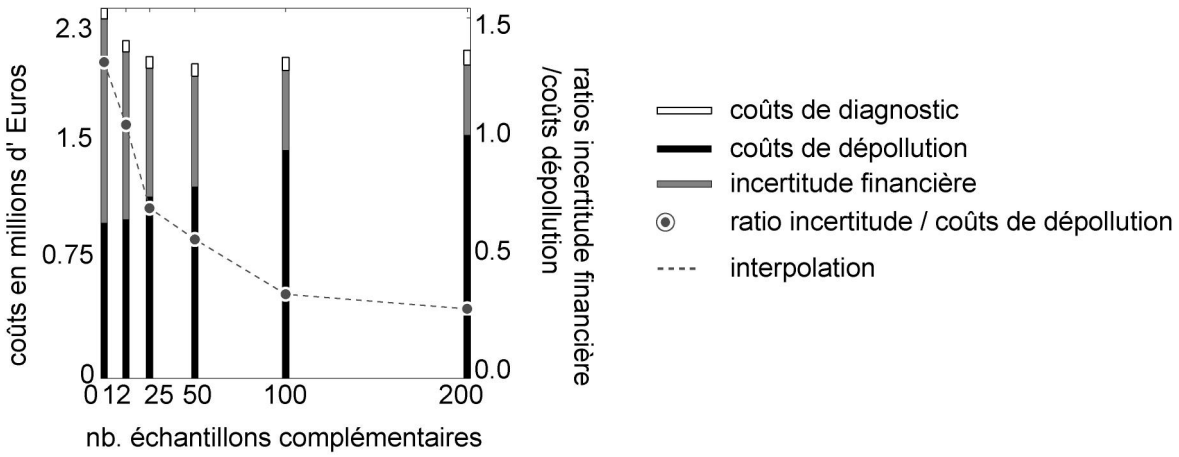


Figure IV-2 Incertitude sur l'estimation des coûts de dépollution, coûts de diagnostic et coûts globaux de réhabilitation (diagnostic + dépollution) en fonction du nombre d'échantillons qui pourraient être collectés en phase de reconnaissance complémentaire. Ces calculs prévisionnels montrent une diminution significative de l'incertitude sur l'estimation des coûts jusqu'à une centaine d'échantillons. Au-delà, le gain de précision devient négligeable, l'effet d'information, liés à un tri systématique des sols, devenant un facteur d'incertitude prépondérant. Il est à noter la très faible part des coûts de diagnostic dans les coûts globaux de réhabilitation.

4 ILLUSTRATION

Le site considéré présente une pollution multiple (6 polluants) sur un domaine d'environ 1 ha de superficie. Il s'agit d'une ancienne décharge sauvage où des produits pétroliers divers ont été déversés pendant de nombreuses années à partir des années 30. L'étude géostatistique a porté sur la modélisation de la répartition spatiale des teneurs en polluants dans les sols dans le cadre d'une EDR (définition du terme source pour un modèle de transport de polluants par les eaux souterraines). Une campagne d'échantillonnage complémentaire étant prévue dans l'EDR, une approche géostatistique a été suivie également pour établir un plan d'échantillonnage qui devait permettre de lever au mieux les incertitudes sur les limites de pollution reconnues à partir des données existantes. Le nombre maximal d'échantillons pour cette campagne avait été défini à l'avance indépendamment de l'étude géostatistique.

4.1 Contexte

Sur une période d'une dizaine d'années, les campagnes successives de reconnaissance menées sur l'ancienne décharge ont abouti à la réalisation de sondages de différents types (pelle mécanique, tarière mécanique) et à l'implantation de piézomètres. L'analyse en laboratoire des 58 échantillons de sols récoltés au cours de ces campagnes a mis en évidence des teneurs élevées en 6 substances, supérieures aux valeurs de définition de la source sol (VDSS) : hydrocarbures totaux, somme des 10 HAP (Som_10HAP), benzène, somme des xylènes (Som_Xylène), arsenic et cuivre.

Le niveau de la nappe est situé vers 4.5 m de profondeur au droit du site. L'aquifère des alluvions grossières constitue un aquifère très productif (perméabilité supérieure à 10^{-3} m/s). Le sous-sol est constitué de terre végétale, d'alluvions quaternaires composées de matériaux alternant entre argiles et sables limoneux et d'alluvions quaternaires grossières sablo-graveleuses jusqu'à une profondeur de 10 m. Vers 10 m, on note la présence d'une couche d'argile brune d'épaisseur proche de 1 m.

L'EDR avait pour objectif d'évaluer l'impact de la pollution des sols sur la nappe sous-jacente et d'établir le risque sanitaire que cette pollution représente pour les captages d'eau potable avoisinants. En fonction de ce risque, des mesures de réhabilitation adaptées au site devaient être proposées : traitement des sols si risque reconnu pour les captages, sinon fermeture du site et définition d'un périmètre de sûreté avec interdiction d'implanter de nouveaux captages d'eau à l'intérieur de ce périmètre.

Pour évaluer le risque d'atteinte des captages avoisinants, un modèle d'écoulement et de transport a été établi pour étudier à la fois l'épuisement du stock de polluants dans les sols au cours du temps et la migration possible de ces polluants jusqu'aux captages. L'approche géostatistique a permis de simuler la répartition spatiale des teneurs en polluants dans les sols à l'intérieur du site pour en déduire le terme source du modèle de transport. Dans une phase préliminaire, la géostatistique a été utilisée également pour établir le plan d'échantillonnage qui devait permettre de compléter les données disponibles provenant des campagnes de reconnaissance antérieures. L'illustration porte sur cette partie de l'étude. Elle comprend les étapes suivantes conformes à l'approche décrite au § 3.

- Analyse exploratoire des données existantes.
- Définition de critères de rationalisation de l'échantillonnage complémentaire.
- En rapport avec les critères de rationalisation retenus, production de cartes d'incertitude sur l'état de contamination des sols à partir des données existantes (avant reconnaissance complémentaire).
- Calcul d'un plan d'échantillonnage rationalisé et production de cartes avec l'emplacement des prélèvements complémentaires.

4.2 Analyse exploratoire des données existantes

Les données existantes sont les teneurs en 6 polluants (hydrocarbures totaux, Som_10HAP, benzène, Som_Xylène, arsenic et cuivre) mesurées au laboratoire ainsi que des informations sur l'état d'imprégnation des sols en hydrocarbures.

Ces données proviennent de 58 échantillons de sols prélevés dans deux types de sondages.

- 17 sondages à la pelle mécanique, jusqu'à une profondeur de 4.5 m, avec prise d'échantillons de sols « ponctuels » tous les 1.5 m (à 1.5, 3 et 4,5 m).
- 10 sondages à la tarière mécanique, jusqu'à une profondeur de 6 m, avec prise d'échantillons de sols « ponctuels » tous les 2 m (à 2, 4 et 6 m).

Un même protocole opératoire ayant été suivi pour constituer les échantillons de taille réduite dans ces deux types de sondages, le support peut être considéré comme identique pour toutes les données de teneur. Ce point est confirmé par l'analyse statistique qui ne permet pas de différencier des populations de teneurs par type d'ouvrage. Le plan du site et l'emplacement de ces sondages existants se trouvent à la Figure IV-3.

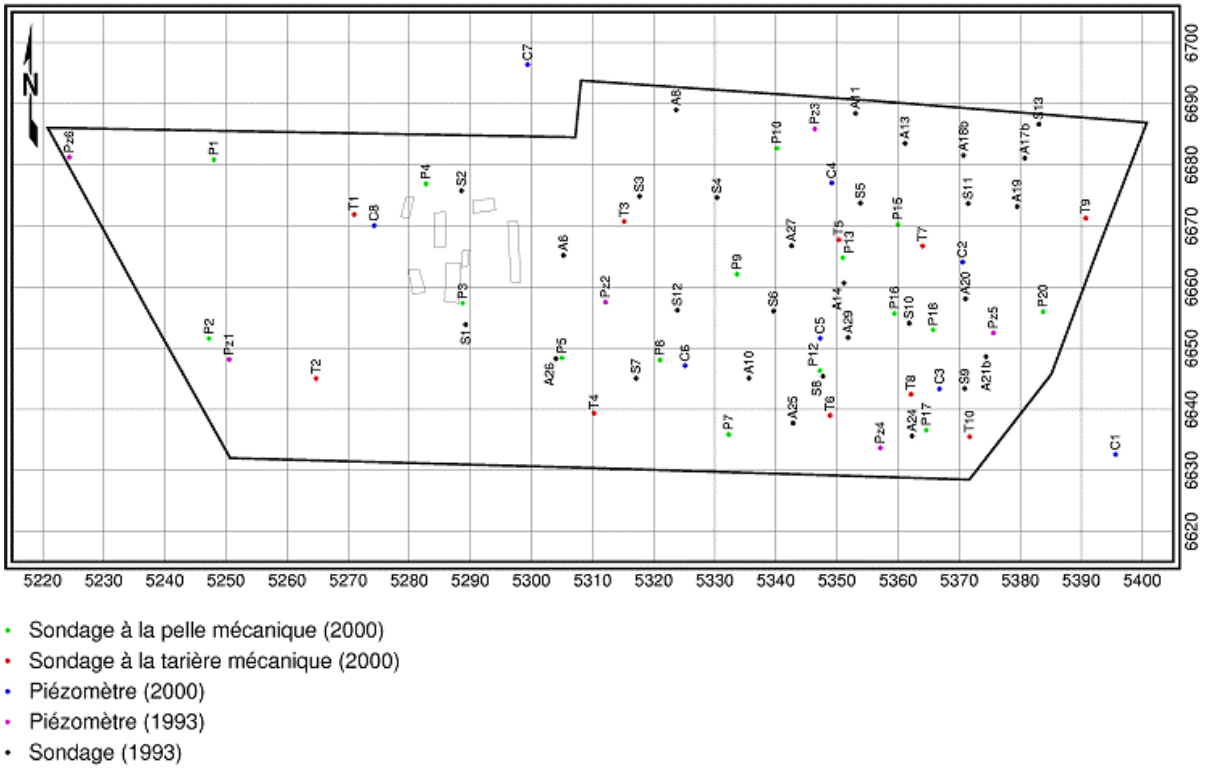


Figure IV-3 Localisation sur le plan du site des sondages existants.

L'information sur le niveau d'imprégnation des sols par des hydrocarbures est fournie par les coupes géologiques. Une distinction peut ainsi être faite entre les sols non imprégnés (aucune marque visuelle de pollution) et ceux peu à très imprégnés (présence de traces d'hydrocarbures voire d'une phase huile importante mobile ou non). Elle aboutit au classement des sols en deux catégories. Par manque de fiabilité de ces observations, une distinction plus fine entre sols peu et très imprégnés a été abandonnée.

Des statistiques conditionnelles[22] ont été calculées sur les teneurs en ces 6 polluants en distinguant les échantillons selon leur état d'imprégnation (Tableau IV-2 et Tableau IV-3). La comparaison des deux catégories de sols peut se faire à partir des moyennes, des valeurs maximales ou de la proportion de valeurs fortes (quantiles) calculées pour chaque catégorie. Ces statistiques confirment une contamination plus importante des sols imprégnés avec des teneurs dans les sols plus fortes à beaucoup plus fortes selon le polluant. Seule exception possible à cette tendance, les xylènes dont les teneurs les plus fortes ont été mesurées sur des échantillons de sols qui seraient non imprégnés. Il est

[22] Les statistiques conditionnelles sont des statistiques élémentaires calculées par classe. Ici, nous distinguons la classe des échantillons non imprégnés par des hydrocarbures de celle des échantillons peu à très imprégnés par des hydrocarbures.

à noter aussi que pour tous les polluants, des teneurs très faibles ont été mesurées sur des sols imprégnés. Ces deux résultats montrent la limite des observations visuelles mais suscitent également un doute sur la représentativité réelle des échantillons de sols analysés.

Tableau IV-2 Résumé statistique sur les teneurs en polluants pour des échantillons non imprégnés par des hydrocarbures (unité mg/kg ms) (voir §I2.3.3Statistiques élémentaires, p. 26 pour les définitions)

	HCT	Som_10HAP	Benzène	Som_Xylène	Arsenic	Cuivre
Nb	34	33	34	34	29	29
Moyenne	7.8E+03	1.8E+00	7.0E-02	1.1E+01	7.2E+00	2.3E+01
Ecart-type	6.0E+03	6.0E+00	3.5E-01	5.6E+01	2.9E+00	7.7E+01
Coefficient Variation	5.6E+03	3.3E+00	5.2E+00	5.3E+00	4.1E-01	3.3E+00
Minimum	0.0E+00	0.0E+00	0.0E+00	0.0E+00	2.9E+00	2.4E+00
Maximum	2.6E+05	3.1E+01	2.0E+00	3.3E+02	1.4E+01	4.2E+02
Intervalle	2.6E+05	3.1E+01	2.0E+00	3.3E+02	1.1E+01	4.2E+02
5ème Quantile	0.0E+00	0.0E+00	0.0E+00	0.0E+00	3.4E+00	2.5E+00
10ème Quantile	0.0E+00	0.0E+00	0.0E+00	0.0E+00	4.0E+00	2.8E+00
25ème Quantile	2.5E+01	0.0E+00	0.0E+00	0.0E+00	5.0E+00	3.9E+00
Médiane	8.6E+01	1.0E-02	0.0E+00	0.0E+00	6.7E+00	5.3E+00
75ème Quantile	4.0E+02	2.3E-01	0.0E+00	0.0E+00	8.9E+00	8.4E+00
90ème Quantile	1.0E+03	7.0E+00	0.0E+00	1.6E+00	1.3E+01	3.7E+01
95ème Quantile	6.5E+04	1.9E+01	6.9E-01	1.1E+02	1.4E+01	2.4E+02

Tableau IV-3 Résumé statistique sur les teneurs en polluants pour des échantillons peu à très imprégnés par des hydrocarbures (unité mg/kg ms)

	HCT	Som_10HAP	Benzène	Som_Xylène	Arsenic	Cuivre
Nb	34	33	34	34	29	29
Moyenne	7.8E+03	1.8E+00	7.0E-02	1.1E+01	7.2E+00	2.3E+01
Ecart-type	6.0E+03	6.0E+00	3.5E-01	5.6E+01	2.9E+00	7.7E+01
Coefficient Variation	5.6E+03	3.3E+00	5.2E+00	5.3E+00	4.1E-01	3.3E+00
Minimum	0.0E+00	0.0E+00	0.0E+00	0.0E+00	2.9E+00	2.4E+00
Maximum	2.6E+05	3.1E+01	2.0E+00	3.3E+02	1.4E+01	4.2E+02
Intervalle	2.6E+05	3.1E+01	2.0E+00	3.3E+02	1.1E+01	4.2E+02
5ème Quantile	0.0E+00	0.0E+00	0.0E+00	0.0E+00	3.4E+00	2.5E+00
10ème Quantile	0.0E+00	0.0E+00	0.0E+00	0.0E+00	4.0E+00	2.8E+00
25ème Quantile	2.5E+01	0.0E+00	0.0E+00	0.0E+00	5.0E+00	3.9E+00
Médiane	8.6E+01	1.0E-02	0.0E+00	0.0E+00	6.7E+00	5.3E+00
75ème Quantile	4.0E+02	2.3E-01	0.0E+00	0.0E+00	8.9E+00	8.4E+00
90ème Quantile	1.0E+03	7.0E+00	0.0E+00	1.6E+00	1.3E+01	3.7E+01
95ème Quantile	6.5E+04	1.9E+01	6.9E-01	1.1E+02	1.4E+01	2.4E+02

4.3 Définition de critères de rationalisation de l'échantillonnage

Pour juger de l'état de connaissance de la contamination des sols, les teneurs en polluants ont été évaluées par rapport à des valeurs guides[23]. Plutôt que de savoir si nous pouvons estimer avec précision les teneurs en polluants en un point donné, l'utilisation de valeurs guides nous permet de savoir si nous pouvons déterminer avec confiance si les sols sont localement fortement pollués ou non. Minimiser ce risque de mauvaise classification des sols a été choisi comme critère de rationalisation. Ainsi, l'état de contamination des sols est jugé incertain si le risque de mauvaise classification des sols est grand pour au moins un des six polluants.

4.4 Production de cartes de la contamination et estimation de l'incertitude

Le critère de rationalisation de l'échantillonnage choisi suppose de pouvoir estimer, à partir des données disponibles sur chaque polluant, une carte 3D de la probabilité de dépassement de sa valeur guide. Plusieurs méthodes géostatistiques permettent d'estimer une probabilité de dépassement. Citons l'estimation par krigeage, généralement après transformation des données (par ex. transformation log, gaussienne ou par indicatrice), ou la simulation conditionnelle, qui sont les méthodes les plus courantes déjà présentées dans les thèmes 1 et 2.

Pour chacun des 6 polluants, il a été déduit de la carte 3D de probabilité de dépassement une carte 2D synthétique des probabilités les plus fortes calculées sur la verticale (de 0 à 4.5 m de profondeur). A titre d'exemple, la Figure IV-4 montre la carte ainsi obtenue pour les hydrocarbures totaux.

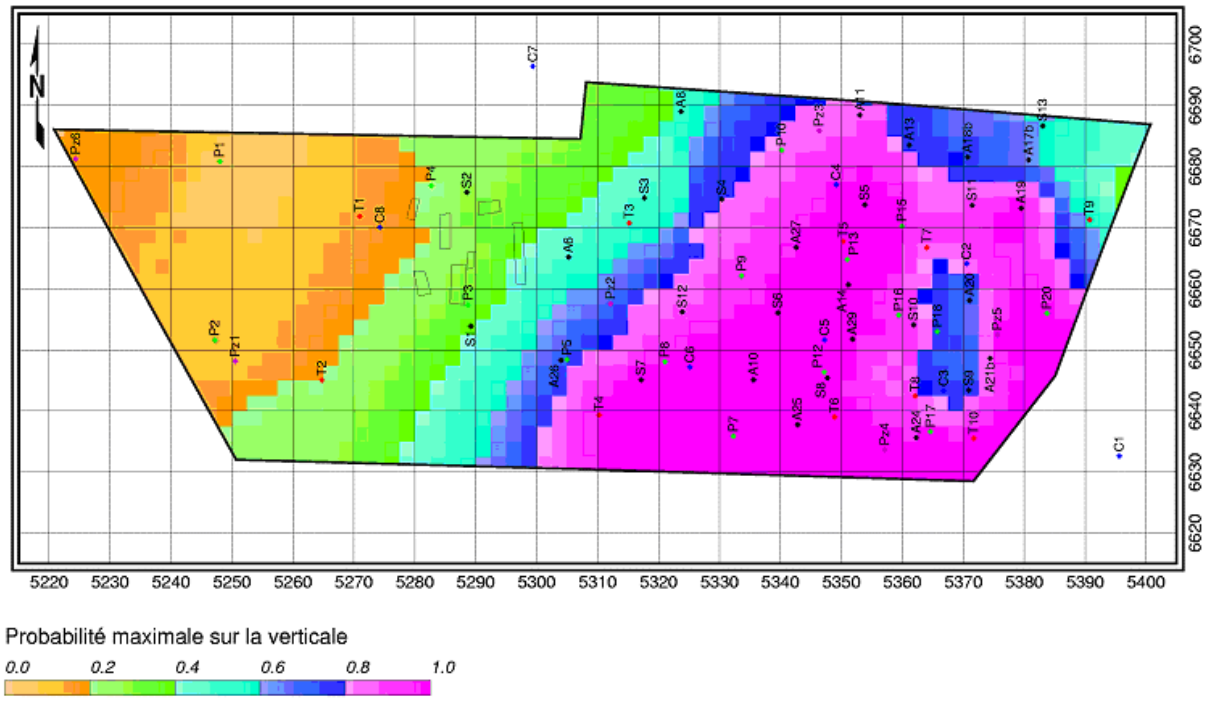


Figure IV-4 Carte de probabilité de dépassement de la VDSS (2500 mg/kg ms) par les hydrocarbures totaux, établie à partir des données existantes. Les probabilités représentées sont les probabilités maximales calculées sur la verticale (0 à 4.5 m de profondeur)

Ces cartes décrivent l'incertitude sur l'estimation de l'état de contamination des sols. Le risque de contamination (teneurs supérieures aux VDSS) est grand dans les zones de plus fortes probabilités. Il

23 Les valeurs guides choisies sont les valeurs de définition de source sol (VDSS).

est faible dans les zones de plus faibles probabilités. Les probabilités intermédiaires traduisent l'existence, verticalement sur toute la hauteur de terrains, d'une incertitude non négligeable sur la classification des sols comme fortement contaminés (teneurs supérieures aux VDSS) ou non.

Il est à noter les points suivants.

1. Il aurait été utile ici de tenir compte de l'état d'imprégnation des sols qui peut être associé à des niveaux de contamination très contrastés des sols, au moins pour certains polluants. Les observations visuelles de ce type étant très nombreuses, elles permettraient d'améliorer l'estimation des probabilités de dépassement des valeurs guides. Une tentative de cartographie de l'état d'imprégnation des sols a montré cependant de nombreuses discontinuités dues à la fois à l'imprécision (subjectivité) des observations visuelles et à un mauvais positionnement vertical de ces observations. L'idée d'exploiter ces données a dû alors être abandonnée.
2. L'existence de corrélations entre polluants n'a pas été prise en compte dans cette première partie de l'étude géostatistique où l'objectif était de déterminer assez sommairement les zones de plus grande incertitude. Ces corrélations ont en revanche été prises en compte dans la seconde partie (voir § 4.6) où l'objectif était d'évaluer plus finement le stock de polluant et sa composition.

4.5 Calcul d'un plan d'échantillonnage rationalisé (nombre, localisation)

Une rationalisation du nombre et de la localisation des points d'échantillonnage complémentaire a été obtenue de la façon suivante.

Le nombre de nouveaux sondages était soumis à une limite budgétaire. Le budget disponible permettait la réalisation de 15 à 20 sondages à la pelle mécanique ainsi que l'analyse en laboratoire des échantillons prélevés. Compte tenu de l'incertitude estimée sur l'état de contamination des sols à partir des données existantes, il est apparu nécessaire de réaliser la totalité des sondages prévu dans l'enveloppe budgétaire.

Les points d'échantillonnage ont été positionnés pour réduire au mieux l'incertitude sur les limites de la source de contamination. Les cartes de probabilité établies pour les 6 polluants d'intérêt montrent toutes des zones de probabilités intermédiaires (zones incertaines) de plus ou moins grande extension comme sur la Figure IV-4. En accord avec les responsables du site, il a été décidé de définir comme incertaines, pour chaque polluant, les zones où la probabilité de dépassement de la valeur guide est comprise entre 0.4 et 0.6, c'est-à-dire ni suffisamment faible pour penser que les sols sont plutôt faiblement contaminés sur toute leur hauteur, ni suffisamment forte pour penser que des sols fortement contaminés existent à certaines profondeurs. L'union des zones incertaines ainsi calculées pour chaque polluant détermine la zone incertaine globale. Il a finalement été choisi de placer les sondages complémentaires selon une répartition régulière qui recouvre toute cette zone incertaine (Figure IV-5).

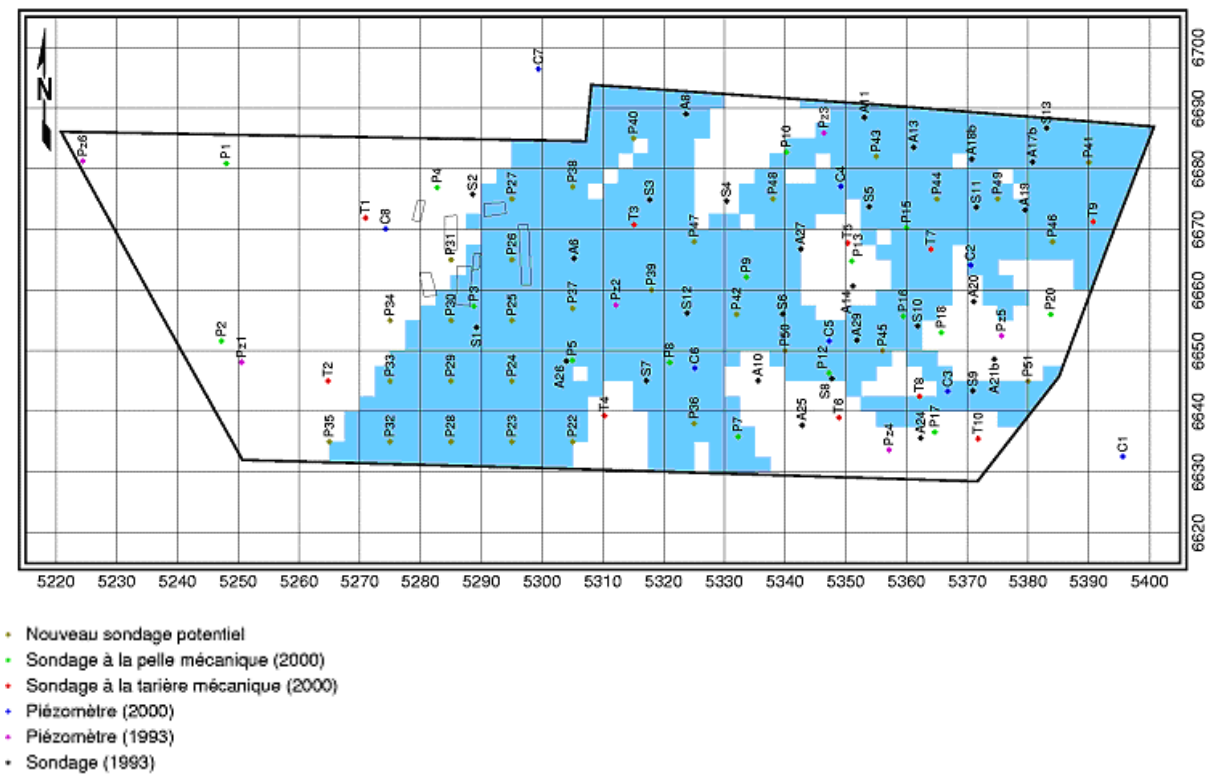


Figure IV-5 : Localisation de la zone incertaine globale (en bleu) et des sondages complémentaires possibles.

Avec un nombre de sondages complémentaires à la limite du budget disponible, une stratégie de creusement des sondages a été proposée pour pouvoir suivre visuellement l'état d'imprégnation des sols, à l'avancement de la reconnaissance, et décider sur le terrain de la nécessité ou non de forer tous les sondages. Trois raisons justifiaient une telle approche.

- L'extension de la pollution vers l'ouest concerne exclusivement des hydrocarbures.
- Les teneurs en polluant tendent à diminuer vers l'ouest.
- L'absence d'imprégnation dans les zones peut s'apparenter à des niveaux de contamination faibles en hydrocarbures.

La stratégie d'échantillonnage a donc été de réaliser un front d'avancement des sondages, de l'est vers l'ouest, les sondages étant positionnés aux emplacements prévus sur la Figure IV-5. Comme attendu, les premiers sondages dans la zone de plus fortes probabilités ont montré des sols très imprégnés. Au fur et à mesure de l'avancement vers l'ouest, les imprégnations se sont révélées moins importantes. La campagne de sondages a été stoppée lorsque plus aucune imprégnation n'a été observée sur les sols. Un nombre total de 17 sondages a ainsi été foré. L'emplacement de ces sondages, où des échantillons de sols ont été prélevés pour analyse en laboratoire, est représenté sur la Figure IV-6.

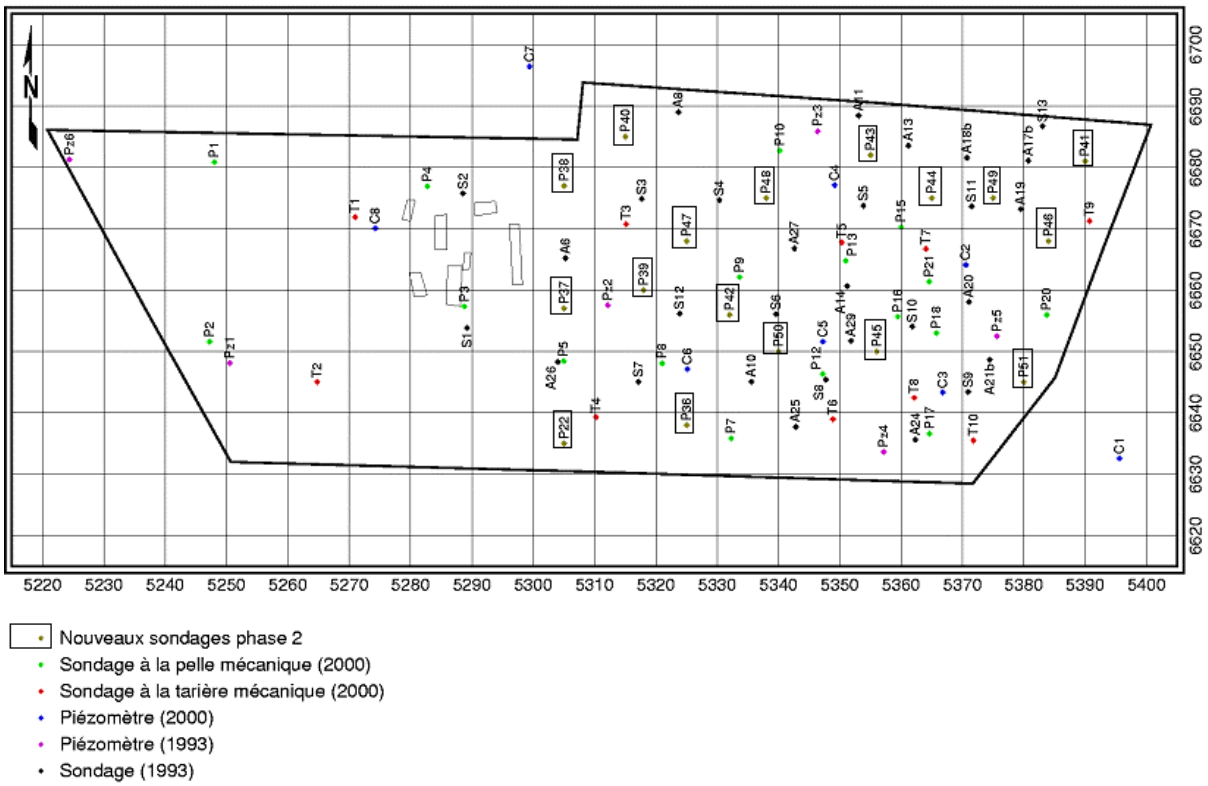


Figure IV-6 : Localisation des sondages réellement forés après suivi de l'état d'imprégnation des sols au cours de la reconnaissance complémentaire.

4.6 Valorisation des résultats d'échantillonnage complémentaire

Les 17 nouveaux sondages creusés à la pelle mécanique jusqu'à une profondeur de 4.5 m ont donné lieu au prélèvement d'échantillons de sols tous les 1.5 m (à 1.5, 3 et 4.5 m). Parmi les échantillons de sols prélevés, 32 ont fait l'objet d'analyses en laboratoire. Ces nouvelles données ont été utilisées en complément des données acquises lors des campagnes précédentes pour mettre à jour les cartes de probabilité et pour produire des cartes de teneurs qui ont été introduites dans les modèles d'écoulement et de transport.

A la différence de la première étape (§ 4.4), les nouvelles cartes de probabilités prennent en compte les corrélations entre polluants. La Figure IV-7 représente la carte des probabilités maximales sur la verticale calculées pour les hydrocarbures totaux après mise à jour du modèle. Par comparaison à la carte des probabilités estimée sans les données complémentaires (Figure IV-4), nous constatons que la zone incertaine (de couleur bleu ciel) est bien plus réduite sur la nouvelle carte de probabilités qui intègre les données complémentaires. Cette carte paraît également plus contrastée avec des probabilités toutes supérieures à 0.8 dans la partie est et toutes inférieures à 0.2 dans la partie ouest.

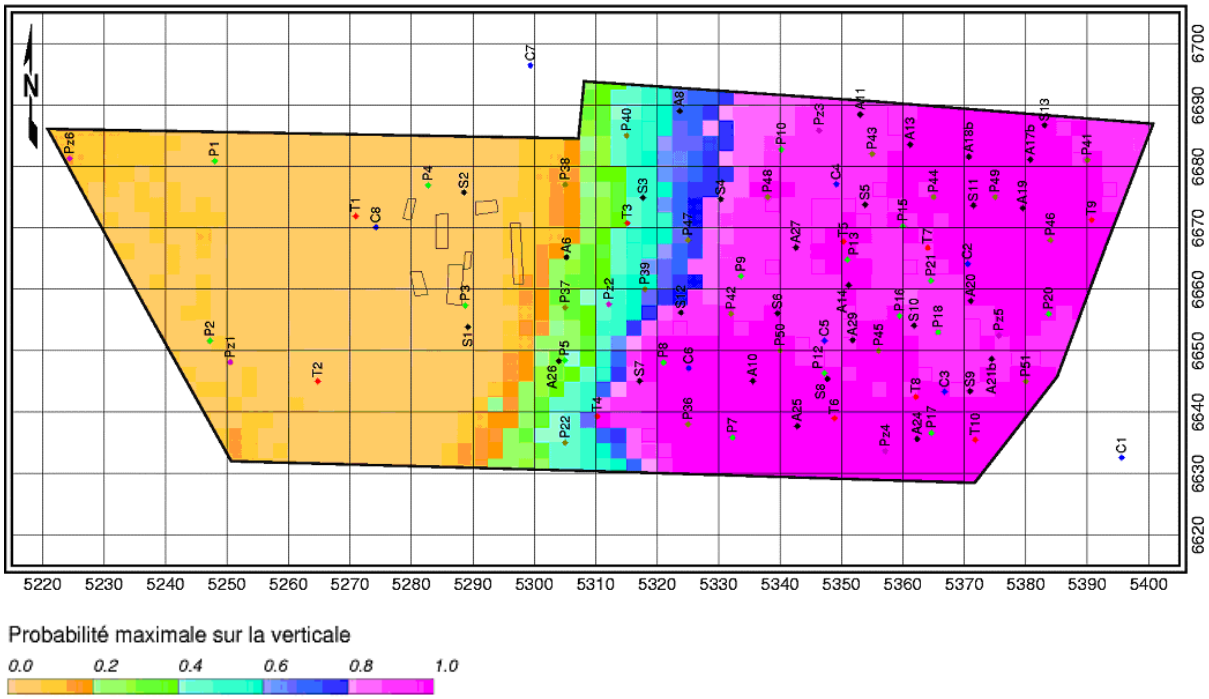


Figure IV-7 Carte de probabilité de dépassement de la VDSS (2500 mg/kg ms) par les hydrocarbures totaux, établie à partir des données anciennes et complémentaires. Les probabilités représentées sont les probabilités maximales calculées sur la verticale.

CONCLUSION

L'étude des sites pollués soulève de nombreux problèmes qui relèvent de la géostatistique :

- analyse de données spatiales,
- cartographie, délimitation de zones, calcul de volumes,
- quantification de l'incertitude sur des estimations, prise en compte pour l'aide à la décision,
- intégration d'informations multiples de nature, nombre et précision diverses,
- stratégie d'échantillonnage,
- etc.

Ces problèmes se posent au quotidien aux professionnels des sites et sols pollués, tant du côté des bureaux d'étude qui doivent évaluer au mieux l'état de contamination des sites et proposer à leurs clients des solutions de réhabilitation avantageuses et justes, que des instances administratives qui engagent leur responsabilité en acceptant ou non les solutions qui leur sont proposées.

Dans ce contexte, ***GeoSiPol*** avait pour première mission d'établir un manuel méthodologique avec un triple objectif :

1. replacer la géostatistique dans un cadre métier sur différents thèmes jugés importants pour l'étude des sites et sols pollués,
2. rendre compréhensible par tous le principe d'une approche géostatistique adaptée à chaque thème et en justifier l'intérêt et l'apport,
3. fournir aux bureaux d'études et aux institutionnels le matériel nécessaire pour qu'à leur tour ils puissent proposer et défendre la pratique d'approches géostatistiques auprès des industriels ou autres détenteurs de sites et des autorités publiques.

L'exercice a conduit à la rédaction du manuel méthodologique, axé sur quatre thèmes jugés prioritaires par les membres adhérents de ***GeoSiPol***. D'autres thèmes sont ressortis comme importants et pourraient faire l'objet de présentations méthodologique similaires dans le futur.

La rédaction collective de ce manuel, par des personnes à la fois « expertes » en géostatistique mais aussi très impliquées dans le domaine de l'environnement, leur revue critique par un grand nombre de professionnels des sites pollués et leur validation par tous ont permis d'atteindre ces objectifs.

Sur tous les thèmes abordés, la question n'est pas ou plus de savoir si la géostatistique peut être utile et sous quelles conditions. Elle y trouve en effet sa place en apportant à la fois des méthodes de travail et des outils précieux. Si des questions demeurent au moment d'une étude, elles portent sur la définition des objectifs et sur la quantité d'information nécessaire pour les atteindre et les atteindre sûrement. La géostatistique permet d'évaluer cette quantité d'information minimale et de la traduire en une incertitude sur la réalisation des objectifs. Il n'est pas juste de croire que d'autres approches, généralement plus simples, permettraient de faire cette économie de données.

L'étape suivante apparaît au moins aussi importante que ce manuel. Il s'agit de mener les actions qui permettront la mise en pratique de la géostatistique et sa reconnaissance large. Si ces actions incombent aux membres adhérents du groupe, ***GeoSiPol*** se prépare aussi à y contribuer très activement par :

- l'organisation de séminaires pour informer et former aux approches géostatistiques présentées dans le manuel méthodologique,
- l'organisation de « travaux pratiques » sur des cas d'étude réels, pour amener les détenteurs de sites à s'impliquer dans le groupe, pour approfondir certaines approches géostatistiques ou bien encore pour comparer différentes méthodes ou approches,
- l'organisation régulière de réunions ***GeoSiPol*** de discussion et d'échanges où chacun pourra exposer aux autres des cas d'étude, ses retours d'expérience et les difficultés rencontrées,
- la promotion du groupe et de ses travaux à travers son site internet www.geosipol.org, la présence de ***GeoSiPol*** dans les salons ou les conférences, et la diffusion contrôlée des travaux du groupe,
- l'ouverture du groupe à un plus grand nombre de professionnels.

BIBLIOGRAPHIE

LIVRES DE REFERENCE EN GEOSTATISTIQUE

Chilès J.P. and P. Delfiner, 1999, Geostatistics: modeling spatial uncertainty. Wiley.

Goovaerts P.,1997, Geostatistics for natural resources evaluation. Oxford University Press.

Isaaks E.H. and R.M. Srivastava, 1989, An introduction to applied geostatistics. Oxford University Press.

Lantuéjoul C., 2002, Geostatistical simulation: models and algorithms. Springer.

Myers J.C., 1997, Geostatistical error management: quantifying uncertainty for environmental sampling and mapping. Wiley.

Olea R.A., 1991, Geostatistical glossary and multilingual dictionary, Oxford University Press.

Rivoirard J., 1994, Introduction to disjunctive kriging and non-linear geostatistics. Oxford University Press.

Saporta G.,1990, Probabilités, analyse des données et statistique. Technip.

PUBLICATIONS CITEES DANS LE MANUEL

Demougeot-Renard H., 2002, De la reconnaissance à la réhabilitation des sols industriels pollués: estimations géostatistiques pour une optimisation multicritère. Thèse de doctorat n° 14615, Ecole Polytechnique Fédérale, Zürich, Suisse. (Thèse disponible : http://www.vdf.ethz.ch/)

Demougeot-Renard H., C. de Fouquet, P. Renard, 2004, Forecasting the number of soil samples required to reduce remediation cost uncertainty. Journal of Environmental Quality, vol. 33, pp. 1694-1702.

Demougeot-Renard H., C. de Fouquet, 2004, Geostatistical approach for assessing soil volumes requiring remediation: validation using lead-polluted soils underlying a former smelting works. Environmental Science and Technology, vol 38, n°19, pp. 5120-5126.

Deverly F., 1984, Echantillonnage et géostatistique. Thèse de Doct.-Ing. en Sciences et Techniques Minières, Option Géostatistique, Ecole des Mines de Paris. [http://cg.ensmp.fr/bibliotheque]

Jeannée N., 2001, Caractérisation géostatistique de pollutions industrielles de sols. Cas des HAP sur d'anciens sites de cokeries. Thèse de Doctorat en Géostatistique, Ecole des Mines de Paris. [http://cg.ensmp.fr/bibliotheque]

Jeannée N., C. de Fouquet, 2003, Apport d'informations qualitatives pour l'estimation des teneurs en milieux hétérogènes : cas d'une pollution de sols par des hydrocarbures aromatiques polycycliques. C.R. Geoscience 335, pp 441-449.

SITES INTERNET

- Groupe de travail français ***GeoSiPol*** : http://www.geosipol.org/
- Publications du Centre de Géostatistique de l'Ecole des Mines de Paris : http://cg.ensmp.fr/
- Publications de FSS International : http://www.fssintl.com/
- Publications de Géovariances : http://www.geovariances.com/
- Forum européen de géostatistique et statistique spatiale : http://www.ai-geostats.org/
- Association **geoENVia** organisatrice des conférences européennes en géostatistique et environnement geoENV : http://www.geoenvia.org/

INDEX DES TERMES GÉOSTATISTIQUES

Cet index reprend les principaux termes statistiques et géostatistiques utilisés dans le manuel et leurs synonymes. Il renvoie, pour chacun d'entre eux, aux pages où le terme est explicité. La traduction anglaise des termes est indiquée entre parenthèses.

Analyse en composantes principales (principal component analysis), *96*

Analyse exploratoire des données (exploratory data analysis, *20*

Changement de support (change of support), *53*, *64*

Coefficient de corrélation (correlation coefficient), *91*

Cokrigeage (cokriging), *94*

Echantillonnage hétérotopique (heterotopic sampling), *87*

Effet d'information (information effect), *57*

Effet de pépite (nugget effect), *31*

Information indirecte, information secondaire, information auxiliaire (soft data), *88*

Isotropie (isotropy), *29*

Krigeage (kriging), *33*

Modèle de variogramme (variogram model), *32*

Nuage de corrélation, diagramme de dispersion (scatterdiagram, scatterplot, crossplot), *90*

Nuée variographique (variogram cloud), *27*

Simulations (simulations), réalisations (realizations), *61*

Stationnarité (stationarity), *29*

Support (support), *21*

Validation croisée (cross validation), *33*

Variabilité spatiale (spatial variability), distribution spatiale (spatial distribution), structure spatiale (spatial structure), répartition spatiale, corrélation spatiale (spatial correlation), *18*, *53*

Variogramme croisé (cross variogram), *94*

Variogramme expérimental (experimental variogram), *28*