Caractérisation géostatistique de pollutions industrielles de sols: cas des hydrocarbures aromatiques polycycliques sur d'anciens sites de cokeries

Nicolas Jeannée

ECOLE DES MINES
DE PARIS

THESE

pour obtenir le grade de

Docteur de l'Ecole des Mines de Paris

Spécialité « **Géostatistique** »

présentée et soutenue publiquement

par

Nicolas Jeannée

le 30 mai 2001

Caractérisation géostatistique de pollutions industrielles de sols
Cas des hydrocarbures aromatiques polycycliques sur d'anciens sites de cokeries

Directeur de thèse : Chantal de Fouquet

Jury

MM.	Ghislain	de Marsily	Président
	Jean-Paul	Chilès	Rapporteur externe
	Michel	Jauzein	Rapporteur externe
Mme	Chantal	de Fouquet	Examinateur
MM.	Frédéric	Goldschmidt	Examinateur
	Jacques	Laversanne	Examinateur
	Jacques	Rivoirard	Examinateur

"Si la géostatistique a simplement réussi à donner l'habitude aux gens de distinguer les problèmes différents ou les effets physiques différents, de ne plus confondre un effet de support avec une erreur d'échantillonnage ou des choses comme ça, [...] eh bien je crois que ce serait déjà pas mal, si on était arrivé à ça."

Georges Matheron

(Conférence organisée par Géovariances, Ecole des Mines de Paris, 1987)

En essayant continuellement on finit par réussir.
Donc : plus ça rate, plus on a de chances que ça marche.

Devise Shadok

Remerciements

Je tiens tout d'abord à remercier sincèrement Chantal de Fouquet pour l'encadrement qu'elle a apporté à ce travail, son enthousiasme et sa rigueur scientifique.

Je désire également exprimer toute ma reconnaissance à Jacques Rivoirard. Plusieurs des outils utilisés lui doivent d'exister et ses remarques ont eu un poids décisif au cours de l'évolution de ce travail.

Ghislain de Marsily m'a offert l'opportunité de découvrir les sciences de l'eau et de la terre en m'accueillant au sein de son DEA il y a cinq ans. Je le remercie aujourd'hui pour l'honneur qu'il m'a fait en acceptant de présider le jury de ma thèse.

Mes plus sincères remerciements vont aux deux personnes qui ont accepté d'être rapporteurs de ce travail, Jean-Paul Chilès et Michel Jauzein, pour leurs nombreuses remarques techniques et un recul que je leur envie.

Ayant été à l'origine de la collaboration entre le Centre de Géostatistique et le Centre National de Recherche sur les Sites et Sols Pollués, Emmanuel Ledoux est en partie responsable de tout cela... Je l'en remercie chaleureusement, ainsi que pour le regard critique qu'il a régulièrement apporté à l'occasion de réunions d'avancement.

Paul Lecomte fut l'un des initiateurs de ce travail en tant que directeur du Centre National de Recherche sur les Sites et Sols Pollués. Son expertise dans le domaine des sols pollués, ses craintes parfois et conseils souvent ont contribué à ce que ce travail, en apportant un savoir concret aux praticiens, atteigne l'un de ses objectifs essentiels.
Un grand merci à Frédéric Goldschmidt, son successeur, et à travers lui aux membres du CNRSSP, passés et actuels, qui m'ont fourni les données à la base de ce travail et m'ont permis tout au long de ces quatre années d'apprendre énormément.

Une des plus belles récompenses pour ce travail m'a été donnée par Jacques Laversanne, de Charbonnages en France, lorsqu'il m'a remercié pour l'avancée et les outils que ce travail apportait à la problématique des sites et sols pollués. Je le remercie pour ses interventions, qui ont permis de rapprocher ce travail du terrain.

Si j'éprouve une fierté particulière à avoir été assistant du projet 1 (d'accord nous n'étions souvent que deux...), c'est en grande partie grâce à Philippe Wavrer ; il a commencé par m'apprendre à mettre des gants et faire des trous pour finir par me relire (et me fournir quelques photos, aussi). Parce que c'est aussi cela, encadrer, je le remercie tout particulièrement.
Philippe est parti, d'autres ont suivi et le projet 1 a continué grâce à Laurent Eisenlohr puis Ignace

Salpeteur que je remercie pour leurs remarques et apports à ce travail.

Parce que grâce à son regard acéré et ses commentaires avisés, de visage maintes fois ce mémoire a changé, je tiens à remercier sincèrement Nicolas Bez (et aussi pour avoir éliminé de telles envolées...).
D'autres relectures plus ponctuelles furent également constructives, et je remercie pour cela Michel Schmitt et Hans Wackernagel.

Merci également à Christian Lantuéjoul, dont j'ai pu à maintes reprises bénéficier des connaissances, de la pédagogie et des programmes.

Un immense merci à Françoise Poirier (même si "there is not a what"...) pour ses traductions, ses encouragements et dans un registre slightly différent pour l'inestimable découverte du cinéma classique que je lui dois.

Merci à Nathalie Dietrich et Isabelle Schmitt, aux autres permanents du centre et aux petits (mais brillants !) thésards qui me suivent pour les relations que nous avons eues. Je ne saurais oublier Laurent Bertino & David Geraets, qui m'ont supporté pendant de longues heures de bureau (ouais enfin moi aussi j'ai dû vous supporter hein...) et avant eux Chris Roth pour ses encouragements et ses "putain mais c'est quoi ces variogrammes de merde ? ? ?", Bertrand Iooss et Estelle Schuhler.

Il y a cinq ans, lorsque pour la première fois l'occasion m'a été donnée de rédiger les remerciements d'un travail, j'avais conclu par un elliptique "Merci finalement à tous ceux que je n'ai pas nommé mais qui savent tout ce que je leur dois". Ils sont la face immergée de l'iceberg (bien qu'étant nettement moins froids...), et cette fois-ci je vais être un peu plus long. Cachés derrière chaque page, en embuscade sous chaque figure, sans eux peut-être ce travail aurait-il quand même vu le jour, mais je n'aurais sûrement pas eu le même plaisir à le réaliser...

Alors ça y est le bal est ouvert : Laurent (tu avais raison, cela valait la peine... Elle est pas belle ? ? ! ?), Malek (koulchi labès ! ! On surfera encore longtemps côte à côte...), Delphine (putain bordel de merde on en a quand même passé, des bons moments !), Thomas ("ah mais pardon, un moment ! !", pour tes solos, ta vision du monde et le reste), Diz (pour toutes tes disableries et surtout pour le reste), Guillaume (témoin des origines à nos jours de ce que je suis), Cédric (valeureux "premier de cordée", bienheureux au pays du soleil et que j'espère rejoindre un jour), Eric (mon ami et canard favori), ma petite Cécile, Anne, Jérôme, Virginie, Cathy, Marie-Pierre, Valéry "blondin", Stéphanie ma b.p. et Gaël, Thierry (qui est maintenant bronzé vraiment de partout) et Stéphanie, Valéry "blondin", Bruno, Arnaud et il y en a d'autres bien entendu, beaucoup d'autres, et je les remercie aussi...

Bon ça y est c'est presque fini, restent les inclassables, celles qui occupent une place un peu à part. Je tiens à remercier Stéphanie (car si je suis arrivé au début de cette thèse, c'est encore un peu grâce à toi) et finalement Tatiana (car si je suis arrivé à la fin de cette thèse, c'est aussi pour et grâce à toi).

Pour finir, je tiens à adresser le plus grand merci à ma famille, qui me prouve depuis des années que les proverbes ont parfois tort et que, bien que loin des yeux...

Table des matières

Première partie

Introduction

Introduction

Le développement industriel a depuis 150 ans entraîné de multiples pollutions, d'origines accidentelles autant que diffuses ; sols, nappes souterraines et atmosphère sont concernés. La prise de conscience des risques sanitaires posés par de telles pollutions est apparue tardivement, il y a une vingtaine d'années, suite à des accidents graves fortement médiatisés[1]. Ce lourd passé industriel doit aujourd'hui être géré. En France, on comptabilise entre 200 000 et 300 000 anciens sites industriels et d'activités de service, dont un millier appellent une action à court terme au vu des dangers immédiats qu'ils représentent. L'étude et la réhabilitation de ces sites pollués représentent un enjeu majeur de par leur nombre élevé, les risques sanitaires graves qu'ils représentent et les coûts importants impliqués.

Audit de site et réhabilitation sont basés sur un calcul de concentration des polluants présents sur le site, suivi le cas échéant d'une détermination des zones à dépolluer. Ces estimations, nécessaires à toute intervention, restent souvent empreintes d'empirisme malgré l'ampleur des moyens techniques et financiers mis en œuvre. Toute erreur à cette étape peut avoir des conséquences graves en termes sanitaires ou financiers.

La pollution des sols se développe dans un milieu généralement complexe et hétérogène. Les sources pouvant être multiples, la modélisation fine s'avère délicate. La reconnaissance d'un site repose donc fréquemment sur l'*interprétation* des informations liées à l'historique du site, éventuellement complétée par le prélèvement orienté de quelques échantillons. Pourtant, dans le cas de pollutions anciennes - qui constituent une part importante du problème - cet historique est souvent connu de façon incomplète. La fiabilité d'un diagnostic de pollution de site à partir de prélèvements ponctuels est par ailleurs difficile à valider : si un audit concluant à une pollution ne suffit pas à en déterminer les limites et en contrôler les évolutions, un audit négatif ne peut exclure une présence de pollution. Le questionnement fondamental de la valeur d'un résultat est souvent omis, bien que la connaissance de l'incertitude liée aux estimations soit essentielle.

L'intégration d'une description *objective* de l'état de pollution d'un site permettrait de limiter les risques liés à une interprétation trop précoce du contexte du site. Là se situe notre objectif : enrichir l'approche actuelle par une réflexion méthodologique fournissant, à partir d'un échantillonnage réfléchi du site, des méthodes permettant d'en décrire objectivement le niveau de pollution ainsi que les incertitudes liées à cette estimation. A ce stade le praticien ne manquera pas de rappeler qu'il

[1] Les exemples sont nombreux. Citons l'accident de SEVESO en Italie le 10 juillet 1976 où un emballement réactionnel dans une unité de chlorophénols a entraîné le rejet à l'atmosphère de dioxines, ou l'affaire du "Love Canal" aux USA en 1980, lors de laquelle la découverte de 20 000 tonnes de déchets toxiques au fond d'une rivière a conduit à l'évacuation des habitations voisines.

est dommageable de se priver d'informations liées à l'historique de la pollution, à son évolution, à ses sources possibles en ne se basant que sur quelques analyses ponctuelles. Nous lui donnons raison, il est nécessaire de prendre en compte cette information historique essentielle. Cependant, il est conceptuellement important de ne pas utiliser cette information trop en amont, car cela reviendrait à accorder une place privilégiée à une modélisation du phénomène - que celle-ci provienne d'équations physiques, soit construite par l'expérience ou liée aux interactions géochimiques entre les différents composés en présence - au détriment des *données*.

Les méthodes géostatistiques visent à décrire des phénomènes naturels corrélés dans l'espace et éventuellement le temps, à quantifier l'incertitude d'estimations de ces phénomènes réalisées à partir d'un échantillonnage en général très fragmentaire. Elles sont appliquées dans des domaines variés : mine, pétrole, hydrogéologie, topographie, météorologie, pédologie, halieutique, finance, etc, ainsi que l'environnement qui nous intéresse plus spécifiquement. L'utilisation de la géostatistique est relativement répandue dans ce domaine, en particulier pour l'étude de pollution des sols - nombreuses publications, congrès internationaux spécialisés, etc. Cette utilisation se concentre souvent sur la mise au point de méthodes plus ou moins élaborées d'estimation d'une pollution, en prenant en compte par exemple des informations auxiliaires, mais les difficultés inhérentes à ce type de variables sont plus rarement discutées : variables fortement contrastées, incertitudes liées à l'échantillonnage de sols souvent hétérogènes. Nous nous proposons d'aborder ces difficultés, en étudiant en particulier l'étape essentielle de l'inférence d'une structure spatiale à partir d'échantillons ponctuels. Des outils permettent ensuite d'améliorer la compréhension de ces pollutions, d'estimer les concentrations en place et de sélectionner les zones de dépassement de seuils de pollution.

La **première partie** du mémoire décrit le problème posé, le principe de l'audit de site et les difficultés rencontrées en pratique lors de la caractérisation d'une pollution de sol. Les deux-tiers des pollutions industrielles de sols recensées en France proviennent d'hydrocarbures et de solvants chlorés, et sont pour une bonne part issues d'industries liées à la pyrolyse de la houille. Ce mémoire est basé sur l'étude de la pollution en Hydrocarbures Aromatiques Polycycliques (HAP) de deux anciens sites de cokeries, dont les caractéristiques sont présentées. Les objectifs du travail, qui découlent des problèmes énoncés, sont détaillés. Phase essentielle et délicate de toute étude géostatistique, l'inférence d'une structure spatiale occupe une position centrale dans ce mémoire. Un chapitre en reprend la problématique et compare plusieurs outils variographiques afin de déterminer ceux qui sont appropriés aux variables fortement contrastées rencontrées sur les sites.

La pertinence de l'utilisation de méthodes géostatistiques dépend de la mise en évidence ou non d'une structure spatiale pour les concentrations en polluants. Nous avons pu mettre en évidence de telles structures sur l'un des sites, auquel est consacrée la **seconde partie** du mémoire. Elle commence par une présentation du site et une analyse exploratoire détaillée des polluants qui illustre comment, par des considérations simples et sans modélisation, il est possible d'apporter nombre d'informations relatives à l'état de la pollution. La prise en compte du caractère spatialisé intervient ensuite ; une comparaison pratique des outils variographiques, qui nécessitent des hypothèses plus ou moins fortes, est effectuée. Suite à cela, nous présentons une estimation des concentrations en place qui intègre les informations apportées par des variables corrélées à la concentration. En particulier, l'utilisation de données auxiliaires peu coûteuses, telle que la présence d'indices organoleptiques, est discutée. Un chapitre est consacré au calcul de probabilités de dépassement de seuils ; sa pertinence en vue d'une sélection de zones à dépolluer est illustrée et plusieurs méthodes de calcul sont comparées. Le choix important d'un support de dépollution cohérent avec les techniques à

mettre en œuvre et les objectifs de la réhabilitation est discuté.

Dans certains cas, l'absence de structure spatiale ôte toute pertinence à une estimation locale des concentrations ainsi qu'à la sélection de zones à traiter. Il est néanmoins alors essentiel de pouvoir détailler les différentes sources d'hétérogénéités des concentrations, afin d'étudier si certaines peuvent être réduites. Cette étude est effectuée pour le second site, auquel est consacrée la **troisième partie**. Il ressort notamment que l'analyse chimique classique d'un échantillon ponctuel de sol s'avère peu représentative de la concentration au point de prélèvement ; ces analyses sont comparées à des investigations plus légères par analyse des gaz intersticiels et utilisation de kits chimiques.

Sous-jacent à l'ensemble du mémoire, l'aspect méthodologique a pour but d'attirer l'attention du praticien sur plusieurs points auxquels il doit réfléchir en pratique face à un site, de donner des clés de réponse aux questions qu'il pourra se poser. En outre, au cours du mémoire les aspects liés à la méthodologie géostatistique sont autant que faire se peut dissociés de l'application concrète de ces méthodes et des résultats qui en découlent pour les sites particuliers traités.

Deuxième partie

Position du problème et éléments méthodologiques

Chapitre 1

Position du problème

Sommaire

Partant de l'approche préconisée en France, ce chapitre pose le problème de la caractérisation d'une pollution industrielle de sols. Le cas particulier de la pollution d'un ancien site de cokerie par des hydrocarbures aromatiques polycycliques est présenté. Le contexte de la thèse ainsi que ses objectifs sont exposés.

"On définira un site contaminé comme un espace où se sont exercées ou s'exercent des activités de production, de transformation, de transport, de service [...] et qui, du fait de négligence, de défaut de conception ou de maintenance, conduit à l'apparition de dommages et risques immédiats ou différés pour les usagers, les riverains actuels ou futurs et l'environnement" [Ricour (1993)].

1.1 Etat des lieux des friches industrielles en France

Les pays industrialisés n'ont pris conscience des problèmes de pollution que depuis 25 ans environ, tandis que leur développement industriel remonte dans de nombreux cas à plus d'un siècle et demi. Cela se traduit en France par l'adoption de la loi du 15 juillet 1975 relative à l'élimination des déchets et à la récupération des matériaux, puis de la loi du 19 juillet 1976 relative aux installations classées pour la protection de l'environnement. Ajoutons à ces dates celle du 2 février 1995, avec l'adoption de la loi relative au principe "pollueur-payeur"[1]. Le changement progressif des mentalités qui découle de ces lois influence de plus en plus les nouveaux projets industriels, et ce dès leur conception ; néanmoins, il reste à gérer un passé industriel aux répercussions environnementales souvent dramatiques.

[1] Article 1 de la loi Barnier, introduit à l'article L. 200-1, alinéa 5, du code rural : "le principe pollueur-payeur, selon lequel les frais résultant des mesures de prévention, de réduction de la pollution, et de lutte contre celle-ci doivent être supportés par le pollueur".

Un recensement systématique des sites potentiellement pollués est en cours. Cet *inventaire national* a répertorié à ce jour 896 sites considérés comme appelant à court terme une action administrative pour remédier aux dangers qu'ils présentent. Ils constituent la base du programme d'intervention des DRIRE (Direction Régionale de l'Industrie, de la Recherche et de l'Environnement). S'ajoute à cela un *inventaire historique* des anciens sites industriels et d'activités de service qui, à la différence de l'inventaire national, a vocation à être exhaustif et répertorié dans la base de données BASIAS (Base des Anciens Sites Industriels et d'Activités de Service). Quand l'ensemble de la France sera couvert, ce qui est prévu pour 2005, on attend 200 000 à 300 000 sites recensés. Finalement, un *inventaire des sites en activité* est en cours, avec un objectif de 1300 études de sols à réaliser d'ici 2001 [Hugon & Lubek (2000)]. Parallèlement à ces recensements, la recherche appliquée aux phénomènes de pollution de sites s'est développée, contribuant à une meilleure compréhension des phénomènes.

Pour caractériser une pollution, plusieurs approches ont vu le jour. En France, l'*audit de site* se base sur une étude des risques qui repose sur la détermination de trois facteurs :

- la *source du risque*, c'est-à-dire les produits polluants présents, leurs teneurs et leurs propriétés physico-chimiques,
- les *voies de transfert*, qui permettent d'apprécier les évolutions et modifications potentielles de la pollution en fonction des conditions du milieu - spéciation, présence d'une nappe -,
- les *cibles* de la pollution : fréquentation du site, exposition, etc.

L'audit d'un site débute par l'établissement d'un diagnostic environnemental. Cette *évaluation simplifiée des risques* comprend le rassemblement des informations disponibles sur le site et le plus souvent une phase d'investigation sommaire [BRGM (2000a)]. Elle permet ainsi la détermination des trois facteurs mentionnés et le classement du site en fonction du risque qu'il représente : site nécessitant des investigations approfondies, site devant faire l'objet d'une surveillance appropriée ou site sans problème, dit "à banaliser". Si nécessaire, une *évaluation détaillée des risques* ou étude d'impact est alors réalisée [BRGM (2000b)], et en fonction de celle-ci une réhabilitation du site est envisagée.

1.2 Problèmes rencontrés

1.2.1 Sols pollués

A des fins réglementaires, le sol est considéré comme "l'ensemble de la zone non saturée comprise entre la surface topographique et le niveau de battement de la nappe" [Pellet & Laville-Timsit (1993)].

La pollution des sols s'inscrit dans des échelles de temps très longues ; la formation d'un sol est un processus lent, ce qui en fait une ressource non renouvelable pour l'échelle de temps beaucoup plus courte des activités humaines. Par ailleurs, constituant le principal réceptacle des pollutions d'origine industrielle, les sols ont accumulé des quantités considérables de polluants à proximité des zones industrielles et urbaines. Contrairement aux pollutions de l'eau ou de l'air, qui s'exercent dans des milieux relativement homogènes, la pollution des sols se développe dans un milieu généralement complexe et hétérogène ; rares sont les sols présentant des caractéristiques homogènes sur de grandes

zones [Gailland & Gissler (1994)]. Il en découle un contraste élevé des concentrations. En outre, produits polluants et localisations varient en fonction de l'histoire du site, cette hétérogénéité pouvant être accrue par un remaniement des sites, superficiel ou en profondeur.

Par conséquent, la fiabilité d'un diagnostic de pollution de site à partir de quelques prélèvements est par nature difficile à valider : de même qu'un audit concluant à une pollution ne permet pas toujours d'en déterminer les limites et d'en contrôler les évolutions, un audit négatif n'est fréquemment pas la garantie d'une absence de pollution. Il existe donc à la base une difficulté intrinsèque d'obtention d'une information technique objective et fiable, pourtant indispensable pour une gestion efficace du patrimoine [Hugon & Lubek (2000)].

1.2.2 Approche de l'échantillonnage

Audit et réhabilitation d'un site nécessitent une quantification de la pollution. Cela implique la localisation des produits polluants par classes de teneurs, ainsi qu'une délimitation des "taches" de pollution. L'échantillonnage de données précisant les concentrations en place est donc nécessaire.

Tout échantillonnage implique un coût. Ce coût résulte uniquement de sa réalisation dans le meilleur des cas, et il s'y ajoute celui de la dépollution lorsque celle-ci est détectée. Cette perspective, qui diffère par rapport à celle de domaines tels que la prospection minière, n'est pas sans conséquence sur la motivation de certains intervenants à échantillonner de façon détaillée ces sites. Par ailleurs, bien que nous n'y ayons pas été confronté dans ce travail, on constate parfois lors d'investigations l'implication de plusieurs intervenants dont l'expérience et les intérêts divergent souvent - propriétaire, autorité, bureau d'études, société de dépollution. Il en résulte des pratiques différentes qui ne peuvent que compliquer l'investigation.

Concernant la stratégie d'échantillonnage, l'utilisation des informations liées à l'historique est fréquemment privilégiée, fournissant à moindre coût des informations liées à la pollution. Cet historique est néanmoins souvent mal connu dans le cas de sites ayant un long passé industriel, et son utilisation exclusive génère un risque important de ne pas détecter certaines taches de pollution. Usuellement, l'échantillonnage vise à confirmer cette information historique ; cet échantillonnage étant orienté, préférentiel, il n'atténue pas le risque de non détection.

Les difficultés d'homogénéisation de matériaux hétérogènes rendent les analyses chimiques peu représentatives de la pollution en place, même à l'échelle de l'échantillon - carotte de sondage, fosse. Cela a suscité un intérêt croissant ces dernières années pour la mise en œuvre de méthodes permettant une meilleure représentativité de la pollution telles que les analyses de gaz, dont on attend beaucoup. Des méthodes moins lourdes telles que les kits chimiques ou enzymatiques sont également étudiées.

1.2.3 Sélection de zones à dépolluer

Suite à l'audit d'un site et selon l'ampleur des risques qu'il présente pour la santé humaine et l'environnement, son propriétaire se verra juridiquement imposer une dépollution permettant de réduire ces risques à un niveau acceptable. Cette dépollution nécessite l'estimation du volume

de sol contaminé, estimation actuellement réalisée de façon assez sommaire par les acteurs de la dépollution. L'absence de toute indication de la précision de cette estimation conduit à une incertitude de l'ampleur des coûts financiers liés à la dépollution.

En mine, lors d'une sélection des zones de teneur supérieure à une coupure, il existe un écart systématique entre les prévisions de teneurs et de volumes et les résultats d'exploitation, toujours défavorables. Cette dégradation des résultats se retrouve lorsque l'on est confronté à un site pollué. Elle découle de deux effets [Matheron (1984)] :

– **Effet de support**
L'histogramme des concentrations a même moyenne, que les concentrations soient calculées sur des échantillons ou sur des blocs de plusieurs mètres cubes. Par contre, les variances diffèrent selon le volume considéré, le *support*. La proportion de valeurs supérieures à un seuil fixé en dépend donc également (voir figure 1.1). Cet *effet de support* est indépendant de toute erreur d'échantillonnage. Sa prise en compte est essentielle, dans la mesure où l'unité de sélection utilisée lors de la dépollution est fréquemment bien supérieure à celle utilisée lors de l'échantillonnage.

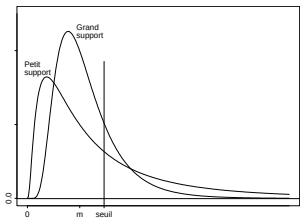


FIG. 1.1 – Histogramme des concentrations en fonction de la taille du support.

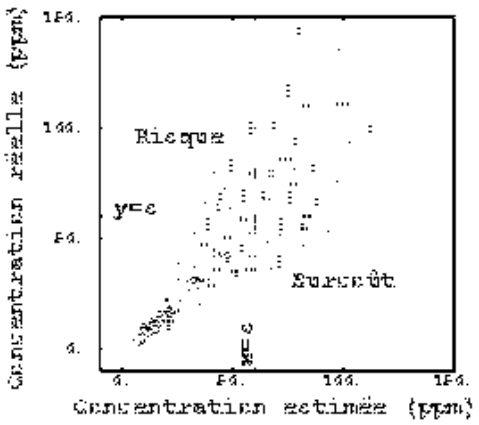


FIG. 1.2 – Illustration de l'effet d'information.

– **Effet d'information**
En pratique, la sélection ne s'effectue pas à partir de la concentration réelle, inconnue, mais au contraire à partir d'une estimation de cette concentration. La différence entre cette estimation et la concentration réelle, inévitable quelle que soit la qualité de l'estimateur et dont découle une perte d'efficacité de la sélection, est l'*effet d'information*. Il se traduit par deux types d'erreurs lors de la sélection des valeurs supérieures à un seuil s (voir figure 1.2). Si la concentration estimée est inférieure au seuil s tandis que la concentration réelle est supérieure, le point correspondant est supposé à tort comme non pollué, ce qui représente un risque sanitaire. D'un autre côté, si la concentration est estimée supérieure à s tandis qu'elle est en réalité inférieure, la dépollution inutile du point concerné est synonyme de surcoût.

1.3 Contexte de la thèse

Toute erreur de quantification de la pollution lors d'un audit de site peut avoir des conséquences graves en termes sanitaires ou financiers. Moyennant une modélisation appropriée, la géostatistique permet une estimation des concentrations en place, tout en fournissant une indication sur la précision de l'estimation, les incertitudes sur les teneurs et tonnages estimés ainsi que l'évaluation du risque de laisser en place des zones de concentration supérieure à un seuil d'intervention fixé.

Afin de mettre en œuvre ces méthodes et d'étudier sous quelles conditions elles améliorent une caractérisation de pollutions industrielles de sols, nous nous intéressons au cas particulier de la pollution de deux anciens sites de cokeries par des hydrocarbures aromatiques polycycliques.

1.3.1 Sites de cokeries

Destinées à produire du coke - charbon carbonisé à haute température - utilisé par l'industrie sidérurgique pour les hauts fourneaux, les cokeries étaient relativement répandues au début du 20$^{\text{ème}}$ siècle dans les pays industrialisés. Quelques-unes de ces usines sont encore en activité aujourd'hui.

La cokéfaction consiste à chauffer un mélange de 2 à 3 types de charbon jusqu'à 1000°C environ, afin de fournir du coke[2]. A titre d'exemple, une cokerie classique produit par tonne de houille : 735 à 800 kg de coke, 300 à 340 m^3 de gaz, 30 à 35 kg de goudron brut, 7 à 12 kg de benzol et 5 à 10 kg de sulfate d'ammonium [Piel et al. (1997)]. Ces produits sont normalement récupérés pour être traités séparément.

Un four à coke ne constituant pas un réacteur continu et étanche, il y a émission d'hydrocarbures par fuites, lors du chargement et du défournement des fours, de l'extinction humide du coke, de la maintenance, etc. Ces émissions de gaz et de poussières sont essentiellement diffuses.

1.3.2 Pollutions par des HAP

Les Hydrocarbures Aromatiques Polycycliques (HAP) sont produits en quantité par l'activité humaine, essentiellement au cours des processus de pyrolyse et de combustion incomplète de matières organiques mis en œuvre dans l'industrie, le transport et le chauffage. Ils constituent en particulier 45 à 70 % des goudrons de houille, co-produits des activités anciennes de la pyrolyse du charbon - cokeries et usines à gaz - et se retrouvent dans les sols de ces anciennes friches contaminées par des goudrons.

Les HAP regroupent des substances chimiques constituées de 2 à 6 cycles benzéniques juxtaposés. Ils sont classés en HAP légers - 2 et 3 cycles - ou lourds - 4 à 6 cycles -, de poids atomiques croissants et ayant des caractéristiques physico-chimiques et toxicologiques différentes. Les pouvoirs mutagène[3] et cancérogène des HAP, qui apparaissent à partir de 4 cycles et sont particulièrement

[2] Le fonctionnement d'une cokerie ainsi que son impact environnemental sont décrits dans l'annexe B.
[3] Cette toxicité agit sur les chromosomes.

marqués pour les HAP à 5 et 6 cycles, expliquent l'attention portée à ces composés. Ils ont également des effets nocifs sur le système immunitaire, effets qui sont corrélés avec le pouvoir cancérogène.

Le benzo(a)pyrène est le plus étudié des HAP, étant souvent considéré comme leur traceur [Bouchez et al. (1996)]. Bien qu'il en existe beaucoup plus, les analyses chimiques réalisées lors d'études environnementales ne concernent généralement qu'une sélection de HAP[4]. Souvent, il s'agit de la liste des 16 HAP sélectionnés par l'agence américaine pour la protection de l'environnement (US EPA), même si d'autres sélections existent : 10 HAP pour la liste néerlandaise, 6 pour l'OMS (voir tableau 1.1).

Liste US EPA	Abrév.	Cycles	Poids moléculaire	Liste Pays-Bas	Liste OMS
Naphtalène	Nap	2	128.18	*	
Acénaphtylène	Acy	3	152.20		
Acénaphtène	Ace	3	154.21		
Fluorène	Fle	3	166.22		
Anthracène	Ant	3	178.24	*	
Phénanthrène	Phe	3	178.24	*	
Fluoranthène	Flt	4	202.26	*	*
Pyrène	Pyr	4	202.26		
Benzo(a)anthracène	Baa	4	228.30	*	
Chrysène	Cry	4	228.30	*	
Benzo(a)pyrène	Bap	5	252.32	*	*
Benzo(b)fluoranthène	Bbf	5	252.32		*
Benzo(k)fluoranthène	Bkf	5	252.32	*	*
Dibenz(a,h)anthracène	Dba	5	278.36		
Benzo(ghi)pérylène	Bgh	6	276.34	*	*
Indéno(1,2,3-c,d)pyrène	Inp	6	276.34	*	*

TAB. 1.1 – Liste des HAP retenus par l'US EPA (16), les Pays-Bas (10) et l'OMS (6). Abréviations, nombres de cycles et poids moléculaire en $g.mol^{-1}$ [d'après Montgomery (1996)].

En mettant en œuvre différents outils statistiques, Oosterbaan-Eritzpokhoff (2000) montre que les 16 HAP de la liste US EPA peuvent se regrouper en trois catégories : légers, intermédiaires et lourds (voir tableau 1.2), les corrélations entre les concentrations en HAP d'une même catégorie étant fortes, particulièrement pour les HAP intermédiaires et lourds. Dans chaque catégorie un HAP peut servir de substance indicatrice satisfaisante de la pollution : le naphtalène pour les légers, le fluoranthène pour les intermédiaires et le benzo(a)pyrène pour les lourds. Le dibenz(a,h)anthracène étant moins bien corrélé au benzo(a)pyrène que les autres HAP lourds, l'auteur recommande son étude complémentaire, étant donné son caractère très cancérogène.

Solides à température ambiante, les HAP se trouvent dans les goudrons sous forme liquide, pâteuse ou solide. Seuls les HAP les plus légers, notamment le naphtalène, sont solubles, susceptibles

[4]Le protocole de préparation et d'analyse des HAP utilisé pour ce travail est décrit dans l'annexe A.

Catégorie	Poids moléculaire	Composés	Indicateur
HAP légers	< 170	Nap, Acy, Ace, Fle	Nap
HAP intermédiaires	de 170 à 210	Ant, Phe, Flt	Flt
HAP lourds	> 210	Pyr, Baa, Cry, Bap, Bbf, Bkf, Dba, Bgh, Inp	Bap (Dba)

TAB. 1.2 – Associations des 16 HAP de la liste US EPA en trois catégories, pour des sites de cokeries [d'après Oosterbaan-Eritzpokhoff (2000)].

d'être volatilisés et biodégradés[5]. Les HAP sont très stables et s'adsorbent[6] en grande partie à la matière solide, ce qui rend leur migration vers les couches inférieures faible.

Outre les HAP, le phénol et les composés phénoliques dérivés sont également produits par la pyrolyse du charbon, ce qui a motivé leur suivi sur les sites investigués. Les phénols agissent sur le système nerveux central - affaiblissement, problèmes de vision, pertes de conscience -, provoquent des dermatoses et des problèmes oculaires et peuvent se révéler être un poison violent, létal pour l'homme à partir de 14 $mg.kg^{-1}$. Alors que le phénol est très soluble dans l'eau, ses dérivés le sont très peu. Etant par conséquent plus fortement adsorbés à la matière organique - sous forme de "résines phénoliques" -, ce sont surtout ces derniers que l'on retrouve dans les sols. Les phénols sont plus biodégradables que les HAP, ce phénomène étant cependant partiellement inhibé si leur présence est trop importante dans le sol.

1.3.3 Cadre scientifique

Ce travail est une collaboration entre le Centre de Géostatistique de l'Ecole des Mines de Paris et le Centre National de Recherche sur les Sites et Sols Pollués (CNRSSP). Ce dernier est né en 1996 d'une volonté politique nationale et d'une initiative collective pour parvenir à résoudre à relativement court terme les problèmes que pose l'existence de nombreux sites pollués sur le territoire national. Par un accord de collaboration entre différents partenaires industriels, universitaires et associatifs, il travaille sur différents thèmes. Outre celui concernant l'investigation et la caractérisation des sites pollués, auquel cette thèse est rattachée, le CNRSSP s'intéresse à travers ses programmes au comportement des polluants dans les sols et les aquifères, au comportement des sédiments et des écosystèmes aquatiques ainsi qu'à l'impact des polluants sur la santé humaine et les écosystèmes.

Ce cadre de recherche, non directement industriel, fut un avantage ; il a permis la mise en œuvre de mailles d'échantillonnage régulières, la réalisation de resserrements de sondages et l'examen détaillé des données : variabilité due à l'homogénéisation délicate et à petite échelle. Par contre, le coût élevé des opérations a imposé un nombre d'échantillons restreint, des analyses chimiques uniques plutôt que par triplicat[7] et la difficulté de programmer une campagne de validation des méthodes.

[5] La biodégradation consiste en la décomposition des molécules organiques par les micro-organismes présents dans le sol. Outre la présence de ces micro-organismes, elle dépend essentiellement de la température et de la quantité d'oxygène.

[6] L'adsorption se définit comme la rétention plus ou moins réversible du polluant sur la matrice solide.

[7] Ce procédé couramment préconisé consiste à répéter trois fois l'analyse avant de fournir comme concentration finale la moyenne des trois.

1.4 Objectifs

1.4.1 Caractérisation des HAP et examen de variables qualitatives

Une bonne connaissance de l'état de la pollution est tout d'abord indispensable. Pour cela, les statistiques élémentaires des HAP et les corrélations sont étudiées, ainsi que les liens avec l'information historique.

Lors des campagnes d'investigation, plusieurs indices qualitatifs sont dérivés de la description des échantillons : présence d'odeur, de traces de goudron, de débris de maçonnerie, etc. Une étude détaillée de ces indices est menée. En cas de corrélation avec les concentrations en HAP, l'utilisation des indices qualitatifs permettra de guider l'échantillonnage de ces derniers et d'améliorer à moindre coût leur estimation. On comparera par ailleurs les concentrations en HAP à des analyses de gaz et des mesures par kits chimiques réalisées sur les sites suivis.

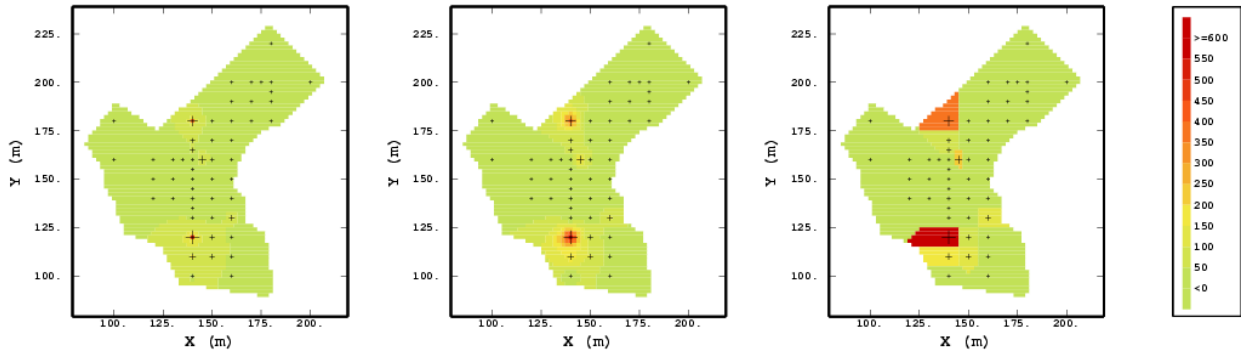
1.4.2 Estimation des concentrations en place

La quantification du niveau de pollution implique une estimation des concentrations en place. Le praticien pourra s'interroger sur l'intérêt d'une méthode d'interpolation géostatistique, alors qu'il existe plusieurs techniques d'interpolations simples et rapides. Il m'a semblé utile d'illustrer ici les résultats fournis par deux méthodes classiques d'interpolation : l'inverse des distances et le plus proche voisin. L'interpolation par inverse des distances estime la concentration en un point x_0 par une combinaison linéaire des concentrations des N données disponibles x_i dont les coefficients pour chaque point i sont inversement proportionnels à la distance entre x_0 et x_i. Autrement dit, plus une donnée x_i est proche de x_0 et plus elle a d'influence sur la valeur interpolée. Il est possible de considérer une puissance de la distance d, fréquemment 2. L'interpolation par plus proche voisin affecte quant à elle à x_0 la valeur de la variable au point de donnée le plus proche.

Les deux méthodes sont illustrées à la figure 1.3 pour le benzo(a)pyrène sur un des sites. Deux exemples de puissance sont donnés pour l'inverse des distances. Bien que les valeurs les plus fortes se retrouvent sur les différentes interpolations, les cartes ont des allures sensiblement différentes : très lisse dans le cas de l'inverse des distances, en escalier dans le cas du plus proche voisin[8], l'inverse des distances quadratiques constituant un cas intermédiaire. Par conséquent, l'extension des taches de pollution varie fortement en fonction de la méthode d'interpolation choisie. L'extrapolation des valeurs situées en bordure est visible.

Le choix de la méthode d'interpolation est un problème délicat. Sur quoi le baser, comment le justifier ? Dans le meilleur des cas, il est effectué par référence aux propriétés physiques attendues de la variable ; on choisira par exemple un interpolateur plutôt lisse tel que l'inverse des distances si l'on doit interpoler une piézométrie. Cette démarche ne permet cependant pas de répondre à la question fondamentale suivante : quelle valeur accorder au résultat de l'interpolation ?

[8] L'interpolation par plus proche voisin présente souvent l'inconvénient d'augmenter le *biais conditionnel* : dans le cas de phénomènes avec effets de bord - qui présentent des zones de transition entre valeurs faibles et fortes -, elle surévalue les concentrations estimées au voisinage des données de concentration élevée et les sous-évalue près des concentrations faibles.



(a) Bap - $1/d$ (b) Bap - $1/d^2$ (c) Bap - Plus proche voisin

FIG. 1.3 – Interpolations classiques pour le benzo(a)pyrène sur le site Y, par inverse des distances, de puissances égales à 1 et 2, et plus proche voisin. Représentation proportionnelle des concentrations analysées. Echelle en $mg.kg^{-1}$.

Moyennant certaines hypothèses et l'existence de structures spatiales pour les HAP, la géostatistique apporte avec le krigeage une réponse à ces questions ; elle permet en outre de quantifier l'incertitude associée à l'estimation. L'utilisation combinée des HAP corrélés est présentée, ainsi que celle des variables auxiliaires corrélées aux concentrations en HAP. On étudie finalement comment une investigation fine et peu coûteuse de variables auxiliaires peut fournir une première connaissance de l'état de pollution du site et un guide pour l'implantation des prélèvements destinés à une analyse chimique classique des sols, plus coûteuse.

1.4.3 Probabilité de dépassement de seuils de pollution

Lors d'une dépollution de site, il est important de savoir où diriger l'effort de dépollution en fonction du seuil d'intervention et du risque que l'on s'accorde de laisser en place une pollution supérieure au seuil fixé. Pour cela, une idée consiste à sélectionner par seuillage, à partir de l'estimation du polluant[9], les valeurs estimées supérieures à la valeur d'intervention. Une carte d'estimation par krigeage de la concentration représente la meilleure estimation linéaire de cette concentration, au sens de la variance d'erreur minimale. Néanmoins, cette carte gomme les variabilités locales, qui existent mais que les données ne permettent pas d'approcher. Ces variabilités peuvent être simulées, comme l'illustre la figure 1.4 pour la concentration en phénanthrène sur le site Y. Les simulations diffèrent sensiblement de l'estimation de ce HAP par krigeage, bien que chacune des cartes respecte les concentrations analysées aux points expérimentaux. Par conséquent, si l'on ne veut risquer des erreurs de sélection importante et sachant qu'il existe toujours une différence entre valeur estimée et valeur réelle - effet d'information -, les variabilités locales doivent être modélisées.

La concentration exacte restant inconnue, une carte de la probabilité de dépasser un seuil de pollution en chaque point du site permet de sélectionner les zones de concentrations supérieures au

[9] Afin d'alléger le texte, plutôt que de parler d'"estimation de la concentration d'un polluant", ou d'un HAP, nous sous-entendrons fréquemment, lorsque cela ne sera pas source d'ambiguïté, le terme concentration pour nous en tenir à "estimation du polluant".

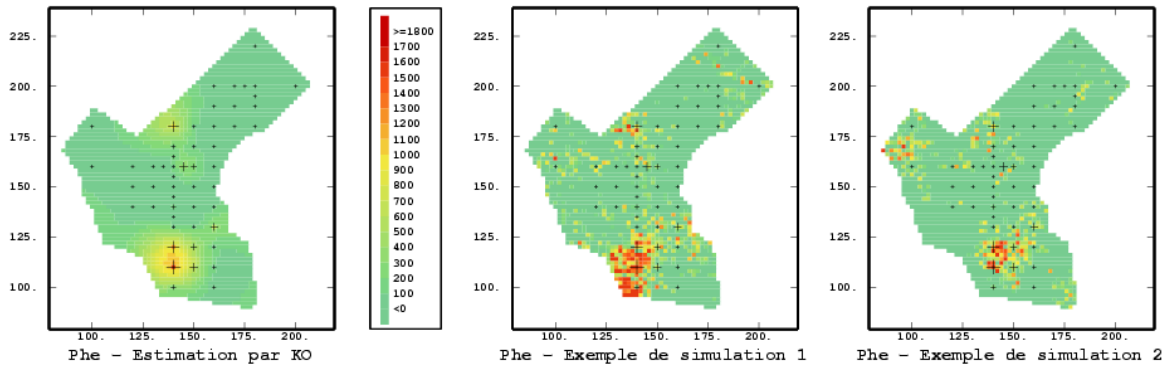


FIG. 1.4 – Estimation par krigeage ordinaire de la concentration en phénanthrène (en ppm) sur le site Y, exemples de simulations. Echelle en $mg.kg^{-1}$.

seuil, tout en ayant connaissance du risque encouru de laisser en place une valeur supérieure à ce seuil, qui existe toujours. De façon assez intuitive, cette carte de probabilité peut être obtenue à partir des simulations. Ces dernières reproduisant les variabilités locales, en considérant un nombre de simulations suffisamment grand, il est possible de se faire une bonne idée du niveau de variabilité. La probabilité de dépassement de seuil s'obtient en calculant en chaque point la proportion des cartes simulées dépassant le seuil. Demougeot-Renard & de Fouquet (2000) présentent un exemple pour l'estimation de volumes de sols contaminés par des HAP. Néanmoins, cette démarche repose sur une hypothèse de loi multigaussienne et l'obtention de simulations conditionnelles est une démarche assez lourde. D'autres méthodes, parfois appelées "non paramétriques", reposent sur le calcul d'indicatrices du dépassement du seuil. Entre autres inconvénients, ces méthodes génèrent une perte d'information - le codage d'une variable continue sous forme d'indicatrice - et nécessitent de tout recommencer chaque fois que le seuil d'intérêt est modifié [Liao (1990), Chilès & Delfiner (1999)] ; pour ces raisons, nous les avons écartées dans ce travail.

Nous utilisons une approche basée sur la modélisation de la loi spatiale. Deux estimations des probabilités de dépassement *ponctuel* des seuils sont envisagées : par *krigeage disjonctif* et par *espérance conditionnelle* [Matheron (1976, 1978b)]. Le terme *ponctuel* signifie que le volume auquel nous nous intéressons est identique à celui des échantillons ; cela présente un intérêt dans la mesure où si contrôle de dépollution il y a, celui-ci est effectué ponctuellement. Néanmoins, ce calcul ne tient pas compte de la taille du support utilisé lors de la dépollution. Pour cela, des modèles de changement de support sont appliqués.

1.4.4 Inférence de la structure spatiale

Sous-jacente à l'ensemble de ce travail, la question de l'inférence de la structure spatiale a constitué une partie importante des développements. A partir d'un échantillonnage fragmentaire de la variable d'intérêt, elle a pour objectif d'approcher sa structure spatiale puis de la modéliser. Dans le cas de variables présentant un fort contraste entre les teneurs, le manque de robustesse du variogramme classique nous a conduit à examiner les propriétés du variogramme déduit de la covariance non centrée et des variogrammes pondérés ; c'est l'objet du prochain chapitre.

Les structures spatiales des HAP présentent systématiquement une variabilité importante à petite échelle qui trouve son origine dans plusieurs sources d'hétérogénéité des concentrations : difficultés d'homogénéisation, variabilité à petite distance. Nous y reviendrons pour le site X, qui a permis d'analyser de façon détaillée ces problèmes.

Chapitre 2

Inférence d'une structure spatiale

Sommaire

L'inférence structurale est une étape cruciale de toute étude géostatistique. Sa problématique peut être résumée en deux questions. Peut-on raisonnablement estimer le variogramme régional à partir du variogramme expérimental ? Par ailleurs, existe-t-il des outils variographiques plus appropriés que d'autres, qui permettent d'améliorer cette estimation ? Nous récapitulons quelques voies permettant d'estimer de façon robuste la structure d'un phénomène. Une attention particulière est portée au variogramme déduit de la covariance non centrée, qui a suscité un travail méthodologique plus approfondi, ainsi qu'aux variogrammes pondérés. L'annexe C reprend les notions géostatistiques nécessaires à la compréhension du mémoire.

2.1 Inférence structurale

La variable régionalisée $z(x)$, connue en N points $x(i)\,,\, i = 1, \ldots, N$, est considérée comme une réalisation d'une fonction aléatoire intrinsèque $Z(x)$. Usuellement, le variogramme régional est approché par le variogramme expérimental classique

$$\gamma^{\star}(h) = \frac{1}{2N(h)} \sum_{x_j - x_i \sim h} [z(x_j) - z(x_i)]^2 \,,$$

$N(h)$ représentant le nombre de couples de points distants de h.

L'instabilité du variogramme expérimental, en particulier dans le cas de distributions dissymétriques, a été illustrée à maintes reprises [voir Srivastava R.& Parker H. (1989), Chilès & Delfiner (1999)]. Elle provient essentiellement du fort contraste entre quelques valeurs très élevées et le reste de la distribution, constituée de valeurs faibles. Cela est a fortiori le cas lorsque le nombre N de données est faible. Dans un cadre "gaussien", les incréments quadratiques $[Z(x_j) - Z(x_i)]^2$

suivent une distribution chi-deux avec un degré de liberté. Cette distribution est dissymétrique, et un grand nombre de couples est nécessaire pour que la moyenne expérimentale soit un bon estimateur de la moyenne théorique. Il est par conséquent apparu nécessaire de trouver des outils structuraux moins instables que le variogramme classique.

2.1.1 Estimateurs robustes de la variable brute

Avant de présenter différents types d'estimateurs introduits pour pallier ces difficultés, revenons sur un point de vocabulaire. Un estimateur est dit *résistant* s'il est peu sensible à des changements, même importants, survenant sur un petit jeu de données ; typiquement, un estimateur sensible aux valeurs extrêmes tel que la moyenne n'est pas résistant, tandis que la médiane l'est. Par ailleurs, par rapport à un modèle donné *a priori*, un estimateur est dit *robuste* s'il ne conduit pas à des résultats aberrants lorsque les hypothèses du modèle ne sont plus tout à fait validées. La distinction entre ces deux notions étant néanmoins en grande partie arbitraire, nous les regrouperons pour ne plus parler que de robustesse.

De nombreux exemples de variogrammes robustes ont été proposés dans la littérature [pour un récapitulatif, voir par exemple Chilès & Delfiner (1999)]. Ils reposent généralement sur les principes suivants :

- Remplacer les incréments quadratiques par des puissances d'ordre inférieur, moins sensibles aux valeurs fortes - pour une puissance égale à 1, on obtient le variogramme d'ordre 1.
- Remplacer la moyenne des incréments par une autre statistique moins sensible aux valeurs fortes, par exemple la médiane.
- Ne pas tenir compte des valeurs d'incréments trop fortes.

L'objectif de ces variogrammes est de permettre une meilleure mise en évidence d'une éventuelle structure. Il faut néanmoins être prudent car, face à des distributions dissymétriques pour les incréments, en utilisant par exemple la médiane au lieu de la moyenne expérimentale, on n'estime plus le même paramètre. Le résultat n'est donc plus un estimateur[1] du variogramme $\gamma(h)$, ce qui doit être pris en compte si l'on ne veut pas que l'estimation de Z soit moins bonne.

D'autres outils tels que la covariance $C(h)$ ou le corrélogramme $\rho(h) = \frac{C(h)}{C(0)}$ sont également utilisés. Cependant, il n'est pas rare en pratique que l'hypothèse de stationnarité d'ordre deux (voir annexe C.4.1) qui leur est nécessaire soit supposée *a priori* ; si celle-ci fait défaut, le risque de biais lors du calcul de la moyenne ou de la variance intervenant explicitement dans ces outils est accru. Nous présentons au paragraphe 2.2 un outil de ce type, le variogramme déduit de la covariance non centrée $C(0) - C(h)$, en analysant l'importance de l'hypothèse de stationnarité d'ordre deux sous-jacente.

2.1.2 Transformées de la variable brute

Plutôt que de calculer la structure expérimentale directement sur la variable brute, il est possible de recourir à une transformée de celle-ci. Par exemple, dans le cas d'une variable dissymétrique,

[1] Afin d'éviter toute ambiguïté, on précisera fréquemment s'il s'agit de l'estimation du variogramme ou de Z.

on transformera la variable de façon à réduire l'influence des valeurs extrêmes. Ces transformées peuvent être utilisées dans l'idée d'explorer une éventuelle présence de structure. Si par contre on cherche à en déduire une estimation du variogramme de la variable brute, seule la vérification de certaines hypothèses permet un retour au variogramme de la variable brute sans risque de biais.

2.1.2.1 Indicatrices

Soit l'indicatrice de $Z(x)$ associée au seuil s

$$\mathbb{I}_s(x) = \mathbb{I}_{Z(x)\geq s} = \begin{cases} 1 & \text{si } Z(x) \geq s \\ 0 & \text{si } Z(x) < s \end{cases}$$

Par construction, l'indicatrice n'est pas sensible aux valeurs fortes, excepté lorsque le seuil devient lui-même élevé. Il en découle l'utilisation possible de tels variogrammes d'indicatrices comme variogrammes robustes. Il existe sous certaines hypothèses une relation entre covariance de l'indicatrice et covariance de la fonction aléatoire (FA) elle-même ; par exemple, si $Z(x)$ est une FA bigaussienne réduite, son corrélogramme $\rho(h)$ est lié à la covariance $C_0(h)$ de l'indicatrice de la médiane $s = 0$ par la relation

$$C_0(h) = \frac{1}{2\pi} \arcsin \rho(h) \tag{2.1}$$

2.1.2.2 Logarithme translaté

Le fort contraste entre valeurs fortes et faibles, qui rend instable le variogramme expérimental, peut être réduit en considérant, plutôt que la variable brute, son logarithme $L(x) = \ln(Z(x))$. Ce faisant, les écarts entre valeurs faibles se trouvent cependant accrus ; à la limite, pour les valeurs nulles de Z ce logarithme n'est plus défini. Il est par conséquent souhaitable de considérer $L(x) = \ln(1 + Z(x))$, qui conserve les valeurs nulles. Finalement, on gagnera souvent à introduire un paramètre m de l'ordre de grandeur de la moyenne de Z

$$L(x) = \ln\left(1 + \frac{Z(x)}{m}\right)$$

La valeur exacte de m est peu importante, ce paramètre servant surtout à réduire et adimensionner la variable. La structure des données log-translatées diffère de celle des données brutes, et le retour à la structure brute n'est pas évident ; Guiblin (1997) propose une formule de retour, mais celle-ci n'est pas exempte de biais. Sans aller jusqu'à envisager ce retour, cette transformée est un outil complémentaire d'analyse de la structure spatiale, et permet de comparer les structures de différentes variables.

2.1.2.3 Transformée gaussienne

En l'absence de données prenant la même valeur - par exemple au seuil de détection si nous étudions une concentration en polluant -, la transformée obtenue par anamorphose gaussienne est

symétrique. Tout comme la précédente, cette transformée diminue le contraste entre valeurs fortes et faibles et peut être utilisée de façon heuristique afin de compléter l'analyse structurale, mais une hypothèse de stationnarité d'ordre deux est nécessaire si l'on veut donner un sens à cette variable ; dans le cas contraire rien ne garantit l'existence d'un seul histogramme stationnaire dans le modèle.

2.2 Variogramme déduit de la covariance non centrée

2.2.1 Définition

Si $Z(x)$ est une FASt2, alors le recours à la covariance *non centrée* régionale

$$C_R(h) = \frac{1}{|D \cap D_{-h}|} \int_{D \cap D_{-h}} z(x) z(x+h)\, dx \tag{2.2}$$

est licite, et celle-ci peut être estimée sans biais par la covariance non centrée expérimentale

$$C^\star(h) = \frac{1}{N(h)} \sum_{x_i - x_j \sim h} z(x_i) z(x_j) \tag{2.3}$$

Il n'en est pas de même pour la covariance centrée $C_c(h)$, qui, contrairement au variogramme et à la covariance non centrée, nécessite explicitement l'estimation de la moyenne $m = \mathrm{E}[Z(x)]$, car $C_c(h) = C(h) - m^2$. De cette estimation simultanée de m et de la covariance non centrée découle un biais [Journel & Huijbregts (1978)].

$C^\star(0) - C^\star(h)$ est un estimateur non biaisé du variogramme, appelé dans la suite *variogramme déduit de la covariance non centrée*. Cet outil a été utilisé avec succès lors d'applications halieutiques à la place du variogramme classique, pour des variables fort contrastées et présentant des structures de portées courtes par rapport à la taille du champ, ne posant pas de problème de stationnarité [Guiblin et al. (1995)].

Nous verrons en pratique au chapitre 4 que $\gamma^\star(h)$ et $C^\star(0) - C^\star(h)$ peuvent présenter des différences sensibles. Cela a motivé une comparaison méthodologique plus poussée de ces deux outils, afin d'analyser dans quelle mesure leur utilisation combinée permet d'améliorer l'inférence de la structure spatiale [Jeannée & de Fouquet (2000b)].

Les différences entre les grandeurs régionales correspondant aux deux outils sont discutées ci-dessous. Ensuite, nous étudions la dispersion des nuées variographiques régionales pour différents types de fonctions aléatoires simulées non conditionnellement. Finalement, nous comparons pour ces fonctions aléatoires la structure expérimentale déduite d'un sous-échantillonnage régulier du champ à la structure régionale.

2.2.2 Comparaison de $\gamma_R(h)$ et $C_R(0) - C_R(h)$

Tout d'abord, l'estimateur $C^\star(0) - C^\star(h)$ possède-t-il les mêmes propriétés que le variogramme classique ? Par exemple, est-il, comme ce dernier, positif ? Plus généralement, que peut-on dire de

la différence entre l'expression de ces deux outils ? Il est important de noter que le terme $C^\star(0)$ de $C^\star(0) - C^\star(h)$, qui dépend du champ, est calculé une fois pour toutes et n'est pas modifié à chaque pas.

Sur l'exemple de la figure 2.1, pour $h = 4$ le variogramme déduit de la covariance non centrée vaut $C^\star(0) - C^\star(4) = \frac{2}{5} - 1 = -\frac{3}{5}$, qui est négatif, tandis que $\gamma^\star(4) = 0$! Cette valeur négative illustre un risque lié à une utilisation de $C^\star(0) - C^\star(h)$ à la place du variogramme classique[2]. Tout comme le variogramme classique, $C^\star(0) - C^\star(h)$ est symétrique.

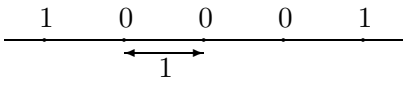


FIG. 2.1 – Variable régionalisée à 1D.

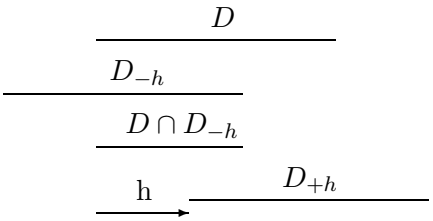


FIG. 2.2 – Domaine D à une dimension, translatés et sous-domaines.

Considérons les grandeurs régionales $\gamma_R(h)$ et $C_R(h)$ pour un domaine D dont certains sous-domaines sont schématisés à une dimension à la figure 2.2. Il découle de la définition de $\gamma_R(h)$ que

$$\gamma_R(h) = s_R^2(h) - C_R(h) \tag{2.4}$$

avec

$$s_R^2(h) = \frac{1}{2|D \cap D_{-h}|} \left[\int_{D \cap D_{-h}} z(x)^2 \, dx + \int_{D \cap D_{+h}} z(x)^2 \, dx \right]. \tag{2.5}$$

Donc, au lieu du terme $C_R(0)$ qui prend en compte l'ensemble du domaine D, le variogramme régional écrit sous la forme $s_R^2(h) - C_R(h)$ n'inclut dans son premier terme que les points intervenant à la distance h.

En outre, dans le cas de variables dissymétriques, de fortes valeurs de $s^2(h)$, causées par des valeurs de z élevées, peuvent accroître artificiellement la continuité des structures observée sur les covariances non centrées [Rivoirard & Bez (1997)].

2.2.3 Modèles de fonctions aléatoires

Nous traitons le cas de deux modèles de FA avec effets de bord, qui peuvent décrire le comportement de polluants, et ensuite deux modèles d'ensemble aléatoires illustrant le cas de variables indicatrices telles que des variables auxiliaires, ou le seuillage de polluants.

[2] Pour pallier le fait que les données en bordure ont un poids moindre il est possible de recourir à une pondération par échantillon de la covariance non centrée, de façon analogue à ce qui sera présenté à la section 2.3 ; nous obtenons alors $C^\star(0) - C^\star(4) = \frac{4}{7} - 1 = -\frac{3}{7}$, plus proche de 0 mais néanmoins encore différent de $\gamma^\star(4)$.

– **FA gaussienne**
Le cas d'une FA gaussienne $Y(x)$ d'espérance 0 et de variance unité sert de référence. Même si un tel modèle ne se rencontre pas directement lorsque l'on s'intéresse à des polluants dont les distributions sont souvent dissymétriques, une anamorphose gaussienne peut dans certains cas être utilisée pour se ramener à ce modèle de référence.
– **FA lognormale**
Une FA lognormale $Z(x)$ peut s'écrire

$$Z(x) = m\ e^{\sigma Y(x) - \frac{\sigma^2}{2}}$$

avec $Y(x)$ une FA gaussienne réduite de covariance $\rho(h)$, σ^2 la variance de $\log[Z(x)]$ - variance logarithmique -, et $m = \mathrm{E}[Z(x)]$. De plus, Z a pour variance

$$\mathrm{Var}[Z(x)] = m^2(e^{\sigma^2} - 1)$$

et comme covariance

$$C_Z(h) = m^2(e^{\sigma^2 \rho(h)} - 1) \tag{2.6}$$

On considère le cas où $m = 1$ et $\sigma = 1.50$, qui correspond à un coefficient de variation de 2.91, réaliste pour des polluants organiques.
– **Gaussienne seuillée**
Soit $Z(x) = \mathbb{1}_{Y(x)<0}$ l'indicatrice de la médiane d'une FA gaussienne $Y(x)$. La covariance de Z se déduit de celle de Y par la relation 2.1.
– **Polygones poissonniens**
Si N droites poissonniennes traversent le champ D, alors en attribuant aléatoirement aux polygones déterminés par ces droites les valeurs 0 et 1, soit $m - \sigma$ et $m + \sigma$ avec $m = \sigma = \frac{1}{2}$, la fonction aléatoire stationnaire résultante a pour covariance [Chilès & Delfiner (1999)]

$$C(h) = \sigma^2 e^{-2\lambda|h|} \tag{2.7}$$

où l'intensité λ du processus vaut $\frac{N}{\mathcal{P}(\mathcal{D})}$, $\mathcal{P}(\mathcal{D})$ étant le périmètre de D.

2.2.4 Simulations

Un modèle sphérique est utilisé pour la simulation des FA gaussiennes sur un carré de coté $L = 30$; ce choix de taille est guidé par la volonté de conserver des temps de calcul raisonnables tout en autorisant une diversité suffisante des structures simulées. La portée a vaudra successivement $L/6$, $L/2$, L et $2L$, la non stationnarité à l'échelle du champ étant donc de plus en plus marquée.

Les simulations sont réalisées par transformée de Fourier discrète [Pardo-Igúzquiza & Chica-Olmo (1993), Chilès (1995)] sur une grille de 31×31 points - schéma fermé. La méthode des bandes tournantes [voir par exemple Lantuéjoul (1994)], également mise en œuvre, conduit à des résultats analogues pour des temps de calcul plus longs. Tous les calculs variographiques sont effectués dans une seule direction, parallèle à un coté de la grille.

2.2.5 Nuées variographiques

Tout comme la nuée du variogramme représente les valeurs $\frac{1}{2}[z(x_j) - z(x_i)]^2$ en fonction de $|x_j - x_i|$, on peut représenter la nuée correspondante pour la covariance non centrée, composée des valeurs $z(x_i)\ z(x_j)$.

Variogramme et covariance non centrée ne sont pas sensibles aux mêmes relations entre les valeurs de la variable régionalisée. Tandis que ce sont les paires "valeur forte - valeur faible" qui apportent les contributions les plus importantes à la nuée du variogramme, pour la covariance non centrée ce sont plutôt les paires "valeur forte - valeur forte" qui influencent la structure observée. Autrement dit, la structure spatiale observable sur la covariance provient essentiellement des valeurs fortes, alors que c'est le contraste entre valeurs fortes et faibles qui gouverne la structure du variogramme.

Afin de caractériser les différences entre les deux outils variographiques, un premier indicateur est la dispersion des nuées régionales du variogramme et de la covariance non centrée. Ceci est réalisé dans le cas des modèles à effets de bord, les nuées étant dégénérées dans le cas des modèles d'ensembles aléatoires, ne prenant que les valeurs 0 et 0.5 ou 1.

2.2.5.1 Comparaisons des nuées

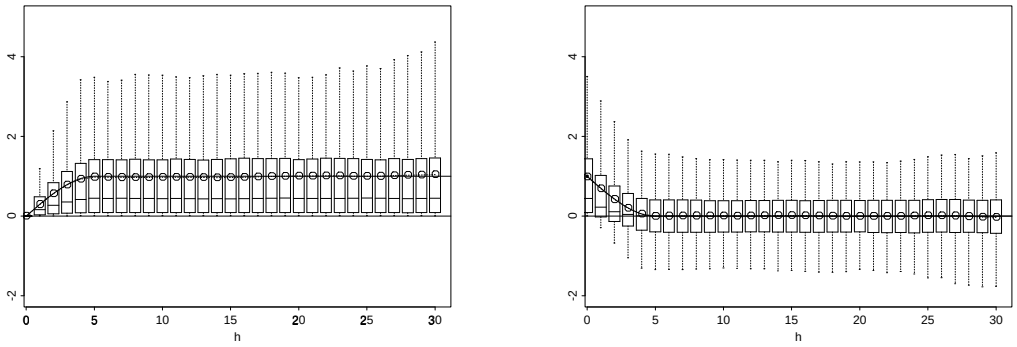
Bien que nous nous intéressions à la dispersion de la nuée pour chaque réalisation des modèles de FA, il est plus synthétique de présenter les résultats sur 100 simulations, ce qui revient à considérer un domaine très large.

Pour une FA gaussienne de modèle sphérique avec $a = L/6$ (voir figure 2.3), L étant la taille du domaine, les boxplots montrent une dispersion beaucoup plus élevée des nuées de $C(h)$ aux petites distances. Pour une régionalisation régulière, si h est petit, $z(x)$ est proche en valeur de $z(x+h)$; il en découle d'une part de faibles valeurs de $\frac{1}{2}[z(x) - z(x+h)]^2$ et d'autre part des valeurs de $z(x)z(x+h)$ qui sont faibles ou fortes en fonction de la valeur de $z(x)$.

Les nuées se stabilisent au-delà de la portée. Pour ces distances, $Z(x)$ et $Z(x+h)$ sont deux variables aléatoires gaussiennes indépendantes, et les variances de $\frac{1}{2}[Z(x)-Z(x+h)]^2$ et $Z(x)Z(x+h)$ sont alors respectivement 2 et 1. Cela explique la plus grande dispersion de la nuée du variogramme par rapport à celle de la covariance. Aux distances les plus grandes, la dispersion augmente à cause du faible nombre de paires.

Les boxplots illustrent la dissymétrie de la nuée du variogramme. Pour la covariance non centrée, cette dissymétrie existe aux petites distances, mais la nuée devient symétrique au-delà de la portée, passant ainsi progressivement d'une distribution chi-deux à l'origine à une distribution symétrique à partir de la portée.

Le cas d'une FA lognormale pour une structure sphérique de portée $a = L/6$ pour la gaussienne (voir figure 2.4) est à comparer avec les résultats précédents ; la variance a changé. Outre la dispersion sensiblement accrue des nuées, le variogramme moyen sur 100 simulations ne reproduit plus le palier du modèle, dont l'expression est donnée par l'équation 2.6. La covariance non centrée moyenne a des fluctuations moindres que $\gamma_R(h)$.



$\gamma_R(h)$ $C_R(h)$

FIG. 2.3 – Statistiques des nuées de $\gamma(h)$ et $C(h)$ pour 100 simulations. FA **gaussienne**, modèle sphérique avec $a = L/6$. Les boxplots reprennent les quantiles à 10, 25, 50, 75 et 90 % de la superposition des nuées de 100 réalisations, les cercles représentent les moyennes. L'échelle verticale est identique pour les deux graphiques.

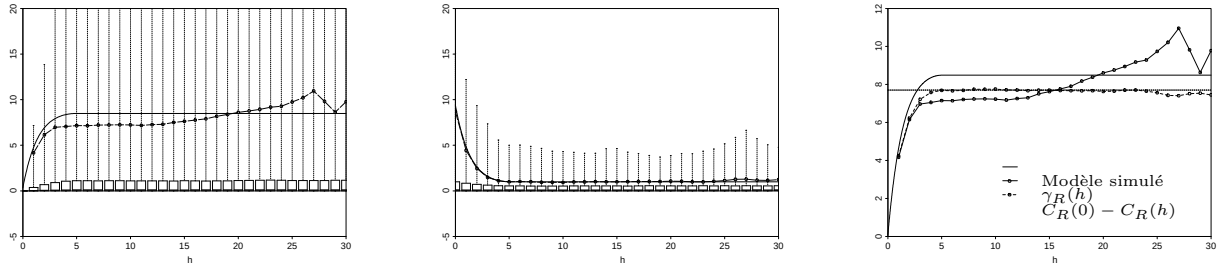
Dans le cas d'une FA lognormale avec un modèle sphérique cette fois de portée $a = 2L$ (voir figure 2.5), la covariance non centrée moyenne tend à indiquer la stationnarité. Une fois repassée en variogramme par $C_R(0) - C_R(h)$, son comportement aux petites distances devient semblable à celui du variogramme classique.

2.2.5.2 Intervalle interquartile

Il est utile de trouver un outil caractérisant la dispersion des nuées. Pour une simulation donnée, l'Intervalle InterQuartile (IIQ) représente pour chaque pas h la différence $q_{75} - q_{25}$ entre le premier et le troisième quartile du boxplot de la nuée pour ce pas. On caractérise dès lors la dispersion de la nuée du variogramme par la moyenne de IIQ sur les 100 simulations[3], autrement dit $\overline{IIQ_{\gamma(h)}}$.

Pour comparer les deux outils, nous considérons le rapport $I(h) = \frac{\overline{IIQ_{\gamma(h)}}}{\overline{IIQ_{C(h)}}}$. Pour chaque structure simulée, un rapport $I(h)$ supérieur (resp. inférieur) à 1 indique que la nuée du variogramme est en moyenne plus (resp. moins) dispersée que celle de la covariance. $I(h)$ est représenté à la figure 2.6 pour les FA gaussienne et lognormale et des portées du modèle sphérique simulé valant

[3]Un autre choix, consistant à considérer l'IIQ pour la superposition des nuées des 100 simulations, conduit à des résultats analogues.



$\gamma_R(h)$ $C_R(h)$

FIG. 2.4 – Statistiques des nuées de $\gamma(h)$ et $C(h)$ pour 100 simulations. FA **lognormale**, modèle sphérique avec $a = L/6$ pour la FA gaussienne. Les boxplots reprennent les quantiles à 10, 25, 50, 75 et 90 %, les cercles représentent les moyennes. L'échelle verticale est identique pour les deux premiers graphiques. Le troisième graphique contient les structures $\gamma_R(h)$ et $C_R(0) - C_R(h)$ simulées, le modèle ainsi que la variance des données.

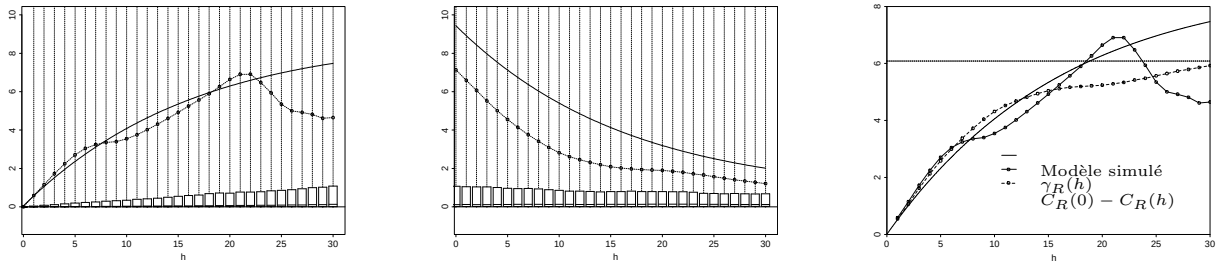
$L/6, L/2, L$ et $2L$.

Pour les deux FA, la nuée du variogramme est moins dispersée que celle de la covariance aux petites distances, quelle que soit la portée considérée. Ensuite, à partir d'une distance inférieure à $a/2$, c'est la nuée de la covariance qui devient moins dispersée. Cette distance est plus courte dans le cas de la FA lognormale. Pour la FA gaussienne, $I(h)$ se stabilise et atteint une valeur identique quelle que soit la portée ; cette stabilisation n'apparaît que pour la portée $a = L/6$, la plus courte, dans le cas lognormal.

En conclusion, pour les deux FA et les différentes portées du modèle simulé, la nuée du variogramme est moins dispersée aux petites distances que celle de la covariance ; ensuite, à partir d'une distance située empiriquement entre le quart et la moitié de la portée du modèle simulé, cette tendance s'inverse. La dispersion importante de la nuée de la covariance $C(h)$ aux petites distances illustre le risque qui existe à utiliser directement cet outil pour l'inférence de la structure spatiale. De plus, dans le cas de distributions dissymétriques pour lesquelles la stationnarité d'ordre deux n'est plus admissible, $C(0) - C(h)$ peut recréer une stationnarité artificielle, ce qui n'est pas le cas du variogramme classique [Rivoirard & Bez (1997)].

2.2.6 Variances d'estimation de $\gamma^\star(h)$ et $C^\star(0) - C^\star(h)$

La structure régionale étant calculée sur la grille 30×30, considérons la structure expérimentale obtenue à partir d'une grille de 11×11 points répartis sur le domaine de coté L (schéma fermé, voir figure 2.7) ; ces conditions d'échantillonnage sont relativement optimistes.



$\gamma_R(h)$ $C_R(h)$

FIG. 2.5 – Statistiques des nuées de $\gamma(h)$ et $C(h)$ pour 100 simulations. FA **lognormale**, modèle sphérique avec $a = 2L$ pour la FA gaussienne. Les boxplots reprennent les quantiles à 10, 25, 50, 75 et 90 %, les cercles représentent les moyennes. L'échelle verticale est identique pour les deux premiers graphiques. Le troisième graphique contient les structures $\gamma_R(h)$ et $C_R(0) - C_R(h)$ simulées, le modèle ainsi que la variance des données.

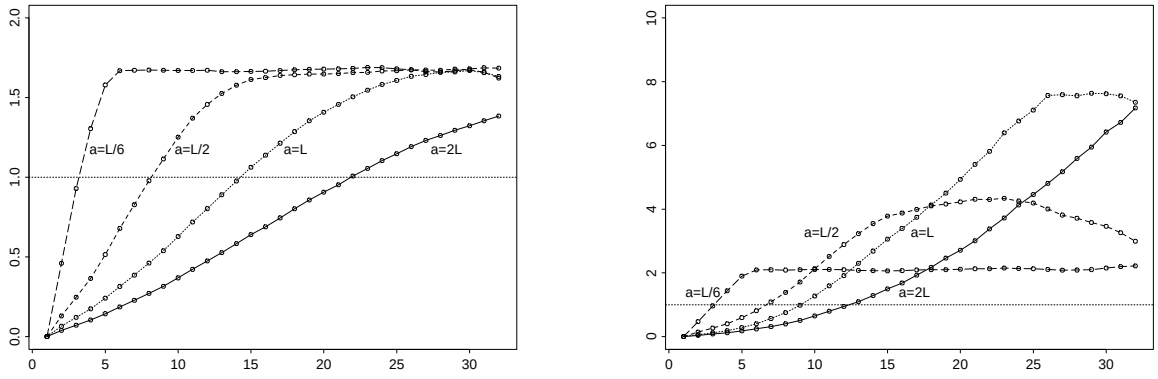
A titre d'exemple, les figures 2.8 et 2.9 illustrent sur *une simulation*, pour le variogramme classique et pour $C(0) - C(h)$, les différences existant entre structures expérimentale et régionale, dans le cas favorable d'une FA gaussienne puis d'une FA lognormale. Pour le cas gaussien et la portée courte $a = 5$, la structure régionale est bien approchée par la structure expérimentale pour le palier, tandis que la portée expérimentale semble légèrement plus élevée. Lorsque la portée est égale au coté du champ, la correspondance entre les structures expérimentale et régionale est très bonne à petite distance, et se détériore aux distances les plus grandes. Les résultats sont très semblables pour le variogramme classique et pour $C(0) - C(h)$.

Concernant la FA lognormale, la correspondance entre les structures est médiocre pour $a = 5$, que ce soit pour la portée, le palier ou le comportement aux petites distances. Les choses se passent légèrement mieux dans le cas $a = 30$, cas où il n'y a pourtant pas stationnarité à l'échelle du champ. Rappelons que ces résultats sont obtenus sur une seule simulation.

Considérons à présent les versions probabilistes $\Gamma_R(h)$ and $\Gamma^\star(h)$ du variogramme régional et du variogramme expérimental d'une FA $Z(x)$. La variance d'estimation du variogramme

$$\sigma^2_{\mathcal{E}}(h) = \mathrm{Var}[\Gamma^\star(h) - \Gamma_R(h)] \tag{2.8}$$

dépend de l'échantillonnage et converge dans le cas ergodique vers 0 lorsque la densité des données augmente. Cette variance d'estimation permet d'évaluer pour une FA donnée si le variogramme régional peut raisonnablement être approché par le variogramme expérimental, et donc si l'inférence structurale peut s'effectuer dans de bonnes conditions. Dans le cas gaussien, son calcul est simple pour le variogramme et la covariance non centrée [Alfaro Sironvalle (1979)] ; les calculs restent simples pour la covariance non centrée dans le cas lognormal. Cependant, nous intéressant également



(a) FA gaussienne (b) FA lognormale

FIG. 2.6 – Rapport $I(h)$ pour 100 simulations. FA gaussienne (a) et lognormale (b), modèle sphérique de portées $a = 2L$, $a = L$, $a = L/2$ et $a = L/6$. $I(h) > 1$ (resp. < 1) indique que la nuée du variogramme est en moyenne plus (resp. moins) dispersée que celle de la covariance.

au variogramme pour une FA lognormale, pour lequel les calculs se compliquent, nous procéderons systématiquement par simulation, par souci d'homogénéité. En effet, comme $\sigma^2_{\mathcal{E}}(h) = \mathrm{E}[(\Gamma^\star(h) - \Gamma_R(h))^2]$, on approche $\sigma^2_{\mathcal{E}}(h)$ par l'erreur quadratique moyenne sur un nombre suffisant de simulations. Cette approximation a été validée pour différents schémas simulés dans le cas gaussien à une et deux dimensions.

Les résultats sont présentés sous forme d'écart-types d'estimation relatifs du variogramme $\frac{\sigma_{\mathcal{E}}(h)}{\gamma(h)}$. Les calculs étant effectués dans une seule direction, parallèlement à un coté de la grille avec une tolérance angulaire nulle, les variances d'estimation obtenues sont plus élevées que celles obtenues avec une tolérance angulaire égale à 90°, *i.e.* un calcul faisant intervenir toutes les directions.

2.2.6.1 FA gaussienne

Les résultats pour le variogramme classique (voir figure 2.10 (a)) sont similaires à ceux obtenus par Chilès & Delfiner (1999) qui utilisent l'approximation [Matheron (1965)]

$$\mathrm{Var}[\Gamma^\star(h) - \Gamma_R(h)] = 4\gamma(h)\sigma_h^2$$

où σ_h^2 est la variance d'estimation de la moyenne de $Z(x)$ sur $D \cap D_{-h}$, calculée à partir des $N(h)$ paires intervenant à la distance h.

Les variances d'estimation sont suffisamment faibles pour permettre une inférence dans de bonnes conditions. Plus la portée du modèle simulé est grande et plus la variance d'estimation est faible et donc les conditions d'inférence bonnes.

La figure 2.10 (b) illustre les variances d'estimation pour $\gamma(h)$ et $C(0) - C(h)$. Le pas de calcul, indiqué pour le cas $a = L/6$, croît monotonement avec les variances d'estimation pour les autres portées. L'ordre de grandeur des variances d'estimation est identique pour les deux outils. Aux petites distances, ces variances sont légèrement plus élevées pour $C(0) - C(h)$, excepté pour $a = L/6$; ensuite, approximativement à partir de la moitié du champ, les conditions d'inférence avec $C(0) - C(h)$ deviennent également meilleures pour les trois autres portées considérées, et ce malgré l'absence de sens de la covariance régionale dans le cas non stationnaire.

2.2.6.2 FA lognormale

Les valeurs élevées des variances d'estimation relatives du variogramme (voir figure 2.11 (a)) illustrent la relation très éloignée entre variogramme expérimental et variogramme régional dans le cas d'une fonction aléatoire lognormale ; l'inférence de la structure spatiale est délicate dans ces conditions. Le nombre de simulations choisi est égal à 5000, au lieu de 500 dans le cas gaussien.

Tout comme dans le cas gaussien, plus la portée augmente et plus les variances d'estimation relatives diminuent. L'ordre de grandeur reste également analogue pour les deux outils, même si leurs différences sont nettement plus marquées ici (voir figure 2.11 (b)). Les conditions d'inférence ne restent meilleures pour le variogramme classique que pour les portées les plus grandes $a = L$ et $2L$, et seulement aux petites distances. Au contraire, pour $a = L/2$ et plus particulièrement $L/6$, les variances d'estimation relatives sont inférieures pour $C(0) - C(h)$ quelle que soit la distance, excepté approximativement à la moitié du champ dans le cas $a = L/2$.

2.2.6.3 Cas des ensembles aléatoires

Considérons comme ensembles aléatoires une FA gaussienne seuillée à la médiane, pour un schéma sphérique de portée $L/6 = 5$, et pour les polygones poissonniens un nombre moyen de droites égal à 40 ; la portée pratique du variogramme exponentiel vaut dans ce cas 4.5.

Comme l'illustre la figure 2.12, les modèles théoriques associés aux deux ensembles aléatoires sont alors très proches, ce qui permet leur comparaison. Deux exemples de simulations sont illustrés à la figure 2.13.

Pour ces deux familles d'ensembles aléatoires, la variance d'estimation du variogramme classique (voir figure 2.14) est légèrement inférieure à celle de la covariance non centrée, l'écart entre les deux variances augmentant à partir de $h \simeq 0.6L$.

Ces résultats ne sont pas surprenants, le variogramme étant plus sensible aux contrastes entre valeurs fortes et faibles, qui caractérisent particulièrement ces ensembles aléatoires.

2.2.7 Conclusions

Le variogramme classique reste dans tous les cas indispensable pour vérifier si l'hypothèse de stationnarité d'ordre deux est admissible. Si la portée observée est grande par rapport au domaine, la covariance non centrée est à éviter. Dans le cas contraire, l'apport de cette covariance non centrée dépend du modèle de fonction aléatoire.

Pour une FA gaussienne, le variogramme déduit de la covariance non centrée confère à l'inférence une meilleure précision aux grandes distances. En effet, les variogrammes expérimentaux présentent fréquemment, même dans le cas stationnaire, des fluctuations importantes dès que la distance devient grande. Le calcul parallèle de $C(0) - C(h)$ est alors utile pour l'estimation du palier.

Pour une FA lognormale, le recours à $C(0) - C(h)$ est encore plus pertinent, dès les petites distances. Cependant, même si cet outil permet une meilleure inférence, les variances d'estimation très élevées ne doivent pas faire oublier que l'on ne peut espérer une connaissance satisfaisante de la structure de la variable.

Le variogramme déduit de la covariance non centrée ne présente pas d'intérêt pour les ensembles aléatoires.

Le variogramme classique reste donc nécessaire pour valider l'hypothèse de stationnarité. Cela n'est pas sans risque, car ce faisant nous nous basons sur l'outil variographique le moins robuste pour valider cette hypothèse. Les variogrammes pondérés constituent de ce point de vue une alternative intéressante.

2.3 Variogrammes pondérés

Les estimateurs robustes du variogramme présentés jusqu'à présent reposent sur une hypothèse de stationnarité d'ordre deux, plus forte que la seule hypothèse intrinsèque nécessaire au variogramme classique. Il n'en est pas de même pour les variogrammes pondérés, qui ne requièrent que cette hypothèse intrinsèque. Développés par Rivoirard (1998, 2000), ils permettent d'accroître la robustesse du calcul de variogramme en prenant en compte la présence d'irrégularités d'échantillonnage et/ou d'éventuels défauts de robustesse des moments d'ordre 2, en particulier lorsque les données présentent de forts contrastes. D'usage encore peu courant, ils sont présentés ci-dessous.

Considérant les pondérateurs W_{ij} associés aux paires (x_i, x_j), nous avons l'expression générale

d'un variogramme pondéré

$$\gamma_p(h) = \frac{1}{2} \frac{\sum\limits_{x_i - x_j \sim h} W_{ij}(z_i - z_j)^2}{\sum\limits_{x_i - x_j \sim h} W_{ij}} \tag{2.9}$$

avec $\sum\limits_{x_i - x_j \sim h} W_{ij}$ le poids du variogramme à la distance h, non nul. On suppose ces poids uniquement fonction de la configuration géométrique des données, et non des valeurs de ces données, ce qui confère au variogramme pondéré la même propriété de non biais que le variogramme classique.

2.3.1 Variogramme pondéré par échantillon

Lorsque celle-ci existe, il est conseillé de prendre en compte l'irrégularité de l'échantillonnage si l'on veut éviter un risque de biais lors de l'estimation. Même dans le cas d'une maille principale régulière, de telles irrégularités surviennent si l'on considère des resserrements locaux de maille, ou des contours de champ non réguliers.

Il est possible d'associer à chaque point x_i un poids w_i inversement proportionnel à une variable traduisant la densité de l'échantillonnage en x_i, comme par exemple la surface d'influence de x_i. On obtient alors le **variogramme pondéré par échantillon** (p.p.e.)

$$\gamma_{ppe}(h) = \frac{1}{2} \frac{\sum\limits_{x_i - x_j \sim h} w_i w_j [z(x_i) - z(x_j)]^2}{\sum\limits_{x_i - x_j \sim h} w_i w_j} \tag{2.10}$$

où la somme des $w_i w_j$ pour tous les couples possibles vaut 1.

Ce variogramme pondéré par échantillon nécessite la détermination des pondérateurs ; nous les voulons inversement proportionnels à la densité d'échantillonnage au point concerné, mais comment estimer cette densité ? Cette question représente à elle seule un vaste domaine [Gordon (1981), Silverman (1986)] et il existe de nombreuses techniques, la plus simple consistant à prendre pour densité en un point le nombre de points situés dans un certain voisinage de celui-ci. Cela nécessite par conséquent le choix de la taille de ce voisinage. Plus généralement, tout choix de pondérateurs inclut une part d'arbitraire.

2.3.2 Variogramme moyen par échantillon

L'équivalent expérimental de l'expression régionale 2.4 s'écrit

$$\gamma^\star(h) = \frac{1}{2N(h)} \sum_{x_i - x_j \sim h} (z_i - z_j)^2 = \underbrace{\frac{1}{2N(h)} \sum_{x_i - x_j \sim h} (z_i^2 + z_j^2)}_{= s^{2\star}(h)} - \underbrace{\frac{1}{N(h)} \sum_{x_i - x_j \sim h} z_i \, z_j}_{= C^\star(h)} \tag{2.11}$$

Dans le cas non stationnaire, pour une distance h donnée l'instabilité du variogramme expérimental provient surtout du premier terme $s^{2\star}(h)$, la variabilité spatiale étant mesurée uniquement à partir des données intervenant à cette distance h. Ces données impliquées variant d'une distance à l'autre, il en découle des fluctuations de la variance des échantillons en fonction de la distance ; cela est particulièrement le cas pour des variables fortement contrastées, à cause de l'instabilité des moments d'ordre deux.

Au contraire, on préférerait mesurer la variabilité spatiale de façon cohérente par rapport à l'ensemble des données. Pour cela, on peut considérer le *variogramme moyen par échantillon* (m.p.e.), construit comme suit :

– Calcul du variogramme "classique" associé à chaque x_i

$$\gamma_i(h) = \frac{1}{2N_i(h)} \sum_{x_i - x_j \sim h} [z(x_i) - z(x_j)]^2 \ ,$$

où $N_i(h)$ est le nombre de points x_j séparés de x_i par h.

– Calcul de la moyenne de ces "variogrammes par échantillon"

$$\gamma_{mpe}(h) = \frac{1}{n(h)} \sum_{i=1}^{N} \gamma_i(h) \ ,$$

où $n(h)$ est le nombre de points x_i intervenant à cette distance h.

Il est en outre possible d'inclure des pondérations différentes pour chaque échantillon, comme cela est le cas pour le variogramme pondéré par échantillon.

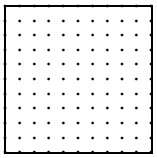
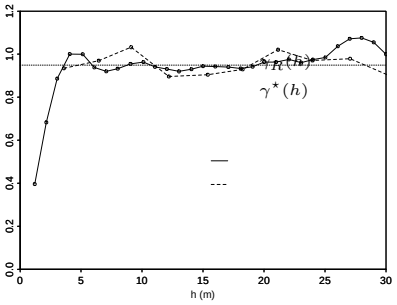
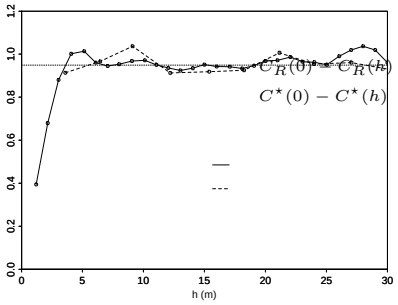


FIG. 2.7 – Grille d'échantillonnage expérimentale de 11×11 points dans un carré de coté L.

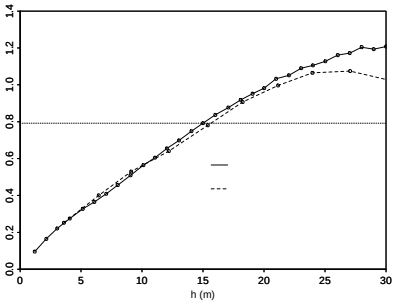


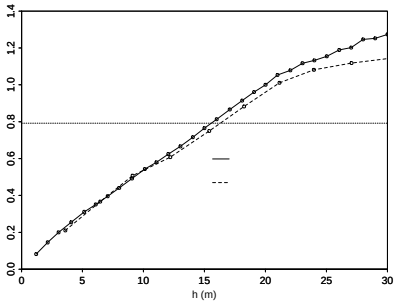


$\gamma_R(h)$
$\gamma^\star(h)$

(a) $a = L/6$

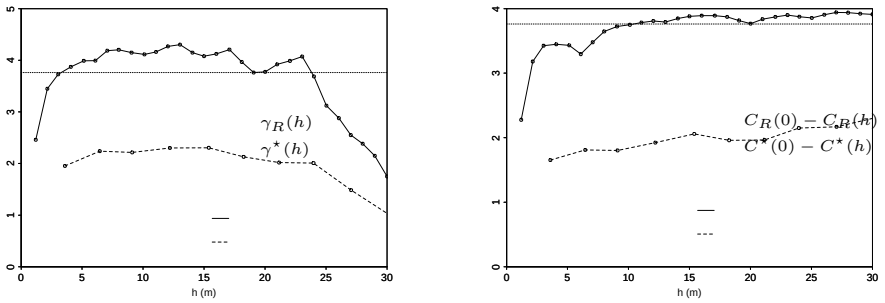
$C_R(0) - C_R(h)$
$C^\star(0) - C^\star(h)$



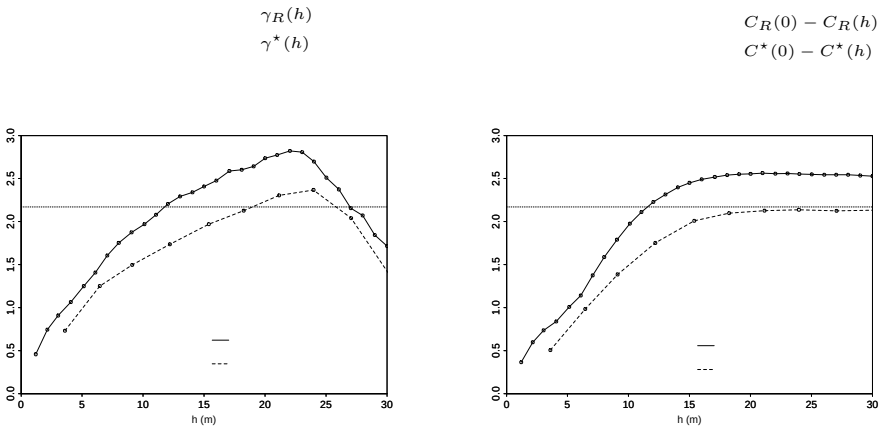


(b) $a = L$

FIG. 2.8 – Comparaison entre structures expérimentale et régionale pour le variogramme classique et $C(0) - C(h)$. Cas d'une FA **gaussienne** simulée avec un schéma sphérique de portée a égale à $L/6 = 5$ (a) et $L = 30$ (b).

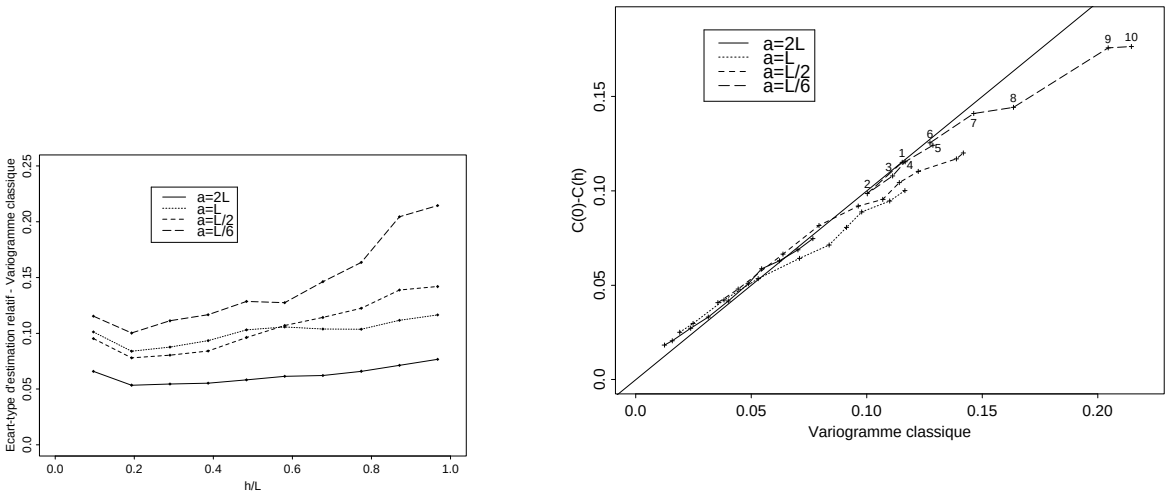


(a) $a = L/6$



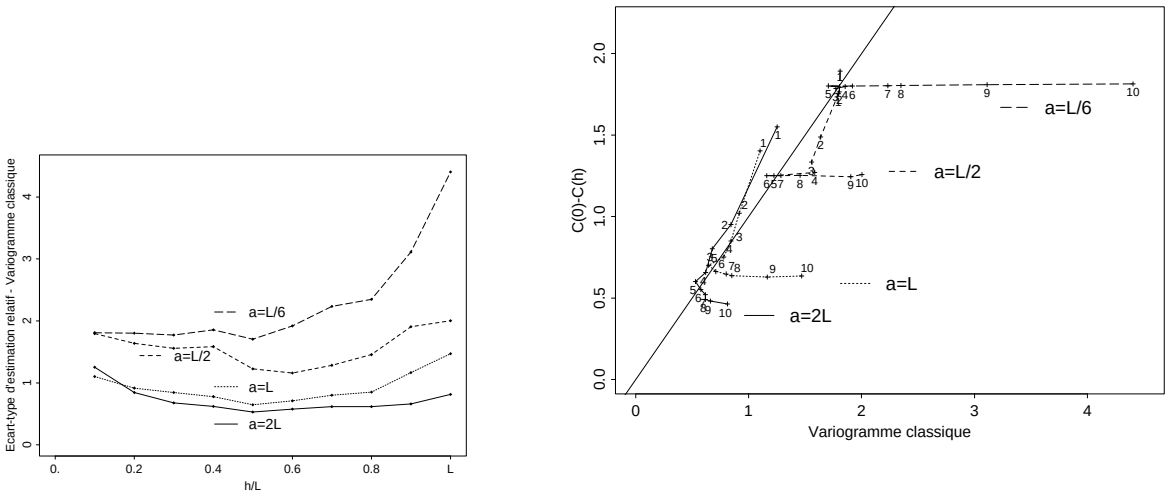
(b) $a = L$

FIG. 2.9 – Comparaison entre structures expérimentale et régionale pour le variogramme classique et $C(0) - C(h)$. Cas d'une FA **lognormale**, modèle sphérique simulé pour la FA gaussienne de portée a égale à $L/6 = 5$ (a) et $L = 30$ (b).



(a) (b)

FIG. 2.10 – Ecart-types d'estimation relatifs pour 500 simulations. Résultats pour le variogramme classique (a) et croisés entre variogramme classique et $C(0) - C(h)$ (b). Pas des calculs expérimentaux $\frac{L}{10}, \frac{2L}{10}, \ldots, L$ indiqués pour $a = L/6$. FA **gaussienne**, modèle sphérique avec $a = 2L$, $a = L$, $a = L/2$ et $a = L/6$.



(a) (b)

FIG. 2.11 – Ecart-types d'estimation relatifs pour 5000 simulations. Résultats pour le variogramme classique (a) et croisés entre variogramme classique et $C(0) - C(h)$ (b). Pas des calculs expérimentaux $\frac{L}{10}, \frac{2L}{10}, \ldots, L$ indiqués. FA **lognormale**, modèle sphérique avec $a = 2L$, $a = L$, $a = L/2$ et $a = L/6$ pour la FA gaussienne.

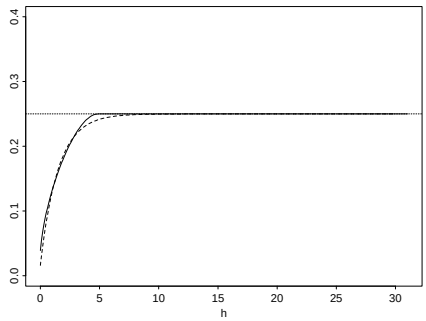
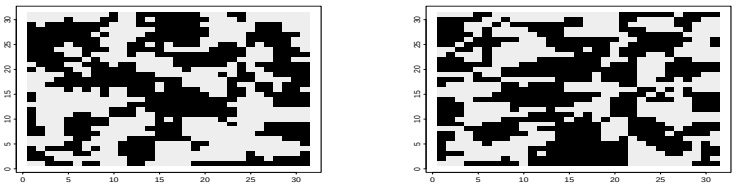


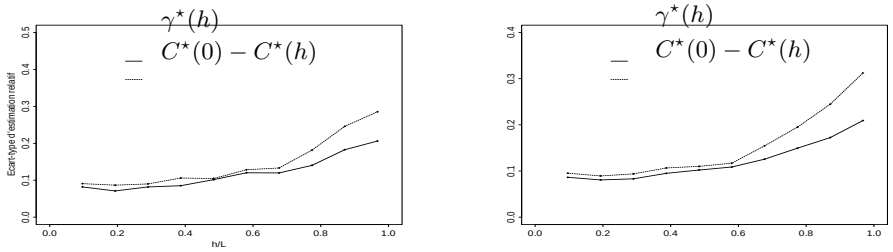
FIG. 2.12 – Comparaison des modèles théorique associés à : "——" la gaussienne seuillée (schéma sphérique de portée 5), et : "- - -" aux polygones poissonniens (40 droites en moyenne, portée pratique 4.5).



($a = L/6$) (40 droites)

(a) Gaussiennes seuillées (b) Polygones poissonniens

FIG. 2.13 – Deux réalisations d'ensembles aléatoires discrétisés sur une grille de 31×31 points.



(a) Polygones poissonniens (b) Gaussiennes seuillées

FIG. 2.14 – Ecart-types d'estimation relatifs pour 500 simulations, variogramme classique et $C(0) - C(h)$. Polygones poissonniens avec un nombre moyen de droites égal à 40 (portée pratique 4.5) (a) et seuillage d'une gaussienne simulée avec un modèle sphérique de portée $a = L/6$ (b).

Troisième partie

Site Y

Chapitre 3

Echantillonnage et analyse exploratoire

Sommaire

Ce chapitre présente le site Y, qui concerne la zone d'épandage d'une ancienne cokerie. La stratégie d'échantillonnage est décrite. L'analyse exploratoire détaillée des données permet de situer le niveau de pollution du site, d'illustrer les corrélations entre HAP, entre différentes techniques de prélèvements, entre HAP et données auxiliaires relevées et finalement entre HAP et informations historiques.

3.1 Site et stratégie d'échantillonnage

3.1.1 Description et historique

La période d'activité de cette ancienne cokerie, située dans la région de Douai (59), s'étend de 1921 à 1973. A partir de la pyrolyse d'environ 130 000 tonnes de charbon, la cokerie produisait annuellement approximativement 100 000 tonnes de coke, 4 000 tonnes de goudron, 1 100 tonnes de sulfate d'ammonium et 400 tonnes de benzol.

Les installations industrielles et annexes ont été démolies au cours des années 1992 et 1993. La figure 3.1 montre la configuration actuelle du site. Cinq campagnes d'échantillonnage des sols, réalisées entre 1990 et 1995, ont permis de déterminer approximativement la zone d'extension des matériaux goudronneux : anciennes mares à goudron Nord et Sud, ancienne argilière située à proximité de la mare à goudron Sud (voir figure 3.2). Suite à la première campagne, un confinement de terres souillées sur une surface d'environ 7000 m^2 a été mis en place au Nord-Est du site.

Dans la suite, la "mare à goudron sud" désigne la mare à goudron à proprement parler et

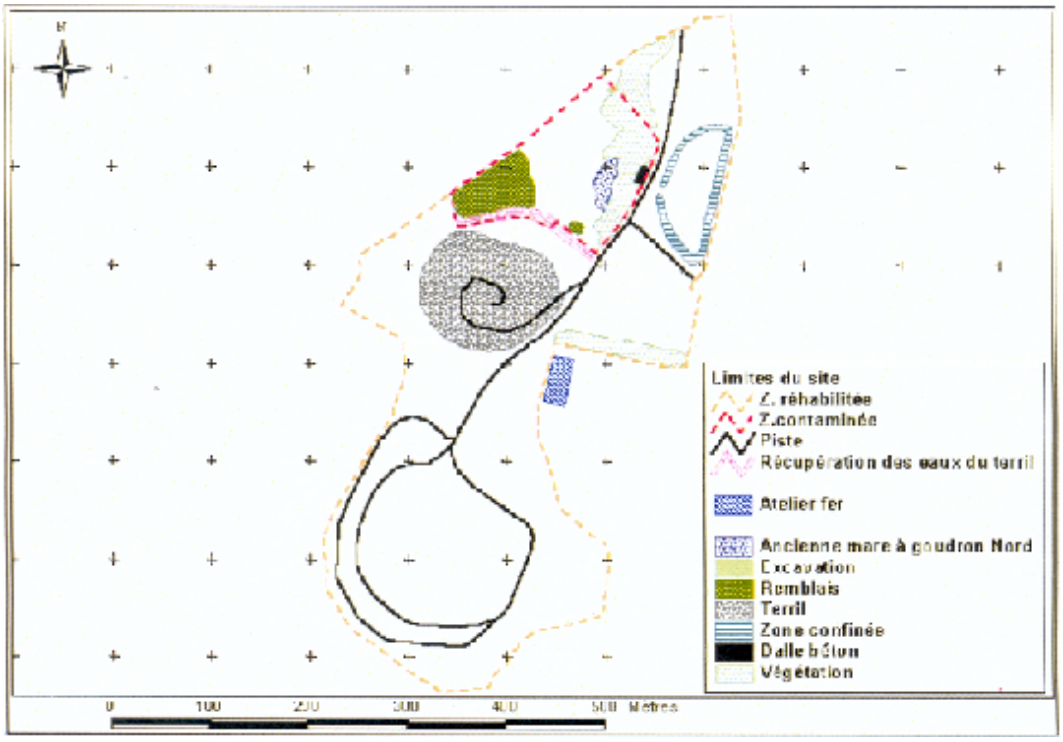


FIG. 3.1 – Site Y - Configuration actuelle [Pitout (2000)].

l'ancienne argilière adjacente. Les deux mares à goudron ont été excavées en 1992. La délimitation de la mare à goudrons Nord est obtenue par géoréférencement [Pitout (2000)] et reprise sur la carte d'implantation des données de la figure 3.2. L'extension de la mare à goudron sud, remblayée, a été déterminée sur la base des campagnes réalisées de 1990 à 1995. Des teneurs en 16 HAP comprises entre moins de 1 $mg.kg^{-1}$ et plus de 11000 $mg.kg^{-1}$ y ont été trouvées. Outre ces mares, aucune indication de remaniement du sol de la zone d'épandage n'a été fournie. L'unité des teneurs, couramment le $mg.kg^{-1}$ ou le $g.t^{-1}$, soit des "parties par million", est symbolisée par *ppm*.

Le profil géologique présente, de la surface vers la profondeur :
- des remblais provenant du démantèlement des installations de la cokerie (jusqu'à 3.7 m au niveau de la cokerie),
- des limons et argiles plus ou moins sableuses d'origine quaternaire et tertiaire (5 à 6 m),
- des craies blanches du Sénonien puis grises du Turonien, constituant l'aquifère principal (50 m environ). Outre une nappe perchée située au Nord de la mare Sud, le toit de la nappe principale se situe à une profondeur d'environ 15 m.

3.1.2 Campagne d'échantillonnage

La campagne CNRSSP a eu lieu en 1997, avec pour objectif l'investigation de la zone d'épandage de la cokerie, située au Nord du terril et à l'Ouest de la zone confinée. La maille principale d'échantillonnage est de 10 m sur la majeure partie de la zone investiguée ; deux croix de son-

dages espacés de 5 m ont été réalisées. La figure 3.2 reprend les 52 points échantillonnés.

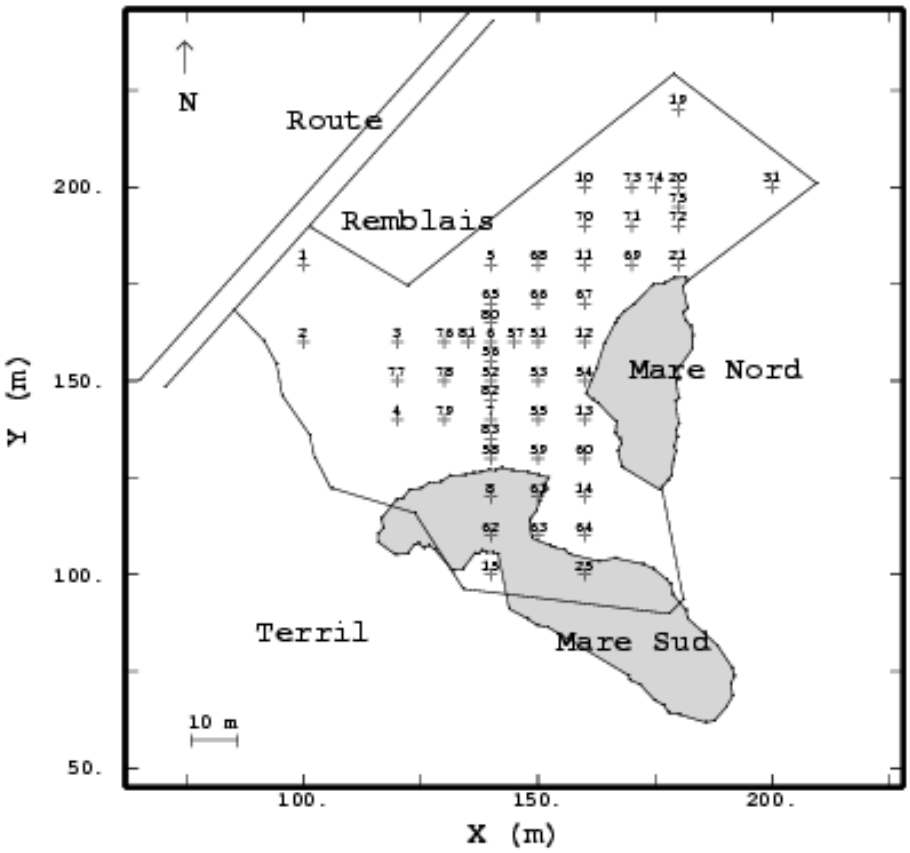


FIG. 3.2 – Site Y - Implantation des données, indication des principales informations historiques.

Outre la mise au point et l'utilisation de méthodes géostatistiques, l'objectif de la campagne consistait à comparer deux techniques d'échantillonnage, mises en œuvre en chaque point (voir figure 3.3) :

- Un **sondage** de 1.5 m de profondeur et 45 mm de diamètre - sondeuse à percussion fonctionnant sans adjonction d'eau ni d'air et utilisant une sonde à lumières[1] -, avec description puis prélèvement le long de la carotte.
- Une **fosse** d'environ 0.5 m de profondeur ouverte à la pelle mécanique - munie d'une benne preneuse de 60 cm de large -, avec description puis prélèvement par rainurage vertical.

Outre les coordonnées ainsi que la profondeur des fosses ou la hauteur de carotte récupérée pour les sondages, nous disposons des analyses[2] des 16 HAP de l'US EPA (en $mg.kg^{-1}$ sec) et d'un indice Phénol (en $mg.kg^{-1}$), ainsi que des masses des fractions granulométriques suivantes : inférieure à 2 mm, comprise entre 2 et 5 mm et supérieure à 5 mm. Conformément à la norme AFNOR en vigueur, seule la fraction granulométrique inférieure à 2 mm est analysée ; nous reviendrons au

[1] Plusieurs photographies illustrent dans l'annexe A.2.3 le matériel utilisé.
[2] Les protocoles analytiques sont décrits dans l'annexe B.2.

paragraphe 3.2.4 sur les implications de ce choix.

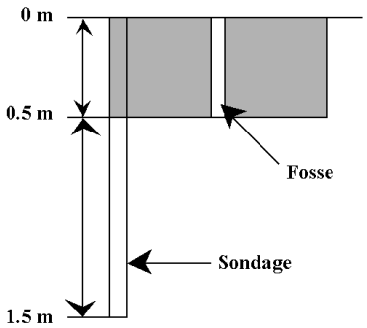


FIG. 3.3 – Stratégie d'échantillonnage des fosses et sondages.

La fosse du point 12 n'a pas été réalisée à cause de la présence de câbles électriques enterrés à cet endroit. D'emblée, notons qu'un tel échantillonnage ne permet pas d'effectuer une estimation du volume des zones contaminées, car l'extension verticale de la pollution n'est pas connue, l'investigation s'arrêtant à une profondeur définie a priori trop faible. Les supports des deux types de prélèvement sont différents. Cependant, le prélèvement d'un kilogramme de sol a été réalisé de façon identique dans les deux cas : par rainurage vertical sur les fosses, ainsi que sur les carottes provenant des sondages, en retenant chacun des matériaux rencontrés de façon à assurer une bonne représentativité de l'échantillon.

Initialement, une mesure de la teneur en polluants organiques volatils devait être réalisée en chaque point. Les mauvaises conditions climatiques et la trop grande taille des fosses, cause de fuites, ont rendu cette mesure impossible ; nous y reviendrons au chapitre 7.

3.2 Analyse exploratoire

L'étude détaillée de chaque HAP permet d'analyser leurs caractéristiques et de déterminer ceux qui sont les plus présents et les plus représentatifs, sur lesquels nous nous concentrerons ensuite.

3.2.1 Fosses et Sondages

3.2.1.1 Statistiques élémentaires

Les statistiques élémentaires ainsi que les quartiles et valeurs extrêmes sont donnés pour les deux modes de prélèvement aux tableaux 3.1 et 3.2.

Les concentrations moyennes en HAP sont systématiquement plus élevées pour les sondages que pour les fosses ; ce n'est pas le cas pour l'indice Phénol. Les écarts-types sont bien supérieurs sur les sondages, excepté pour Nap et Phl. Les coefficients de variation sont particulièrement élevés, avec des valeurs supérieures sur les fosses. La plupart des HAP présentent un nombre important

HAP	Données de fosses				Données de sondages			
	m	σ	σ/m	DIS	m	σ	σ/m	DIS
Nap	55.14	184.16	3.34	0	88.97	184.50	2.07	0
Acy	4.39	14.97	3.41	75	13.58	45.25	3.33	56
Ace	8.57	26.38	3.08	14	30.18	83.77	2.78	10
Fle	15.37	47.22	3.07	41	55.24	146.18	2.65	2
Ant	17.05	55.36	3.25	39	50.36	127.05	2.52	30
Phe	44.32	113.94	2.57	10	139.94	322.23	2.30	0
Flt	65.07	215.68	3.31	14	163.97	435.36	2.66	10
Pyr	41.73	138.68	3.32	20	115.96	345.93	2.98	8
Baa	34.33	116.91	3.41	25	63.31	169.05	2.67	20
Cry	28.37	98.35	3.47	24	61.26	185.09	3.02	20
Bap	20.36	67.02	3.29	25	37.83	96.25	2.54	26
Bbf	20.47	61.79	3.02	29	44.91	107.07	2.38	20
Dba	2.43	8.27	3.40	55	7.06	17.21	2.44	38
Bkf	11.73	39.84	3.40	33	25.32	60.54	2.39	30
Bgh	6.85	17.71	2.59	45	19.40	58.26	3.00	28
Inp	8.96	28.59	3.19	43	19.37	40.88	2.11	28
Phl	15.43	63.74	4.13	14	9.97	25.45	2.55	14

TAB. 3.1 – Site Y - Moyennes m, écart-types σ et coefficients de variation σ/m des concentrations analysées sur les fosses et les sondages ; pourcentages de données inférieures aux seuils de détection (DIS).

de valeurs inférieures au seuil de détection (DIS), égal à 0.1 ppm dans la plupart des cas. Ces concentrations ont été systématiquement ramenées à la valeur du seuil. Il a également été envisagé de les ramener à 0 ; ce choix n'a aucune répercussion sur les statistiques ni sur les structures spatiales, étant donné le niveau de variabilité des concentrations. Par ailleurs, la valeur de ce seuil de détection étant négligeable par rapport aux seuils de dépollution usuellement considérés, égaux à 5 ou 10 ppm par HAP, la mise en œuvre à ce niveau de techniques plus avancées consistant à simuler les concentrations inférieures au seuil de détection n'a pas été jugée utile. Les pourcentages de valeurs inférieures aux seuils de détection sont plus faibles sur les sondages.

Les quantiles à 25 % (voir tableau 3.2) sont tous plus élevés pour les sondages, de même que les quantiles à 75 % et les valeurs maximales - excepté pour Phl et Nap. Les médianes sont par contre plutôt plus élevées sur les fosses, notamment pour les HAP à 4 et 5 cycles. On en déduit une dissymétrie accrue des sondages, avec quelques concentrations nettement plus fortes.

En exprimant la concentration moyenne de chaque HAP en pourcentage de la concentration moyenne de la somme des 16 HAP, on constate la prédominance de Flt, Phe, Nap et Pyr (voir figure 3.4). Nap, le plus léger et seul HAP volatil, reste plutôt localisé dans les fosses, Pyr et Phe plutôt sur les sondages. Pour les deux modes de prélèvement, Ace et Acy sont très peu présents, tout comme Ant et les HAP les plus lourds - Dba, Bkf, Bgh et Inp.

HAP	Données de fosses					Données de sondages				
	Min	25 %	50 %	75 %	Max	Min	25 %	50 %	75 %	Max
Nap	0.17	1.95	5.80	11.80	1230.00	0.86	2.20	8.30	90.20	980.00
Acy	0.10	0.10	0.10	0.10	94.00	0.10	0.10	0.10	1.80	300.00
Ace	0.10	0.13	0.34	1.40	130.00	0.10	0.63	1.90	11.00	540.00
Fle	0.10	0.10	0.30	1.80	250.00	0.10	0.80	1.50	14.00	740.00
Ant	0.10	0.10	0.75	3.30	370.00	0.10	0.10	0.51	6.70	650.00
Phe	0.10	0.45	2.60	12.00	533.00	0.26	1.50	4.00	20.00	1500.00
Flt	0.10	0.34	3.70	18.00	1500.00	0.10	0.49	2.40	50.00	2400.00
Pyr	0.10	0.20	2.30	11.00	960.00	0.10	0.30	1.40	51.00	2200.00
Baa	0.10	0.10	2.10	12.00	820.00	0.10	0.10	0.90	27.00	920.00
Cry	0.10	0.10	2.40	9.30	690.00	0.10	0.20	1.00	26.00	1200.00
Bap	0.10	0.10	1.20	10.00	470.00	0.06	0.10	0.92	18.00	550.00
Bbf	0.10	0.10	1.30	11.00	420.00	0.05	0.13	0.89	22.00	520.00
Dba	0.10	0.10	0.10	1.10	58.00	0.10	0.10	0.20	3.10	95.00
Bkf	0.10	0.10	0.57	4.90	280.00	0.10	0.10	0.44	12.00	300.00
Bgh	0.10	0.10	0.20	3.10	110.00	0.10	0.10	0.42	15.00	400.00
Inp	0.10	0.10	0.42	4.00	200.00	0.10	0.10	0.61	9.00	200.00
Phl	0.10	0.13	0.34	1.50	430.00	0.10	0.19	0.32	1.20	130.00

TAB. 3.2 – Site Y - Extrêma et quartiles des concentrations analysées sur les fosses et les sondages.

3.2.1.2 Histogrammes et implantation

La dissymétrie des histogrammes (voir figures 3.5 et 3.6 respectivement pour les fosses et les sondages) est systématique et très marquée. Une zone de fortes concentrations est visible au Sud pour les fosses comme pour les sondages ; elle correspond à l'ancienne mare à goudrons Sud, pourtant remblayée par des matériaux propres. Par ailleurs, on observe particulièrement pour les sondages une autre zone de fortes teneurs, correspondant au point 5 situé au centre Nord de la zone, à proximité du tas de remblais provenant de l'excavation de la mare à goudron Est, ainsi qu'au Sud-Est de cette zone.

3.2.1.3 Corrélations

Bien qu'indispensable, la description des caractéristiques par HAP est insuffisante, car elle laisse de coté les corrélations entre ces différents éléments, dont l'existence a déjà été montrée [voir par exemple Oosterbaan-Eritzpokhoff (2000)].

Considérons le cas du benzo(a)anthracène Baa et du benzo(a)pyrène Bap, respectivement à 4 et 5 cycles. Le coefficient de corrélation entre ces deux HAP, pour les fosses, est égal à 0.996, valeur qui laisse penser à une corrélation excellente. Ce coefficient de corrélation provient cependant en grande partie de la concentration la plus forte - fosse 8 - (voir figure 3.7(a)). Les nuages indiquent par ailleurs la linéarité de la corrélation entre les concentrations.

Pour affiner l'analyse des corrélations, plusieurs choix sont possibles :
– Une fois observée la bonne correspondance entre les valeurs les plus fortes, un calcul du coefficient de corrélation sans ces valeurs conduit à un coefficient de 0.97, qui reste très

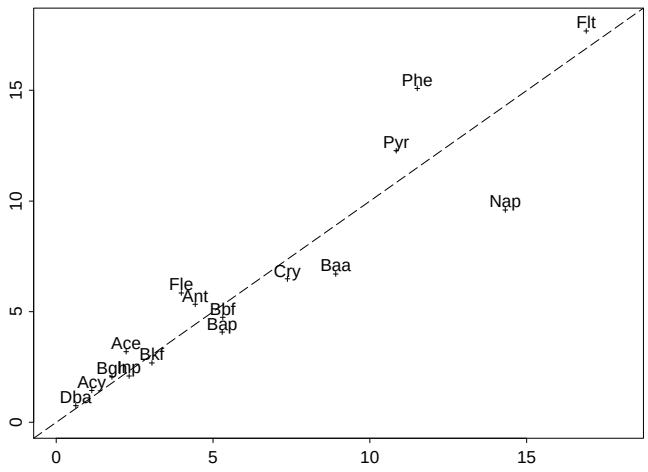


FIG. 3.4 – Site Y - Pourcentages des HAP entre fosses et sondages, par rapport à la teneur totale en HAP.

bon, ce qu'illustre la figure 3.7(b). Cela présuppose l'examen nuage par nuage des valeurs potentiellement très fortes, fastidieux dans le cas d'un nombre de variables élevé.

– Une seconde possibilité consiste à observer la corrélation entre les variables log-translatées de chaque HAP, cette transformation ayant l'intérêt de réduire l'influence des valeurs les plus fortes (voir page 23). Le coefficient de corrélation passe alors à 0.98 et la très bonne corrélation est confirmée par la figure 3.7(c).

Le tableau 3.3 reprend les coefficients de corrélation entre les logarithmes translatés des concentrations en HAP pour les fosses ; les résultats sont analogues pour les sondages. La première remarque concerne le très bon niveau de corrélation entre les différents HAP. A cause de sa détection délicate, Dba se distingue des autres HAP par des coefficients de corrélation moins bons.

Ces corrélations sont d'autant meilleures que les poids atomiques des HAP concernés sont proches, ce qu'illustrent également les nuages de corrélation de la figure 3.8 entre Inp, à 6 cycles, et d'autres HAP de plus en plus légers. On observe la dégradation de la corrélation lorsque le poids

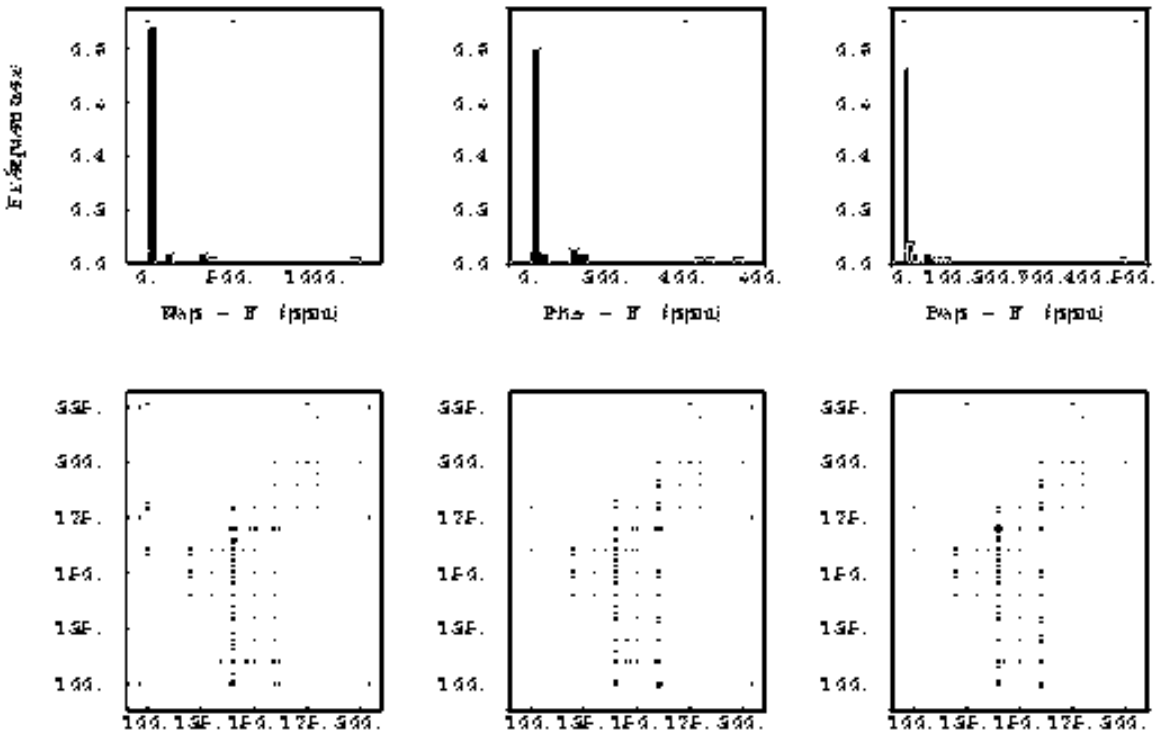


FIG. 3.5 – Site Y - Histogrammes des données de fosses. Les échelles de représentation des croix diffèrent pour chaque figure.

du second HAP devient de plus en plus léger, différant ainsi de plus en plus de celui de Inp ; la correspondance entre les valeurs fortes devient également moins bonne.

3.2.1.4 Utilisation de la somme des 16 HAP

Le risque de l'utilisation de la somme des 16 HAP de l'US EPA comme indicateur de la pollution en HAP a déjà été dénoncé [voir par exemple Lecomte et al. (1998)]. Les HAP les plus présents sur les sites de cokeries et en l'occurrence sur ce site Y sont Flt et Pyr, à 4 cycles, ainsi que Phe à 3 cycles et Nap à 2 cycles ; ces HAP influencent donc particulièrement la somme des 16 HAP (Σ16) au détriment des HAP moins présents, notamment les HAP à 5 et 6 cycles, qui sont les plus toxiques ! Par conséquent, une concentration en Σ16 égale à 500 ppm peut représenter des risques sanitaires très différents selon sa composition. Par exemple, pour les sondages, bien que les valeurs fortes soient globalement respectées, la corrélation est moins bonne entre Inp, HAP à 6 cycles, et Σ16 qu'entre Flt, HAP le plus présent, et Σ16 (voir figure 3.9). Des conclusions tirées de l'utilisation de cette variable Σ16 par exemple lors de l'estimation seraient à prendre avec précaution.

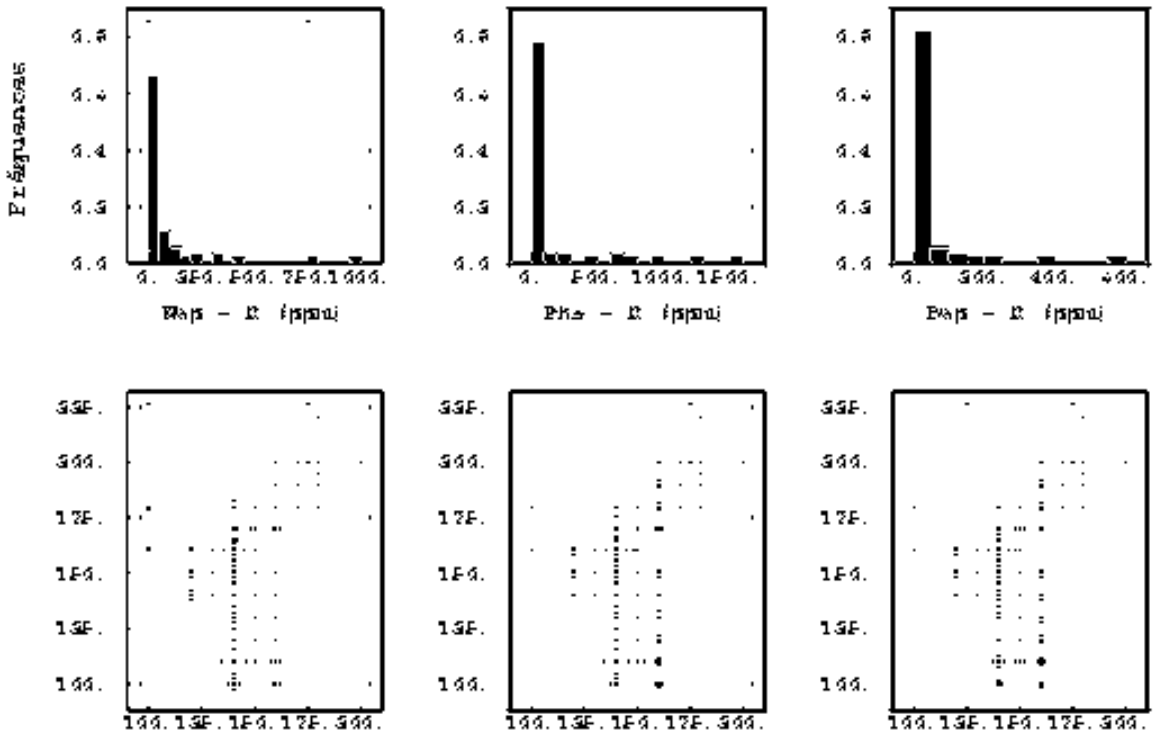
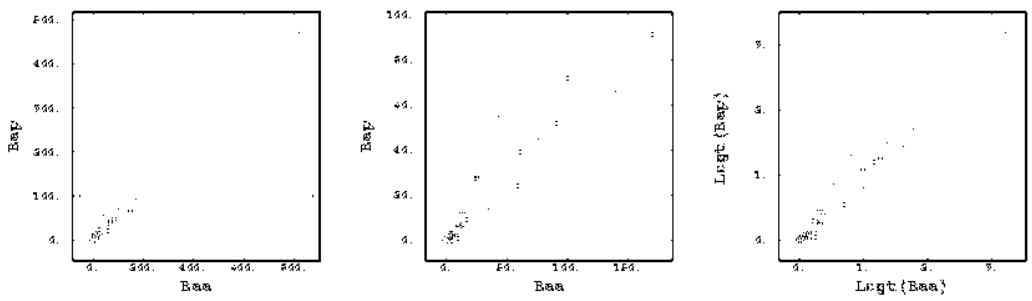


FIG. 3.6 – Site Y - Histogrammes des données de sondages. Les échelles de représentation des croix diffèrent pour chaque figure.

3.2.2 Corrélations entre fosses et sondages

Les deux prélèvements effectués en chaque point proviennent de profondeurs différentes : 0-0.5 m pour les fosses, 0.5-1.5 m pour les sondages. En comparant les concentrations point par point entre les fosses et sur les sondages (voir tableau 3.4), on constate que les corrélations entre les deux profondeurs sont bonnes pour tous les HAP, en variable brute comme en logarithme translaté.

Les nuages de corrélation entre fosses et sondages de la figure 3.10, situés dans la majorité des cas au-dessus de la première bissectrice, montrent qu'à de nombreuses valeurs fortes sur les sondages sont associées des valeurs faibles sur les fosses. La pollution semble donc localisée plutôt en profondeur. Il est possible qu'une couche de matériau sain ait été déposée *a posteriori* sur le site. Cependant, plutôt qu'un effet de la profondeur, les différences entre les techniques de prélèvement sur les fosses et les sondages peuvent également expliquer l'allure des nuages. Néanmoins, l'effet de support implique des concentrations plus dispersées lorsque le support diminue, mais autour d'une moyenne qui reste elle constante, ce qui n'est pas le cas ici. Le prélèvement d'échantillon étant réalisé de façon similaire dans les deux cas, il est peu probable qu'une fraction ait été mieux prélevée dans un cas - ce qui pourrait expliquer les concentrations moyennes différentes. Nous pouvons donc raisonnablement conclure à un réel rôle de la profondeur sur les concentrations, et non de la technique d'échantillonnage.



(a) (b) (c)

FIG. 3.7 – Site Y - Nuages de corrélation entre Baa et Bap sur les fosses : (a) variables brutes, (b) variables brutes avec retrait de la fosse 8, (c) variables log-translatées. Les échelles sont en ppm pour les variables brutes.

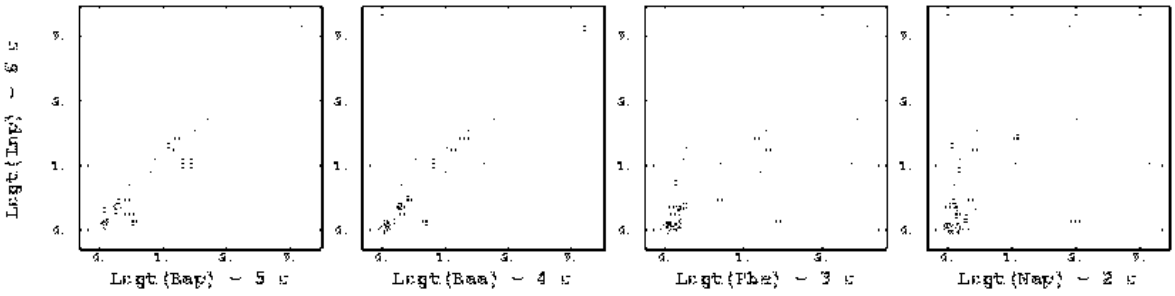


FIG. 3.8 – Site Y - Nuages de corrélation entre Inp et 4 autres HAP, sur les fosses.

3.2.3 Analyse en composantes principales des concentrations

Une analyse en composantes principales[3] est menée sur les concentrations des 16 HAP centrées et réduites. Pour les fosses, la projection des variables sur les premiers facteurs est illustrée à la figure 3.11.

La première composante, qui explique 86.6 % de la variance des données, correspond à un *facteur de taille* ; il traduit la corrélation entre les différents HAP. Par opposition, la seconde composante est souvent appelée *facteur de forme* ; elle explique 11.5 % de la variance, et possède une interprétation physique claire. En effet, les HAP légers à 2 et 3 cycles ont un poids positif sur cette seconde composante, tandis que les HAP de 4 cycles et plus ont un poids négatif, exception faite de Bgh dont le poids est légèrement positif. Cette seconde composante distingue donc HAP légers et lourds.

L'interprétation des composantes suivantes n'est pas instructive et présente peu d'intérêt, les

[3]L'analyse en composantes principales remplace des variables corrélées nombreuses par un nombre réduit de variables non corrélées entre elles, les composantes principales [Saporta (1990)], qui restituent la majeure partie de l'information contenue dans les variables de départ. La première composante est la combinaison linéaire des variables initiales qui restitue une part maximale de la variance des données. La seconde composante, orthogonale, résume ensuite la variance maximale non expliquée par la première composante, et ainsi de suite.

HAP	Nap	Acy	Ace	Fle	Ant	Phe	Flt	Pyr	Baa	Cry	Bap	Bbf	Dba	Bkf	Bgh	Inp
Nap	1.															
Acy	.88	1.														
Ace	.97	.90	1.													
Fle	.93	.97	.94	1.												
Ant	.85	.96	.86	.97	1.											
Phe	.89	.94	.91	.98	.98	1.										
Flt	.78	.93	.80	.93	.99	.95	1.									
Pyr	.76	.93	.79	.92	.98	.94	1.	1.								
Baa	.76	.92	.78	.91	.98	.93	.99	1.	1.							
Cry	.77	.93	.79	.91	.95	.92	.96	.96	.96	1.						
Bap	.72	.90	.74	.87	.94	.89	.97	.98	.98	.96	1.					
Bbf	.67	.88	.72	.85	.94	.89	.97	.98	.98	.95	.98	1.				
Dba	.46	.75	.53	.70	.82	.74	.88	.90	.90	.89	.93	.95	1.			
Bkf	.67	.87	.71	.85	.94	.88	.97	.98	.98	.95	.99	.99	.94	1.		
Bgh	.73	.88	.76	.86	.92	.88	.94	.95	.96	.92	.98	.97	.89	.96	1.	
Inp	.61	.83	.65	.79	.90	.83	.94	.95	.96	.93	.98	.99	.96	.99	.97	1.
Σ16	.87	.96	.88	.97	.99	.98	.98	.98	.98	.96	.96	.94	.82	.94	.94	.91
Phl	.89	.95	.88	.96	.93	.91	.89	.88	.88	.88	.84	.82	.68	.82	.82	.76

TAB. 3.3 – Site Y - Corrélations entre les logarithmes translatés des concentrations des 16 HAP, de leur somme et de l'indice phénol, pour les fosses.

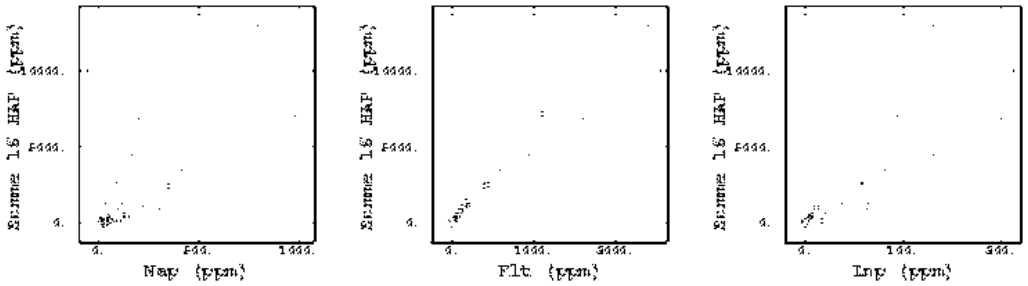


FIG. 3.9 – Site Y - Nuages de corrélation entre la somme des 16 HAP et 3 HAP, pour les sondages.

deux premières composantes restituant à elles seules plus de 98 % de l'information contenue dans les données. Les résultats pour les sondages sont analogues, avec des pourcentages de variance expliquée égaux à 87.1, 6.5, 4 et 1.2 % pour les 4 premiers facteurs. Ces résultats sont cohérents avec ceux obtenus par Oosterbaan-Eritzpokhoff [2000] à partir de 392 données de concentration en HAP provenant de 5 cokeries.

3.2.4 Fraction granulométrique

Chaque échantillon a été prélevé de façon à être le plus représentatif possible de la fosse ou de la carotte d'où il provient. Après séchage à l'air libre, l'échantillon est tamisé à 5 puis 2 mm et l'analyse est effectuée sur un prélèvement de la fraction inférieure à 2 mm. Donc, l'analyse n'est pas réalisée sur l'intégralité de l'échantillon ! Si la masse de la fraction analysée n'est pas constante, la

HAP	Variable brute	Logarithme translaté
Nap	0.87	0.81
Acy	0.95	0.86
Ace	0.58	0.79
Fle	0.90	0.85
Ant	0.84	0.84
Phe	0.82	0.85
Flt	0.84	0.83
Pyr	0.92	0.84
Baa	0.82	0.82
Cry	0.92	0.83
Bap	0.86	0.81
Bbf	0.77	0.80
Dba	0.81	0.71
Bkf	0.77	0.80
Bgh	0.91	0.79
Inp	0.56	0.72
Phl	0.82	0.78

TAB. 3.4 – Site Y - Coefficients de corrélation des concentrations en HAP entre fosses et sondages, en variable brute et en logarithme translaté.

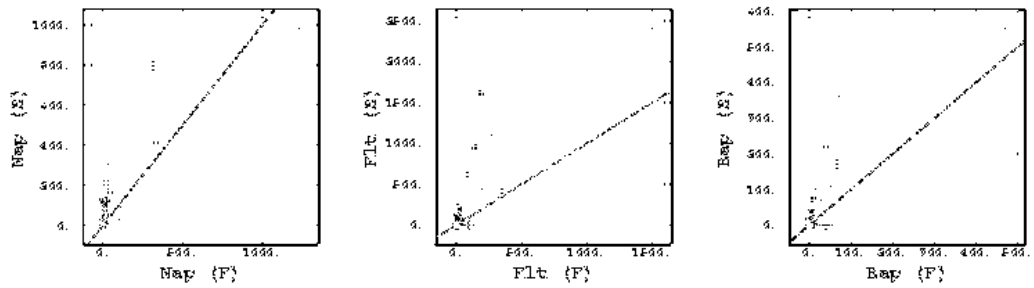


FIG. 3.10 – Site Y - Nuages de corrélation entre fosses (abscisses) et sondages (ordonnées). La première bissectrice est indiquée.

concentration résultante n'est alors plus une variable additive. Travailler en accumulation, qui est le produit de la concentration par la masse de la fraction analysée, conserve la propriété d'additivité. S'il existe en outre une corrélation significative entre la granulométrie et la concentration, le passage en accumulation est d'autant plus nécessaire.

La figure 3.12 montre la variabilité de la masse de la fraction granulométrique inférieure à 2 mm, qui constitue entre 10 et 70 % de la masse totale de l'échantillon. Les concentrations les plus élevées pour les fosses sont obtenues pour des masses relatives de la fraction granulométrique analysée inférieures à 40 %. Pour les sondages, la décroissance des concentrations avec l'augmentation des masses relatives est progressive et systématique. Il semble donc qu'il y ait une relation entre concentration et masse de la fraction granulométrique analysée, ce qui nous conduit à calculer les accumulations et à les comparer aux concentrations. Travailler en accumulation s'avérera nécessaire en cas de corrélation insuffisante entre concentration et accumulation.

Les nuages de corrélation entre concentration et accumulation, illustrés pour trois HAP à la

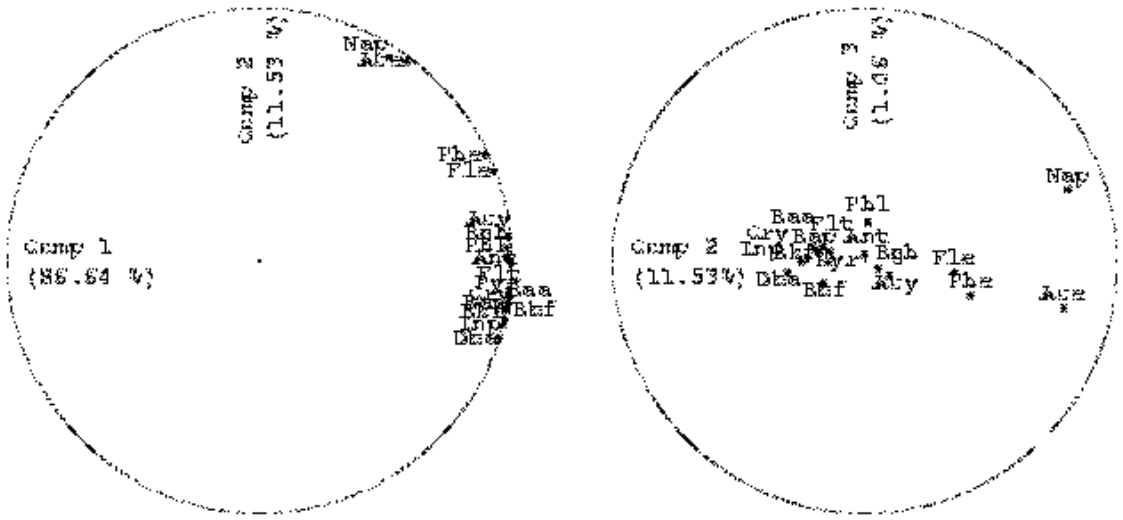


FIG. 3.11 – Site Y - Projection des HAP sur les composantes 1 et 2 puis 2 et 3 de l'analyse en composantes principales réalisée sur les fosses.

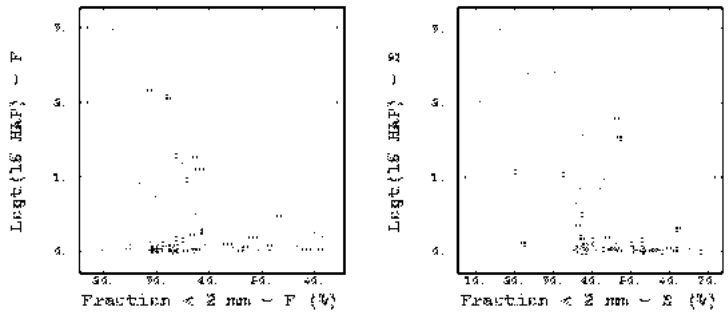
figure 3.13, sont très bons. Cela justifie ici la poursuite du travail en concentration.

Cependant, il n'en va pas de même pour les sondages (voir figure 3.14). Cinq points de concentration élevée ressortent systématiquement du nuage ; ils correspondent au premier mode de l'histogramme de la masse de la fraction granulométrique inférieure à 2 mm.

Donc, aux concentrations les plus élevées sont associées des accumulations qui ne le sont pas. Malgré cela, deux arguments nous poussent à poursuivre le travail en concentration. Tout d'abord, un échantillon de fraction granulométrique inférieure à 2 mm de masse faible, qui indique la présence de particules plus grossières, peut néanmoins présenter des concentrations fortes ; l'importance de ces concentrations élevées est arbitrairement réduite par le produit avec la masse de la fraction inférieure à 2 mm. Cela pose le problème de l'échantillonnage d'un sol hétérogène, fondamental et sur lequel nous reviendrons au chapitre 8. Finalement, travailler en concentration présente l'avantage de conserver des variables dont l'échelle est en $mg.kg^{-1}$, ce qui en rend l'étude plus parlante. Une comparaison entre accumulations et concentrations sera *in fine* réalisée, afin de s'assurer que ce choix n'est pas source de biais.

3.2.5 Indices organoleptiques

Lors de la campagne d'échantillonnage, des indices pédologiques et organoleptiques ont été relevés en chaque point. Informations peu coûteuses qu'il serait intéressant d'intégrer comme *variables auxiliaires* dans notre modélisation des concentrations, ces indices sont-ils de bons indicateurs de la pollution ? Autrement dit, existe-t-il certains indices dont la présence - ou l'absence - correspondrait à l'ensemble des fortes concentrations en HAP, avec éventuellement quelques concentrations plus



(a) Fosses (b) Sondages

FIG. 3.12 – Site Y - Corrélations pour les fosses (a) et les sondages (b) entre le logarithme translaté de la concentration de la somme des 16 HAP et la masse de la fraction granulométrique inférieure à 2mm, en pourcentage de la masse totale de l'échantillon.

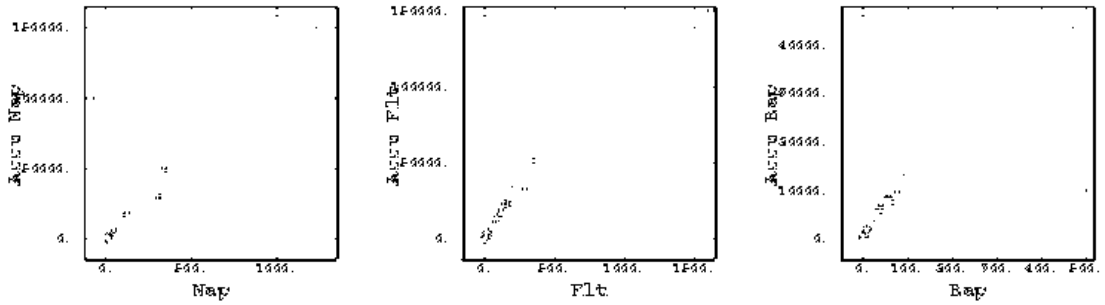


FIG. 3.13 – Site Y - Corrélations pour les **fosses** entre accumulations et concentrations pour 3 HAP.

faibles ? Il est important qu'un tel indice ne "rate" pas de valeurs fortes, car se baser uniquement dessus nous conduirait à manquer certaines taches de pollution. L'utilisation d'un indice comme indicateur des valeurs fortes est dangereux s'il ne les retrouve pas toutes. Notons que nous n'utilisons pas ici de définition a priori de "valeur forte" et "valeur faible" ; l'appréciation de telles valeurs est qualitative.

Pour évaluer la pertinence de chaque indice, plusieurs statistiques ont été comparées : moyennes et médianes par classe (absence/présence), moyenne des rangs par classe et finalement histogrammes de chaque classe. Bien que ces derniers contiennent le plus d'information, il est intéressant de montrer comment, pour un indice, les conclusions peuvent différer selon la statistique choisie.

Huit indices lithologiques et organoleptiques ont été relevés sur les sondages : odeur (absente, légère, forte), goudron, charbon (dans les remblais), charbon (dans le sol), débris de maçonnerie, laitier, couleur verdâtre des limons, craie dans le sol. Notons qu'aucun indice ne s'est avéré utilisable sur les fosses ; le mélange entre sols en place et matériaux rapportés peut être une explication. Il est évident que le codage d'informations naturalistes en indicatrices possède une part certaine de subjectivité ; comment en effet différencier par exemple des "traces de charbon" d'une "présence de fréquentes passées charbonneuses" ? Aussi, la présence d'un indice, même à l'état de traces, est

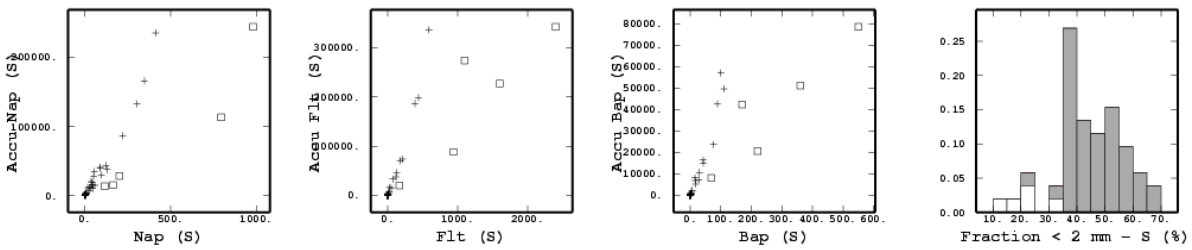


FIG. 3.14 – Site Y - Corrélations pour les **sondages** entre accumulations et concentrations pour 3 HAP. Les points représentés par des carrés sont en blanc sur les histogrammes de la masse de la fraction granulométrique inférieure à 2 mm.

traduite en présence.

Concernant la répartition géographique des indices, notons que le goudron et l'odeur sont cantonnés au Sud et au Nord de la zone centrale. Débris de maçonnerie et remblais charbonneux se répartissent uniformément sur le site, excepté au Nord-Est, seule zone où des traces de craie aient été relevées. Il y a peu de traces de laitier, et une teinte verdâtre des limons a été observée sur la plupart des sondages.

Bien que les matériaux rapportés contiennent les concentrations les plus fortes, nous observons sur le site des limons argileux non exempts de concentrations en HAP importantes.

3.2.5.1 Moyennes et médianes par classe

Commençons par comparer la variation des moyennes et médianes des concentrations en fonction de la présence/absence des indices, et ce pour tous les HAP (voir figures 3.15 et 3.16, où les HAP sont ordonnés par poids croissant, le phénol correspondant au numéro 17). L'allure globale des courbes est liée au niveau de présence des différents HAP. Les HAP 2 (Acy) et dans une moindre mesure 13 (Dba) sont très peu détectés, d'où les faibles valeurs quasi-systématiques des courbes pour ces HAP. Au contraire, on observe dans la plupart des cas un pic des moyennes et médianes pour les HAP 6 à 10, de poids moyens, les plus présents sur les cokeries. Nous ne prenons pas en compte ici les effectifs des classes, bien qu'ils aient une influence directe sur les courbes. Pour pallier cela et compléter l'étude des indices, nous analyserons au paragraphe 3.2.5.3 les histogrammes des concentrations par classe, pour quelques HAP.

Commentons à présent plus en détail ces indices :

- Odeur et goudron : la discrimination est excellente pour ces deux indices, très clairement indicateurs de pollution, à la fois pour la moyenne et la médiane. Pour l'odeur, il semble que l'on puisse tout aussi efficacement mélanger les classes "absence d'odeur" et "légère odeur", cette dernière ne contenant apparemment pas de concentrations élevées - dans le cas contraire ces concentrations auraient des répercussions sur les moyennes.
- Charbon (dans les remblais) : la discrimination est également très bonne, et la présence de charbon est ici liée à une pollution élevée.
- Maçonnerie : la distinction des deux classes est moins nette, excepté pour les HAP de poids

intermédiaire, et également les HAP de poids lourds pour la médiane.

– Laitier, couleur verdâtre, charbon (dans le sol) et craie (dans le sol) : les moyennes pour ces indices sont très similaires. Les moyennes correspondant à l'absence des indices sont systématiquement supérieures, mais la différence entre les deux classes ne semble significative que pour les HAP de poids intermédiaire. Par contre, les médianes diffèrent : pour la couleur verdâtre et le charbon, les choses ne sont pas claires pour les HAP légers, et à partir de Phe (6), c'est la présence de ces indices qui conduit à des concentrations médianes plus élevées. Les valeurs les plus fortes sont donc probablement dans la classe "absence", mais d'autres valeurs de concentrations importantes sont dans la classe "présence" des indices. Pour le laitier, les médianes des deux classes sont quasiment confondues pour tous les HAP ; les valeurs fortes sont donc dans la classe "absence", mais leur nombre réduit entraîne la quasi-égalité des médianes. Finalement, pour la craie la médiane de la classe "absence" est systématiquement plus élevée, mais cela n'est significatif que pour les HAP de poids intermédiaire.

3.2.5.2 Moyenne des rangs

Les concentrations sont rangées par ordre croissant, et se voient attribuer des "rangs" allant de 1 à 52, le nombre de données. Ensuite, la moyenne de ces rangs est calculée d'une part pour les points présentant l'indice, et d'autre part pour ceux où l'indice est absent. Cette statistique présente l'avantage d'être robuste par rapport aux valeurs extrêmes, car seule compte la position relative des concentrations correspondant aux deux classes.

Tous les indices se révèlent discriminants par rapport à la pollution pour la moyenne des rangs, de manière assez significative excepté pour le charbon dans le sol, où l'écart est inférieur à 10%, et pour les HAP légers dans le cas de la craie, de la couleur verdâtre et du laitier (voir figure 3.17). Ce critère accentue donc la discrimination entre les classes absence et présence, par rapport aux moyennes et médianes par classe.

3.2.5.3 Histogrammes par classe

Pour chaque indice, les histogrammes par classe (voir figure 3.18 pour Nap, HAP le plus léger, Flt le plus présent et Bap, à 5 cycles) montrent que :

- Goudron : tous les points présentant du goudron possèdent des concentrations élevées, ce qui rend l'indice intéressant même si certaines concentration fortes ne sont pas détectées.
- Odeur : la présence d'odeur (forte) conduit à de fortes concentrations, mais ces dernières ne sont pas toutes repérées.
- Charbon (remblais) : l'absence de charbon conduit à des concentrations faibles ou moyennes (à une exception près), et la présence à des concentrations fortes, et parfois à des valeurs faibles. Cet indice semble donc utilisable, en tenant compte du fait que ce n'est qu'un indicateur approximatif de la pollution.
- Débris de maçonnerie, laitier : l'absence de débris correspond à des concentrations fortes ou faibles, la présence à des concentrations moyennes. Il semble donc que la classe "présence" soit séparable en deux classes : une pour laquelle cette présence est liée à de faibles concentrations, l'autre étant associée à de fortes concentrations. Cependant, cette dichotomie ne peut s'interpréter de manière intéressante. En effet, géographiquement, on observe uniquement que les données de type "forte concentration-absence de maçonnerie" sont situées dans les zones polluées au Sud et au centre Nord, ce qui est logique, et à proximité de données "forte concentration-présence de maçonnerie". La répartition des données présence-absence est sensiblement identique pour le laitier, mais le nombre de données "présence de laitier" est beaucoup plus faible que pour la maçonnerie. Cela explique la similarité pour ces deux indices des moyennes par classe, et la différence entre les médianes.
- Craie (sol) : à part quelques exceptions, la présence de craie conduit à des concentrations faibles, en particulier pour les HAP lourds. Cette information n'est cependant guère exploitable.
- Charbon (sol), couleur verdâtre : aucune discrimination des concentrations n'est possible avec ces indices, les données des deux classes étant très mélangées. Les moyennes et médianes par classe ainsi que la moyenne des rangs étaient semblables pour ces deux indices.

Il peut sembler étrange que la discrimination ne soit pas meilleure avec l'odeur ; cela s'explique par le fait que seul le naphtalène en ait une. Pour le goudron, certaines concentrations fortes ont été détectées sans que l'on ait pour autant pu observer de goudron sur les carottes correspondantes. L'utilisation de cet indice pourra néanmoins être envisagée, par exemple par cokrigeage avec les teneurs. Un autre indice semble relativement bien adapté à notre objectif : la présence de charbon dans les remblais.

Finalement, même si certains indices tels que la présence de craie, de maçonnerie ou de goudron ne sont pas utilisables seuls, leur croisement peut être envisagé. Cela a été testé, mais sans résultat : la sélection des concentrations fortes en utilisant l'intersection ou la réunion entre plusieurs de ces indices n'en est pas améliorée.

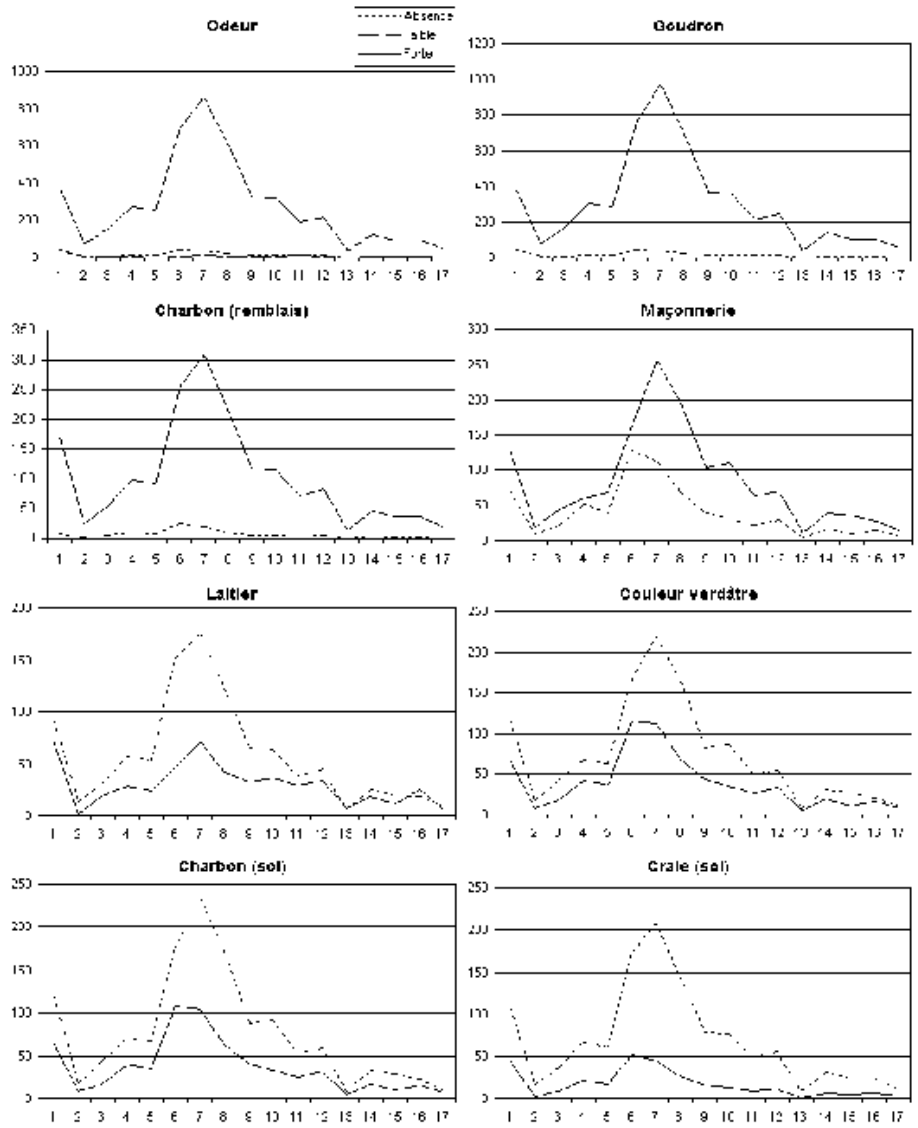


FIG. 3.15 – Site Y - Indices qualitatifs sur les sondages. Les numéros de HAP en abscisse correspondent à l'ordre du tableau 3.1 (1 : Nap, 17 : Phl) ; **moyennes par classe** en ordonnée (ppm). Le trait plein correspond à la présence de l'indice, les tirets à son absence.

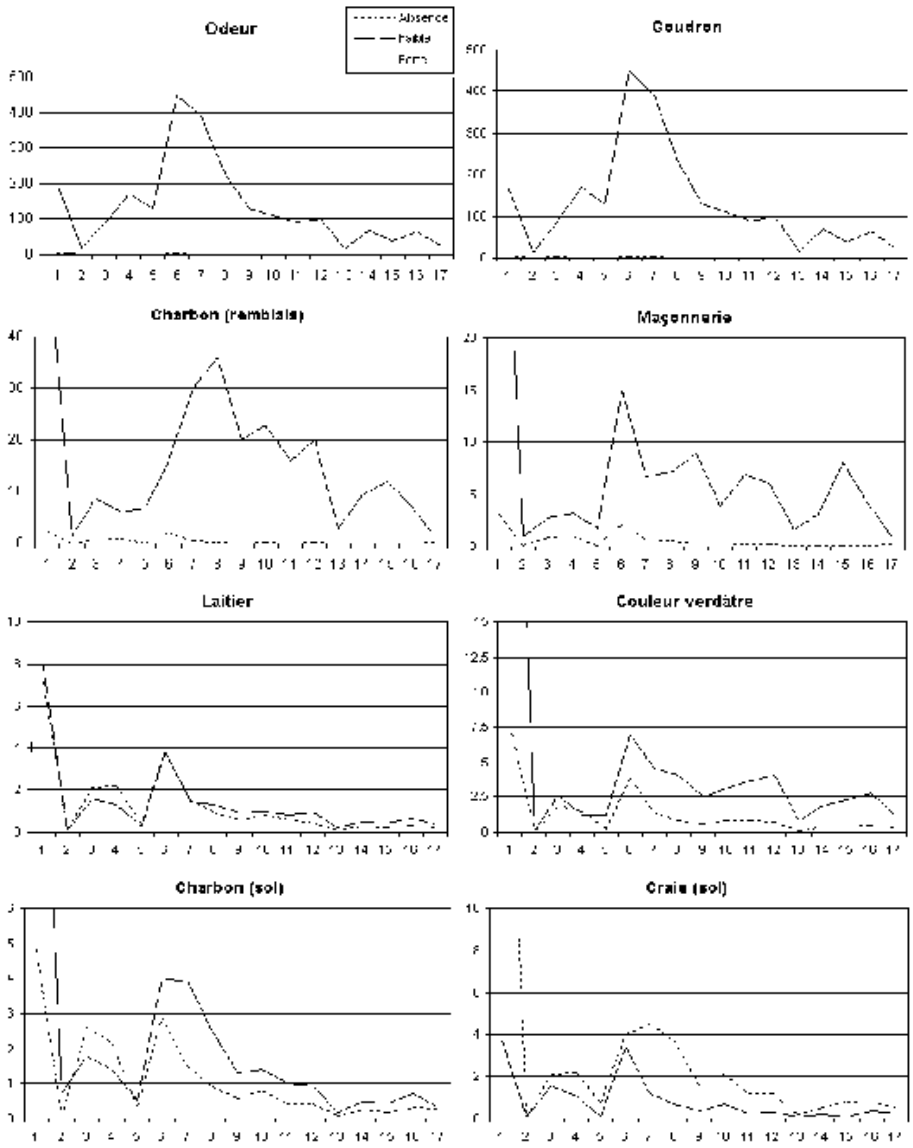


FIG. 3.16 – Site Y - Indices qualitatifs sur les sondages. Les numéros de HAP en abscisse correspondent à l'ordre du tableau 3.1 (1 : Nap, 17 : Phl) ; **médianes par classe** en ordonnée (ppm). Le trait plein correspond à la présence de l'indice, les tirets à son absence.

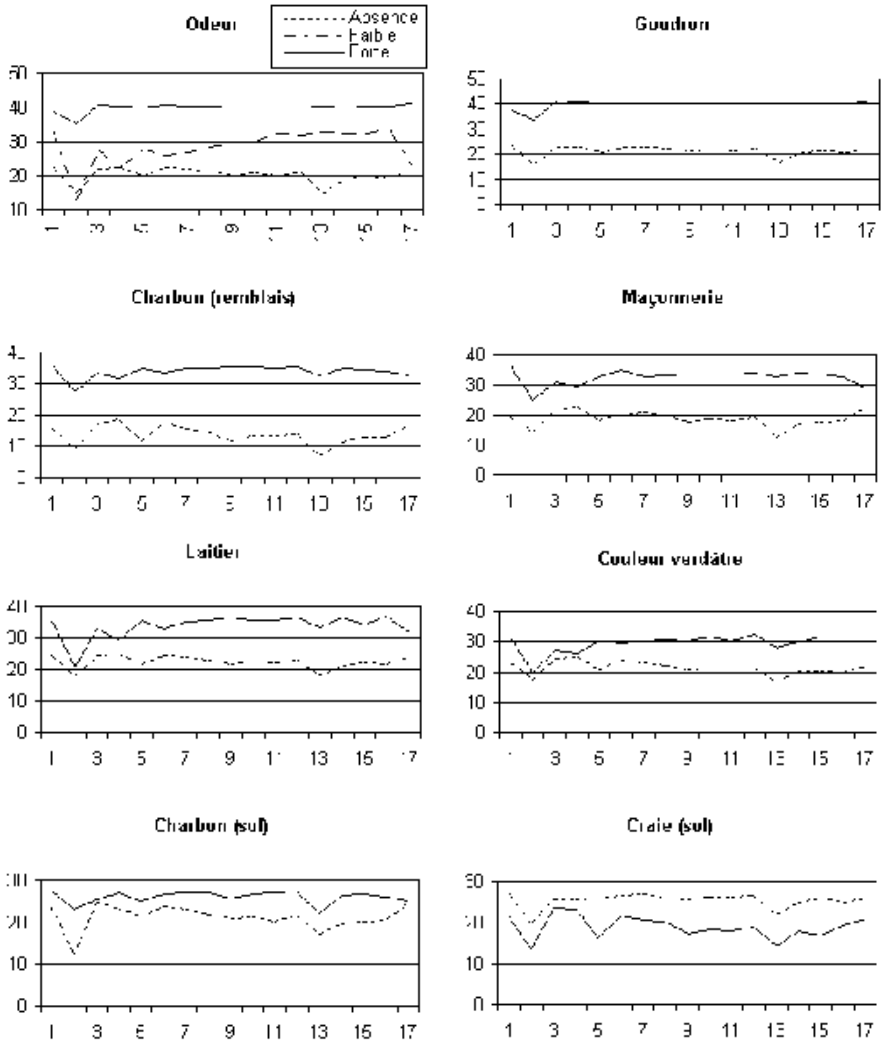


FIG. 3.17 – Site Y - Indices qualitatifs sur les sondages. Les numéros de HAP en abscisse correspondent à l'ordre du tableau 3.1 ; **critère de la moyenne des rangs**. Le trait plein correspond à la présence de l'indice, les tirets à son absence.

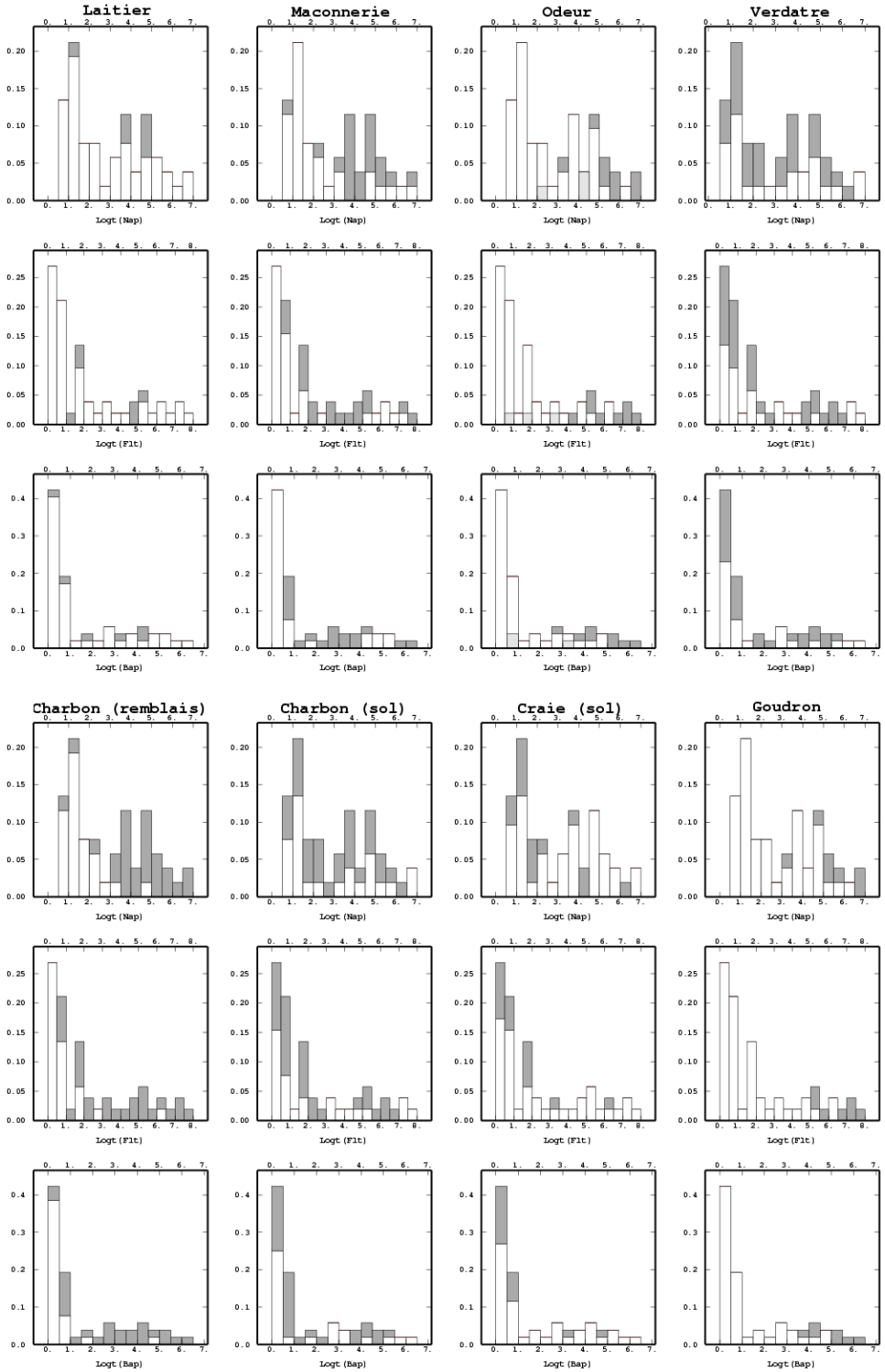


FIG. 3.18 – Site Y - Histogrammes par classe du log-translaté de 3 HAP. Blanc : absence de l'indice ; gris : présence. Pour l'odeur, le gris clair (resp. foncé) indique une légère (resp. forte) présence.

3.2.5.4 Analyse des correspondances

L'information apportée par les indices peut se résumer par leur combinaison. En cas de corrélation avec les concentrations, cette combinaison pourra être utilisée par cokrigeage.

Une AFC[4] a été menée pour les 8 indices des sondages. Les deux premiers facteurs expriment respectivement 33.6% et 23.5% de la variance totale (voir figure 3.19) ; l'information des facteurs suivants n'est pas exploitable.

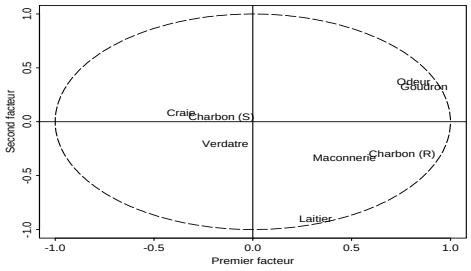


FIG. 3.19 – Site Y - Analyse des correspondances sur les indices qualitatifs des sondages : projection des indices sur les deux premiers facteurs.

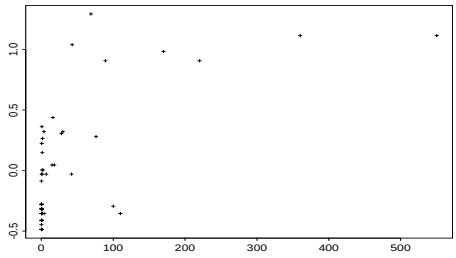


FIG. 3.20 – Site Y - Nuage de corrélation entre le premier facteur auxiliaire et la concentration en Bap sur les sondages.

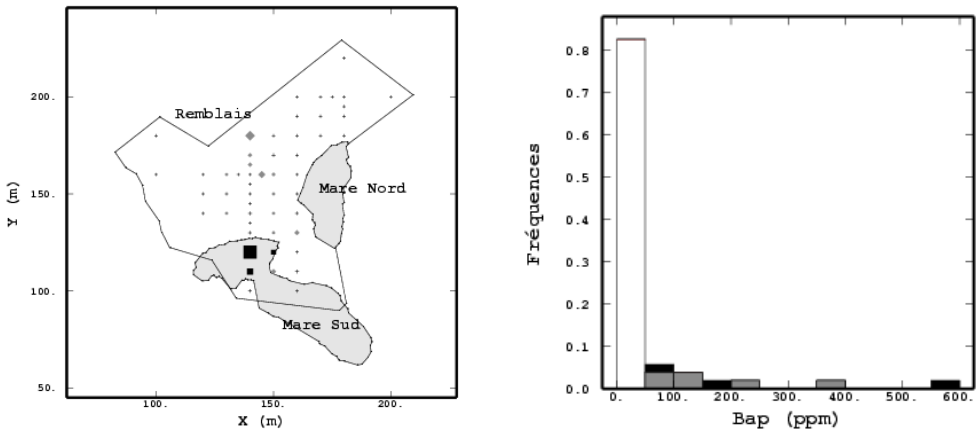
Les trois indices Craie, Couleur Verdâtre et Charbon (S), liés au sol en place, se distinguent des autres indices associés aux matériaux rapportés. Le premier facteur, appelé dans la suite *facteur auxiliaire*, semble donc distinguer le sol en place (valeurs plutôt négatives) des remblais et matériaux rapportés (valeurs élevées). La proximité des indices Odeur et Goudron est cohérente, sur le terrain le goudron étant source d'odeur.

Le nuage de corrélation entre Bap et le facteur auxiliaire (voir figure 3.20) montre que les teneurs faibles correspondent essentiellement à un sol en place, tandis que les concentrations fortes et intermédiaires correspondent aux remblais. Ce résultat important justifiera l'utilisation du facteur auxiliaire par cokrigeage avec les concentrations en HAP. L'étude de la corrélation entre les composantes de l'ACP des concentrations en HAP et le facteur auxiliaire n'a rien apporté.

[4] L'analyse factorielle des correspondances (AFC) est une technique d'exploration de données multivariables qualitatives qui consiste à réduire un nombre élevé de variables à quelques composantes non corrélées qui restituent une partie importante de l'information contenue dans les variables de départ. Dans le cas de variables indicatrices, l'AFC se déduit d'une analyse en composantes principales par transformation de la matrice des données [Greenacre (1984), Lebart et al. (1979)].

3.2.6 Liens avec l'historique

La figure 3.21(a) reprend l'implantation des données, avec l'indication des contours du site accessible et des anciennes mares à goudron. Ces mares ont été excavées, et le résultat de l'excavation de la mare Nord constitue les remblais situés au Nord-Ouest du site. La mare Sud a été remblayée. Que tirer de ces informations sur la localisation des zones *a priori* "polluées" ?



(a) (b)

FIG. 3.21 – Site Y - (a) Implantation des données de Bap sur les sondages, informations données par l'historique du site. (b) Histogramme de Bap, concentrations des données situées dans l'ancienne mare Sud (en noir, représentées sur la carte par des carrés), autres points de concentration supérieure à 50 ppm (en gris, représentés sur la carte par des losanges).

Les informations historiques conduiraient à distinguer les 4 données situées sur la mare Sud des autres. Cependant, d'autres données sont localisées à proximité de la mare, et Pitout (2000) a montré que le contour de la mare Sud est incertain. Du goudron a bien été repéré sur 3 des 4 points de la mare, et pas en dehors. Par contre, une forte odeur a été décelée sur le point le plus au Sud-Ouest, bien qu'il ne soit pas localisé sur la mare.

Par ailleurs, d'autres points du site présentent des concentrations élevées en Bap (voir figure 3.21(b)). Bien que celles-ci s'expliquent par la proximité des remblais ou de mares à goudron, nous ne connaissons pas l'extension des taches de pollution dont ces points proviennent.

En conclusion, les informations relatives à l'historique du site renseignent des zones de fortes concentrations, mais ne suffisent pas à les détecter toutes : ainsi, parmi les 9 sondages pour lesquels la concentration en Bap est supérieure à 50 ppm, seuls 3 sont explicables par l'historique. Il est par conséquent délicat sur un tel site d'utiliser ces informations pour guider l'échantillonnage, et un échantillonnage systématique du site est nécessaire.

3.3 Synthèse

Les concentrations des 16 HAP analysés sont des variables très contrastées, avec une proportion importante de valeurs très faibles, et quelques valeurs extrêmement élevées. Ces variables présentent pour la plupart un fort pourcentage de valeurs inférieures au seuil de détection de l'analyse, qui sont ramenées à cette valeur. Lors de la préparation des échantillons, seul un prélèvement de la fraction granulométrique inférieure à 2 mm est analysé, ce qui n'est pas sans conséquence sur les concentrations des sondages.

Il y a prédominance des HAP à 3 et 4 cycles, notamment le fluoranthène, le phénanthrène et le pyrène, ce qui est classique pour les sites de cokeries. Le naphtalène est très présent sur les fosses, ce qui est explicable par sa volatilité. Le niveau de corrélation entre les HAP est très bon, et d'autant meilleur que les HAP ont un poids atomique proche. Le comportement du phénol semble par ailleurs pouvoir être rapproché de celui du naphtalène.

Les concentrations fortes sont localisées préférentiellement à proximité de la mare à goudron Sud, bien que celle-ci ait été excavée et remblayée avec des matériaux supposés sains. Certaines valeurs fortes ressortent également au Nord de la zone centrale, en particulier sur les sondages. La zone située au Nord-Est du site présente exclusivement des concentrations très faibles.

Deux techniques de prélèvement ont été mises en œuvre, sur deux niveaux de profondeur différents. Les différences de concentration observées sont vraisemblablement attribuables à un effet profondeur plus que lié à la technique. Les concentrations fortes se situent ainsi plutôt en profondeur, une couche de matériaux rapportés sains ayant probablement été étendue sur le site. On perçoit là le risque d'un échantillonnage qui consisterait à s'arrêter verticalement dès qu'une concentration jugée négligeable est mesurée !

La corrélation entre les concentrations en HAP et plusieurs indices qualitatifs prélevés lors de la campagne a été analysée. Deux indices gagneront à être utilisés dans le cas des sondages : la présence de goudron et celle de charbon dans les remblais. En outre, une AFC réalisée sur les indices qualitatifs a mis en évidence l'existence d'un facteur auxiliaire distinguant sol en place et matériaux rapportés. Ce facteur étant corrélé aux concentrations, il sera également utilisé lors de l'estimation des concentrations.

L'utilisation de l'historique est fréquemment préconisée lors d'une étude de risque pour guider la reconnaissance. Nous avons illustré que cette information historique peu détaillée ne permet pas de cibler l'échantillonnage sans risquer de ne pas détecter certaines taches de pollution. Une reconnaissance systématique du site est donc tout à fait nécessaire.

A présent, nous allons avec l'analyse variographique aborder la prise en compte du caractère spatialisé des données.

Chapitre 4

Analyse variographique

Sommaire

Phase essentielle de toute étude géostatistique, l'analyse variographique a constitué un volet important du travail. Outre le travail méthodologique présenté au chapitre 2, plusieurs outils ont été comparés afin de déterminer lesquels sont les plus appropriés à nos variables fort contrastées. Nous illustrons ce travail pour deux HAP sur les sondages, avant d'en appliquer les conclusions aux autres HAP. Plusieurs modèles multivariables permettant la prise en compte des corrélations entre HAP d'une part et entre HAP et variables auxiliaires d'autre part sont discutés.

4.1 Différents modes de calcul du variogramme

L'analyse structurale est souvent délicate pour des variables présentant une dissymétrie marquée. Nous illustrons cela pour deux HAP, tout d'abord à l'aide du variogramme classique ; l'apport de transformées de la variable brute pour la mise en évidence d'une structure est ensuite discuté. Présentant de meilleures propriétés de robustesse, le variogramme déduit de la covariance non centrée - qui nécessite l'hypothèse de stationnarité d'ordre deux de la fonction aléatoire - et les variogrammes pondérés sont ensuite comparés au variogramme classique.

4.1.1 Variogramme classique

4.1.1.1 Variable brute

La nuée variographique du naphtalène Nap sur les sondages (voir figure 4.1), très dispersée, montre le contraste existant entre quelques valeurs fortes et le reste des données. Le retrait des

deux valeurs les plus fortes[1], situées au Sud, entraîne une diminution de la variance d'un facteur supérieur à 3, visible sur le variogramme expérimental.

On ne retient fréquemment du variogramme expérimental que les distances inférieures à la moitié de la taille du champ, le faible nombre de couples au-delà rendant les conditions d'estimation mauvaises[2]. Ici, malgré les irrégularités des contours, on peut considérer que cette distance est de l'ordre de 60 m, et l'on constate effectivement sur la figure 4.1 que le nombre de couples a déjà sensiblement diminué. Il est néanmoins utile de pousser plus loin l'observation du variogramme expérimental. En effet, dans le cas présent, s'arrêter à 60 m nous amènerait à conclure à la stationnarité de la variable, un palier apparaissant vers 40-50 m. Or, l'observation du variogramme aux distances supérieures indique l'existence possible d'une non stationnarité. Le prétendu palier peut n'être du qu'à une fluctuation du variogramme pour ces pas de calculs, qui s'explique par le sur-échantillonnage à 5 m de la zone centrale du champ. Les données de ce resserrement, essentiellement faibles, apportent en effet avec les deux valeurs fortes situées au Sud, à une distance moyenne de 40-50 m, des contributions particulièrement fortes au variogramme expérimental. Nous reviendrons ci-dessous à ce sur-échantillonnage.

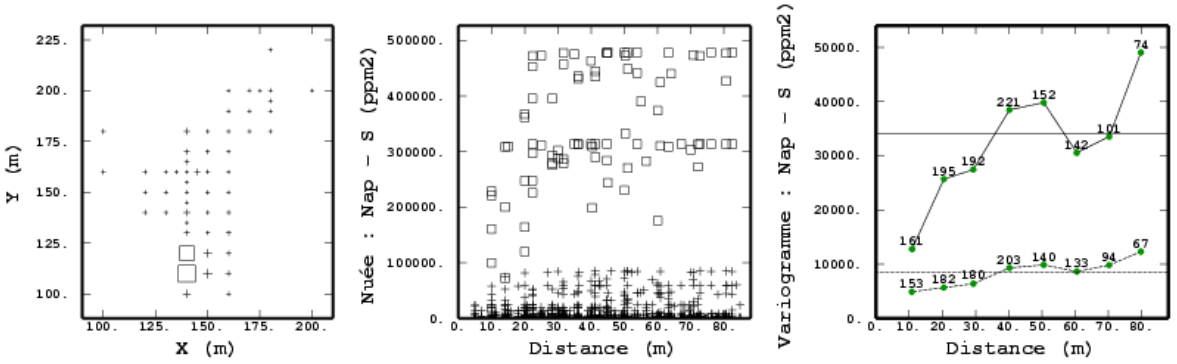


FIG. 4.1 – Site Y - Nap sur les sondages : carte d'implantation, nuée variographique et variogrammes expérimentaux avec (trait plein) et sans (tirets) les deux valeurs fortes, repérées sur la carte et la nuée par des carrés.

Donc, malgré son caractère quelque peu erratique, le variogramme expérimental tend à indiquer l'existence d'une non stationnarité qui persiste après le retrait des deux valeurs fortes. Par ailleurs, tandis que le variogramme expérimental avec toutes les données laisse penser à un effet de pépite relatif faible, cette variabilité relative à petite distance est fortement accrue après retrait des valeurs fortes. Une part importante de la structure semble donc provenir des deux valeurs fortes, soit 4 % des données.

Le découpage du champ conduisant à une reconnaissance médiocre de la direction Est-Ouest, la recherche d'anisotropie a été rapidement abandonnée. Même si cela ne signifie pas qu'elle soit absente, une telle anisotropie ne trouverait aucune justification physique ou liée à l'historique.

[1] Le retrait de valeurs fortes est à déconseiller, ces valeurs étant justement celles qui nous intéressent. Ce retrait a ici pour simple objectif d'illustrer l'influence de ces valeurs peu nombreuses sur le variogramme expérimental.

[2] Matheron (1970) montre que dans le cas d'un schéma linéaire à une dimension sur un intervalle $(0, L)$, la variance d'estimation relative de $\gamma(h)$ reste suffisamment faible pour permettre l'inférence statistique, pour $h \leq \frac{L}{2}$.

Pour Bap sur les sondages (voir figure 4.2), les valeurs fortes ont également une grande influence sur la nuée variographique, et par conséquent sur le variogramme expérimental. La stationnarité est ici acceptable et on observe une portée de l'ordre de 40 m. Ces caractéristiques restent valides après retrait des valeurs les plus fortes, et l'effet de pépite demeure relativement faible.

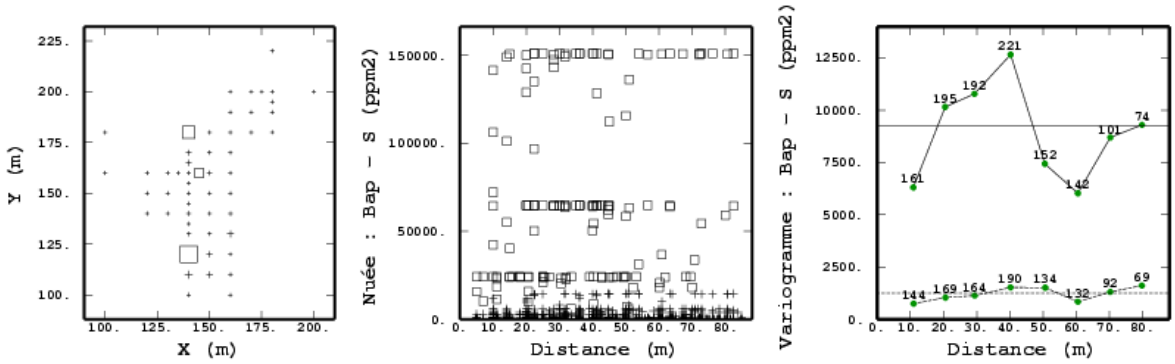


FIG. 4.2 – Site Y - Bap sur les sondages : carte d'implantation, nuée variographique et variogrammes expérimentaux avec (trait plein) et sans (tirets) les deux valeurs fortes, repérées sur la carte et la nuée par des carrés.

4.1.1.2 Effet proportionnel et prise en compte des croix de sondages

L'échantillonnage, réalisé sur une maille régulière de 10 m de coté, présente quelques resserrements à 5 m au centre et au Nord-Est ayant pour objectif d'améliorer la connaissance de la structure aux petites distances (voir figure 4.1). Ceux-ci ont été volontairement implantés dans des zones si possible représentatives du comportement moyen des variables [Bourgine & Niandou (1993)] ; en pratique, on a surtout évité leur implantation à proximité des mares à goudron.

Il convient de prendre en compte le sur-échantillonnage qui découle de ces resserrements si l'on ne veut biaiser le calcul variographique expérimental. Une technique consiste à attribuer à chaque donnée un poids inversement proportionnel à la densité de l'échantillonnage autour de la donnée ; cette technique de *declustering* est discutée à la section 4.1.3.

Il est également possible de calculer dans un premier temps la structure pour la maille principale à 10 m. Les variogrammes ainsi obtenus pour Nap et Bap (voir figure 4.3) ont sensiblement les mêmes caractéristiques que ceux pour l'ensemble des données. Ensuite, les variogrammes des 18 points appartenant aux croix de sondages sont calculés. A ce stade, nous négligeons cependant la différence entre les niveaux de variabilité des deux populations ayant servi aux calculs de variogramme - les zones fort hétérogènes impliquées dans le calcul du variogramme à 10 m, ont été volontairement évitées lors des resserrements ; il y a là un risque de sur-estimation ou de sous-estimation de l'effet de pépite auquel il est essentiel de remédier.

La figure 4.4 montre pour deux HAP l'existence d'un *effet proportionnel* caractéristique des distributions dissymétriques. Matheron (1974) montre qu'un tel effet proportionnel n'est pas incompatible avec une hypothèse de stationnarité globale. Cet effet peut être modélisé en considérant

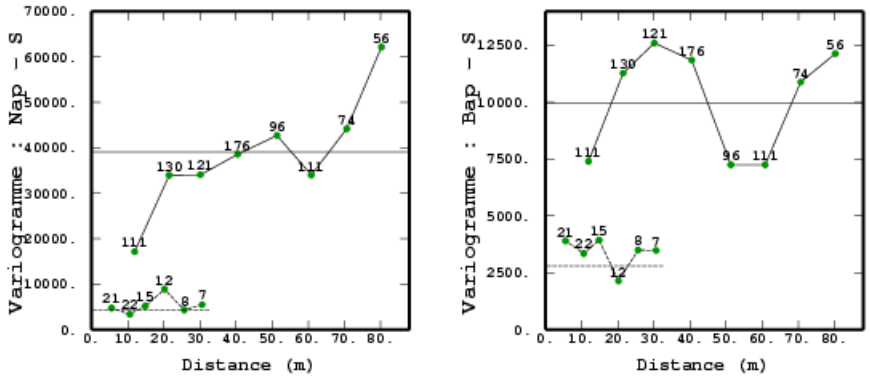


FIG. 4.3 – Site Y - Nap et Bap sur les sondages : variogrammes expérimentaux des données à 10 m (traits pleins), avec indication de la variance, et des données des croix à 5 m (tirets), avec indication de la variance.

localement dans un voisinage V_0 le variogramme

$$\gamma_{V_0}(h) = f(m_{V_0})\,\gamma(h) \tag{4.1}$$

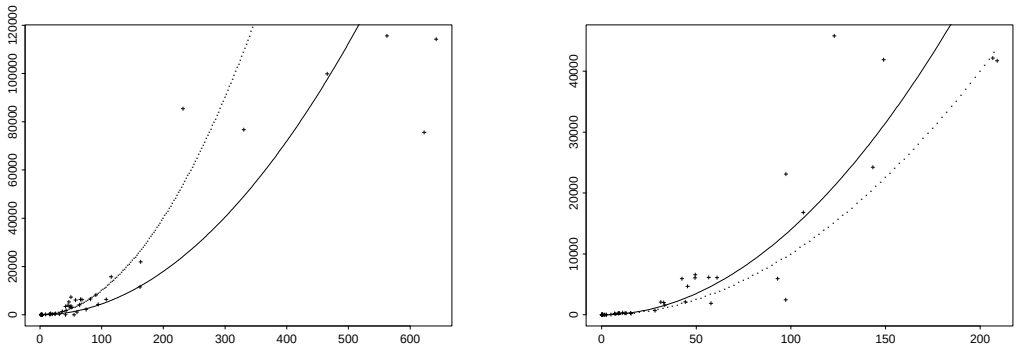
où f est une fonction de la moyenne locale m_{V_0} et $\gamma(h)$ un modèle global [Matheron (1970), Journel & Huijbregts (1978), Chilès & Delfiner (1999)].

Dans le cas lognormal, $f(m)$ est en m^2. En considérant les courbes $f(m) = \alpha\, m^2$, un ajustement de α par moindres carrés conduit pour Nap et Bap à $f_{\text{Nap}}(m) = 0.45\, m^2$ et $f_{\text{Bap}}(m) = 1.4\, m^2$. L'effet proportionnel permet de recaler le variogramme local des croix de sondages $\gamma_l(h)$ au variogramme calculé à partir de la maille à 10 m $\gamma_g(h)$. Il découle de l'équation 4.1 que

$$\frac{\gamma_l(h)}{f(m_l)} = \frac{\gamma_g(h)}{f(m_g)}$$

avec m_l la moyenne locale et m_g la moyenne globale. Une fois la courbe $\sigma^2 = f(m)$ ajustée, le recalage du variogramme local $\gamma_l(h)$ s'obtient directement en multipliant ce dernier par le rapport $\frac{f(m_g)}{f(m_l)}$. Les variogrammes ainsi recalés sont illustrés à la figure 4.5. Les structures des deux HAP s'avèrent donc bien moins continues que sans le recalage ; cette différence de comportement à petite distance a une influence certaine sur les estimations.

En conclusion, les deux variables présentent des variogrammes fluctuants et une variabilité à petite distance importante. On a montré l'influence des quelques valeurs fortes sur les structures, et comment il est possible de prendre en compte les croix de sondages. Les deux variables se comportent différemment aux grandes distances : tandis que le variogramme de Bap atteint un palier pour une distance de 40 m, celui de Nap semble présenter une non stationnarité. Ces conclusions doivent être tempérées vu l'incertitude associée aux structures de variables aussi dissymétriques.



(a) Nap (b) Bap

FIG. 4.4 – Site Y - Nuages de corrélation entre moyennes et variances locales calculées sur des fenêtres glissantes circulaires de rayon égal à 10 m pour Nap (a) et Bap (b) sur les sondages. Meilleur ajustement par moindres carrés de la régression en trait plein, courbe $\sigma^2 = m^2$ en pointillés.

4.1.1.3 Transformées

Nous cherchons ici à valider les observations faites sur la variable brute en recourant à quelques transformées présentées au chapitre 2 : logarithme translaté, transformée gaussienne et indicatrice de la médiane, dont les histogrammes et les variogrammes expérimentaux sont repris à la figure 4.6. Le variogramme d'ordre 1 de la variable brute, également envisagé, n'apporte pas d'information supplémentaire.

Nap ne présente pas de données inférieures au seuil de détection sur les sondages ; la première classe de l'histogramme du logarithme translaté n'est constituée que de concentrations bien détectées. Cet histogramme reste fort dissymétrique. Le variogramme du logarithme translaté est analogue à celui de la variable brute. L'histogramme de la transformée gaussienne de Nap est symétrique, et le variogramme présente une non stationnarité plus marquée que sur la variable brute ou le logarithme translaté. L'effet de pépite est accru, et cela est encore plus vrai pour le variogramme de l'indicatrice de la médiane, cette dernière valant 8.3 ppm pour Nap. Dans le cas d'un modèle gaussien, l'indicatrice de la médiane est l'indicatrice la mieux structurée ; cela n'est manifestement pas le cas ici.

Les transformées de Bap confirment également les conclusions obtenues sur la variable brute, avec un effet de pépite plus important sur la transformée gaussienne et l'indicatrice de la médiane.

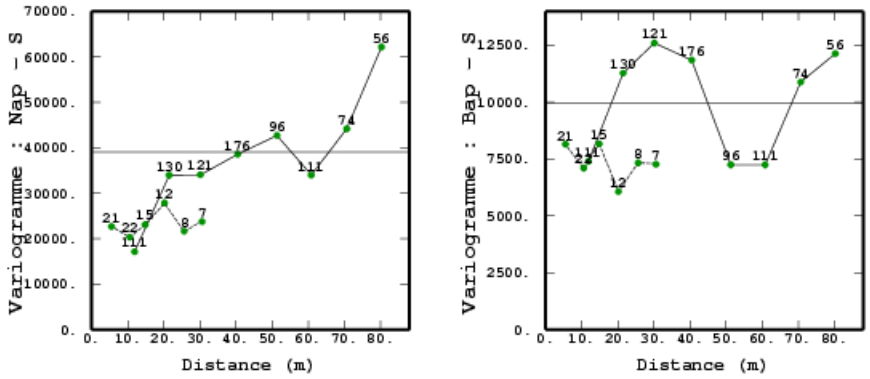


FIG. 4.5 – Site Y - Nap et Bap sur les sondages : variogrammes expérimentaux des données à 10 m (traits pleins), avec indication de la variance, et des données des croix à 5 m (tirets) recalés.

4.1.2 Variogramme déduit de la covariance non centrée

L'étude méthodologique de $C(0) - C(h)$ a montré que son calcul est particulièrement instructif dans le cas de variables fort contrastées, à condition que la stationnarité de la variable sur le champ d'étude puisse raisonnablement être admise.

Pour Nap sur les sondages, les différences entre le variogramme expérimental classique et celui déduit de la covariance non centrée sont sensibles pour la variable brute, que ce soit pour la portée ou le palier (voir figure 4.7). Le comportement aux petites distances, approché par le calcul sur les croix de sondages à 5 m, est analogue pour les deux outils. Contrairement au variogramme classique, $C(0) - C(h)$ tend à indiquer la stationnarité de la variable, en brut comme sur le logarithme translaté. Le comportement moins erratique de $C(0) - C(h)$ ne doit pas ici être un leurre, et il est préférable de s'abstenir d'utiliser cet outil étant donné les réserves sur la stationnarité de la variable.

La stationnarité admissible de Bap pour les sondages rend licite et recommandable le calcul de $C(0) - C(h)$, qui présente moins de fluctuations que le variogramme classique (voir figure 4.8). Les deux outils ont un comportement similaire aux petites distances, en brut comme en logarithme translaté.

Notre conclusion méthodologique est donc vérifiée en pratique : à condition que la stationnarité de la variable soit acceptable, $C(0) - C(h)$ apporte un complément d'information non négligeable par rapport au variogramme classique.

4.1.3 Variogrammes pondérés

Le premier type de pondération affecte un poids à chaque donnée, ce poids pouvant être inversement proportionnel à la densité de l'échantillonnage au voisinage de la donnée. Nous approchons cette densité d'échantillonnage en chaque point par le nombre de points situés dans un voisinage circulaire de rayon égal à 20 m ; le poids affecté au point est l'inverse du nombre de points situés

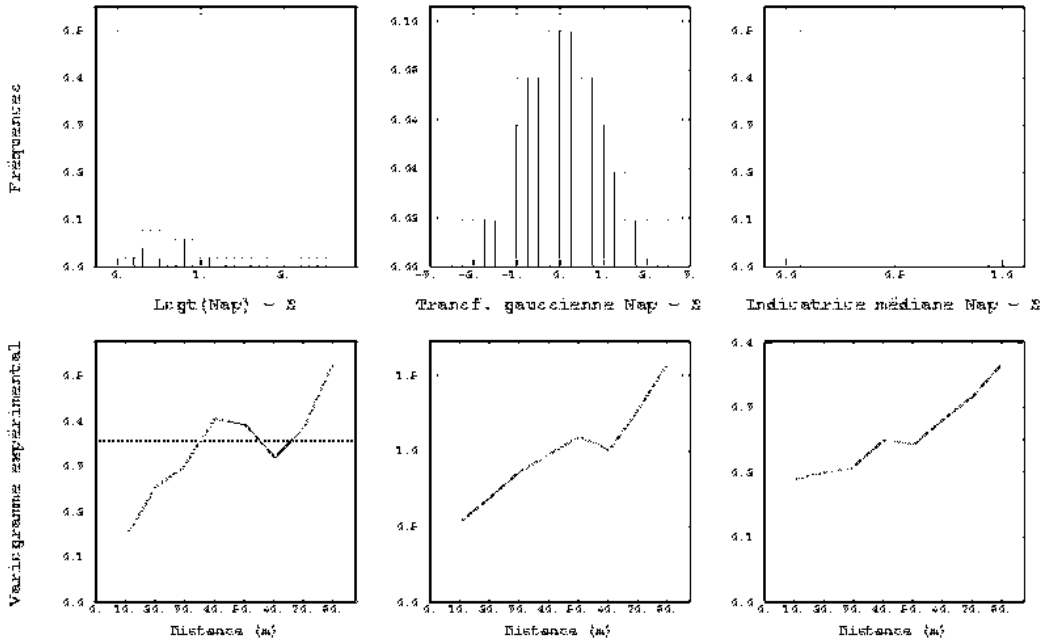


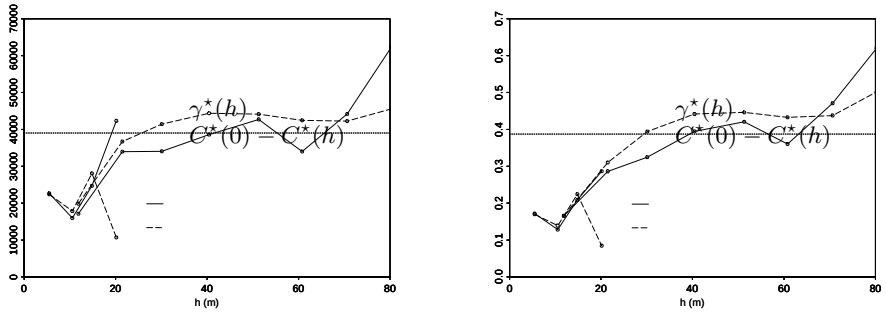
FIG. 4.6 – Site Y - Histogrammes et variogrammes expérimentaux de trois transformées de Nap sur les sondages : logarithme translaté, transformée gaussienne et indicatrice de la médiane.

dans le voisinage. Le variogramme pondéré par échantillon (p.p.e.) résultant est illustré aux figures 4.9 et 4.10 respectivement pour Nap et Bap. Ces figures contiennent également le variogramme classique ainsi que les variogrammes moyens par échantillon (m.p.e.). Il n'y a aucune difficulté méthodologique à pondérer selon le même procédé $C(0) - C(h)$, cette pondération ne modifiant cependant pas l'hypothèse de stationnarité d'ordre deux nécessaire à cet outil.

La variance des données pondérées est plus élevée, les resserrements ayant été réalisés dans une zone peu hétérogène de valeurs plutôt faibles (voir figure 4.9). Les valeurs les plus fortes sont quant à elles situées au Sud, en bordure de champ, dans une zone où la densité d'échantillonnage est plus faible, d'où des pondérateurs élevés. Les deux structures p.p.e. ont un comportement identique aux petites distances. Ensuite, l'influence des resserrements diminuant, elles ont tendance à ressembler, au niveau de variance près, aux structures non pondérées correspondantes.

Les structures moyennes par échantillon sont toutes les deux comprises entre les structures non pondérées. $C(0) - C(h)$ m.p.e. tend à indiquer une stationnarité artificielle, et est tout autant que la version non pondérée à déconseiller dans le cas présent. Pour le variogramme classique m.p.e., les fluctuations sont amoindries. Le variogramme m.p.e. se resserre autour de la covariance non centrée, plus robuste, excepté aux grandes distances où l'on retrouve la croissance du variogramme pondéré.

Comme pour Nap, Bap présente un écart entre les variances pondérées et non pondérées. Le variogramme p.p.e. reste relativement erratique. $C(0) - C(h)$ est quasiment inchangé par la pondération moyenne par échantillon. Le variogramme m.p.e., quant à lui, est moins fluctuant que la version non pondérée et se rapproche de $C(0) - C(h)$.



(a) Variable brute (b) Logarithme translaté

FIG. 4.7 – Site Y - Variogramme classique et déduit de la covariance non centrée pour la variable brute Nap sur les sondages (a) et son logarithme translaté (b). Calculs pour la maille à 10 m et les croix de sondages à 5 m.

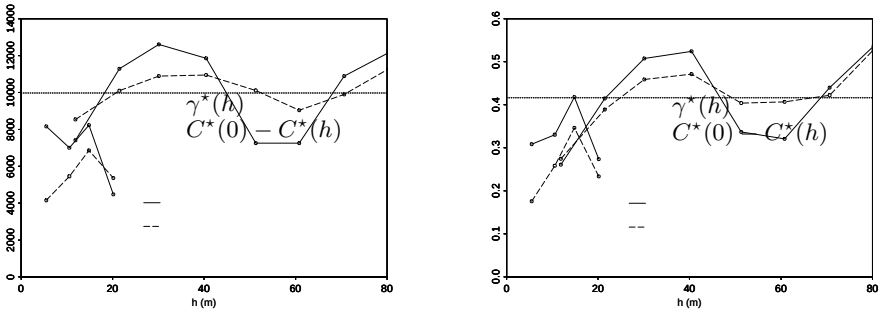
Sans pour autant nécessiter d'hypothèse de stationnarité plus forte que pour le variogramme, le variogramme m.p.e. semble donc posséder les qualités de robustesse de $C(0) - C(h)$.

4.1.4 Synthèse

Le variogramme classique de la teneur a permis de déceler l'existence de structures sur les HAP suivis. Cependant, le peu de robustesse de cet outil a conduit à envisager des calculs complémentaires. Tout d'abord, le calcul pour des transformées de la variable brute permet de confirmer l'existence d'une portée. Leur recours nécessite cependant certaines hypothèses si l'on veut pouvoir revenir en variable brute.

La validation de l'hypothèse de stationnarité d'ordre deux est délicate, seules les distances les plus grandes, peu informées, la mettant en cause. Le choix d'un modèle stationnaire ou non n'influencera que peu l'estimation linéaire des concentrations, qui dépend surtout du comportement de la variable aux petites distances. Il sera par contre nécessaire de trancher pour les méthodes non linéaires du chapitre 6, où déjà l'anamorphose nécessite une hypothèse de stationnarité.

Bien qu'il soit intéressant dans le cas de variables fort contrastées, nous écartons $C(0) - C(h)$ qui requiert la stationnarité d'ordre deux. La pertinence de la pondération par échantillon est intimement liée à la répartition des valeurs faibles et fortes dans les zones sur-échantillonnées et sous-échantillonnées, et le gain en robustesse est limité ; par ailleurs, le choix des pondérateurs est arbitraire. L'utilisation du variogramme moyen par échantillon semble appropriée à nos variables. Par rapport au variogramme classique, son comportement est plus similaire à celui de $C(0) -$



(a) Variable brute (b) Logarithme translaté

FIG. 4.8 – Site Y - Variogramme classique et $C(0)-C(h)$ pour la variable brute Bap sur les sondages (a) et son logarithme translaté (b).

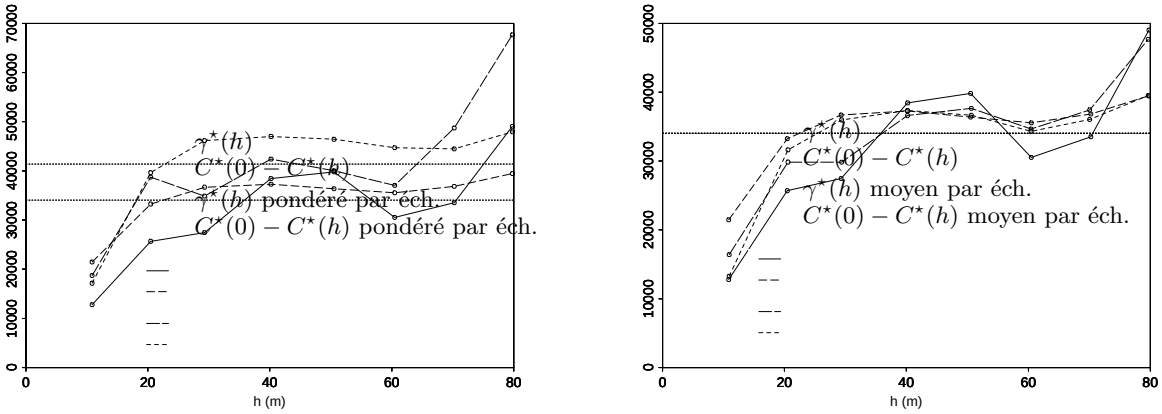
$C(h)$, qui s'est avéré sous réserve de stationnarité plus efficace que le variogramme classique pour des variables fortement contrastées. En outre, le variogramme m.p.e. ne nécessite pas l'hypothèse de stationnarité d'ordre deux, et nous privilégierons donc cet outil. Ce choix présente également l'avantage du calcul sur la teneur, sans qu'il soit nécessaire de recourir à des transformées.

4.2 Application

Afin de simplifier la comparaison graphique, le variogramme m.p.e. de chacun des 16 HAP a été renormé par la variance du HAP concerné. Comme le laissait présager le très bon niveau de corrélation entre les HAP, les structures des différents HAP sur les fosses (voir figure 4.11) sont fort proches, avec une très forte variabilité systématique à petite distance, de l'ordre de 50 % de la variance.

Quelques différences existent : les HAP les plus légers, notamment Nap et Ace, semblent présenter une non stationnarité à l'échelle du champ, qui est à rapprocher de leurs propriétés de volatilité et de solubilité : en fonction de l'hétérogénéité du sol, ils peuvent avoir localement migré. Les autres HAP à 3 cycles, Fle, Ant et Phe, présentent une portée d'environ 20 m. Le point à 80 m des variogrammes expérimentaux peut susciter un questionnement : artefact de calcul, ou indication d'une non stationnarité ? Comme cela a été dit, pour l'estimation linéaire des concentrations cela n'a que peu d'importance. Fle et Phe présentent les variabilités à petite distance les plus faibles.

A partir de 4 cycles, toutes les structures deviennent analogues : une structure de portée égale



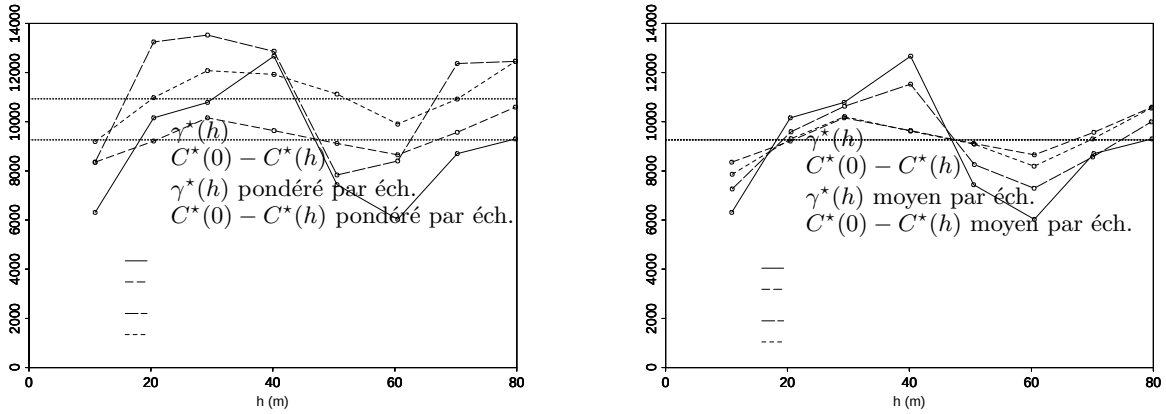
(a) (b)

FIG. 4.9 – Site Y - Variogramme classique et $C(0) - C(h)$ pour Nap sur les sondages : non pondérés et pondérés par échantillon (a), non pondérés et moyens par échantillon (b).

à 20 m, et un décrochement vers 50-60 m. Ce décrochement est expliqué à la figure 4.12 pour le point à 50 m : tandis que pour les HAP à 3 cycles tels que le Phe, on observe plusieurs valeurs fortes au Sud, seule la fosse 8 ressort pour les HAP lourds. Par conséquent, pour ces derniers et par exemple Bap, le nombre de couples incluant une valeur forte est plus faible que pour Phe.

Dans le cas des sondages (voir figure 4.13), la non stationnarité concerne à nouveau Nap, qui ne présente par ailleurs quasiment pas d'effet de pépite, et dans une moindre mesure Fle. La similarité entre les structures des autres HAP est ici encore de mise, mais on observe une portée plus grande que sur les fosses, de l'ordre de 40 m. A quelques exceptions près, telles que Cry ou Bgh, les HAP présentent une meilleure structuration sur les sondages que sur les fosses. Cela peut être rapproché de la présence de remblais et matériaux rapportés, qui concerne plusieurs fosses, tandis que sur les sondages un sol plus majoritairement en place a été observé.

En conclusion, le variogramme m.p.e a mis en évidence des comportements différenciés entre HAP légers d'une part, intermédiaires et lourds d'autre part. Par rapport aux HAP les plus légers, notamment le naphtalène, qui présentent une structure à grande échelle - voire une non stationnarité à l'échelle du champ -, les structures des autres HAP atteignent un palier pour des distances plus courtes. Certaines différences apparaissent également entre fosses et sondages : pour l'ensemble des



(a) (b)

FIG. 4.10 – Site Y - Variogramme classique et $C(0) - C(h)$ pour Bap sur les sondages : non pondérés et pondérés par échantillon (a), non pondérés et moyens par échantillon (b).

HAP, la variabilité à petite distance, non expliquée par l'échantillonnage, est plus faible sur les sondages. Les portées observées pour les HAP à partir de 4 cycles, de l'ordre de 40 m, sont deux fois plus élevées que sur les fosses. Un sol majoritairement en place sur les sondages peut expliquer ces observations.

4.3 Modèles multivariables

Outre la concentration du HAP à estimer, nous connaissons les concentrations des autres HAP analysés et plusieurs variables qualitatives liées aux concentrations. De nombreux modèles multivariables permettent de prendre en compte ces variables corrélées afin d'améliorer l'estimation. Nous présentons leurs propriétés et discutons leur pertinence dans notre contexte.

4.3.1 Corrélations entre HAP

Le modèle linéaire de corégionalisation permet d'assurer un ajustement cohérent des structures simples et croisées. Les variables s'y décomposent linéairement en composantes élémentaires et leurs structures simples et croisées sont donc des combinaisons linéaires de schémas élémentaires. Un premier modèle envisageable, impliquant les 16 HAP, a été évité car il mélange des variables aux propriétés différentes.

4.3.1.1 HAP de même nombre de cycles

Les corrélations entre HAP de poids proche étant très bonnes, un modèle linéaire de corégionalisation entre les HAP de même nombre de cycles est illustré pour les HAP à 5 cycles : Bap, Bbf, Bkf et Dba[3]. Dba étant faiblement présent sur le site, avec 73 % de données inférieures au seuil de détection et une moyenne de 7.06 ppm, il n'a pas été utilisé lors de l'estimation. Les structures croisées se confondent avec l'enveloppe de corrélation maximale - qui correspond au cas d'une corrélation "parfaite", par exemple lorsque les variables sont arithmétiquement liées (voir figure 4.14). Dans ce cas, le cokrigeage apporte peu par rapport au krigeage s'il y a isotopie, *i.e.* si les variables sont informées aux mêmes points.

Lorsque les structures simples et croisées sont toutes proportionnelles et qu'il y a isotopie, le cokrigeage est égal au krigeage. Ce modèle de *corrélation intrinsèque* n'est pas adapté ici, les effets de pépite relatifs différant selon les variables.

4.3.1.2 Composantes de l'ACP

La décomposition, dans le modèle linéaire de corégionalisation, de p teneurs $(Z_i)_{i=1,\dots,p}$ en p composantes élémentaires associées à des échelles différentes du phénomène est possible par *analyse krigeante*. Une analyse en composantes principales (ACP) menée sur les 16 HAP permet de condenser leur information en un nombre réduit de composantes orthogonales. Ces composantes, non corrélées point à point, peuvent néanmoins être spatialement corrélées. Il est possible de contourner ce problème en menant l'ACP sur chacune des composantes associées aux différentes échelles du phénomène, plutôt que sur les teneurs [Wackernagel (1995)].

La non corrélation spatiale des composantes issues de l'ACP est acceptable dans notre cas.

[3]Le comportement de Bbf étant très proche de celui de Bkf, nous aurions pu nous contenter d'utiliser la somme des deux, sachant que leur différenciation pose souvent des problèmes analytiques. Cela conduit à des résultats similaires.

L'essentiel de l'information structurale, contenue dans la première composante, est analogue à celle des HAP les plus présents tels que le fluoranthène, ce qui rend l'utilisation de l'ACP peu probante.

4.3.1.3 Modèle entre fosses et sondages

Finalement, la concentration de chaque HAP étant disponible à la fois sur les fosses et sur les sondages, un modèle linéaire de corégionalisation est envisagé entre ces concentrations, pour tenir compte de leur corrélation (voir figure 4.15).

On ne note pas de diminution de l'effet de pépite relatif sur le variogramme croisé, qui est en outre proche de l'enveloppe de corrélation maximale.

4.3.2 Utilisation des informations qualitatives

4.3.2.1 Indices qualitatifs

Certains indices qualitatifs tels que le goudron sur les sondages discriminent les concentrations fortes (voir chapitre 3). Quel sens donner à un modèle linéaire de corégionalisation entre la concentration en Bap et une indicatrice du goudron ? La meilleure corrélation point à point entre une variable numérique X et une variable qualitative à k modalités s'obtient en attribuant à chacune de ces modalités la moyenne des valeurs de X associées [Saporta (1990)]. On envisage donc un modèle linéaire de corégionalisation entre Bap et la moyenne $M(x)$ de Bap selon qu'il y a présence ou absence de goudron (voir figure 4.16). Cette solution est directement généralisable à une variable qualitative à plus de 2 modalités.

$Z(x)$ peut donc être considérée comme la "dispersée" de sa moyenne $M(x)$ selon que l'indicatrice vaut 0 ou 1. Par exemple, si G est l'indicatrice de la présence de goudron, $M(x)$ peut s'écrire

$$M(x) = \underbrace{\mathrm{E}[Z(x)|G(x)=1]}_{=a}\, G(x) + \underbrace{\mathrm{E}[Z(x)|G(x)=0]}_{=b}\,(1-G(x))$$

a et b étant indépendants de x. Le variogramme de M est alors lié à celui de l'indicatrice G par la relation :

$$\gamma_M(h) = (a-b)^2\,\gamma_G(h)$$

On montre que la relation de Cartier $\mathrm{E}[Z(x)|M(x)] = M(x)$ est vérifiée, ainsi que les relations $\mathrm{E}[Z(x)] = \mathrm{E}[M(x)]$ et $\mathrm{Var}[Z(x)] = \mathrm{Var}[M(x)] + \mathrm{Var}[Z(x) - M(x)]$.

La structure croisée ne présente plus d'effet de pépite, et l'absence de corrélation entre les variabilités à petite distance de la concentration en Bap et des indices qualitatifs gagne donc à être exploitée.

Il est possible d'aller plus loin en tenant compte des résidus $Z(x) - M(x)$. $M(x)$ et $Z(x) - M(x)$ sont non corrélés point à point[4], mais peuvent être corrélés spatialement. En cas de non corrélation

[4] La non corrélation point à point entre valeur probable et résidu demande à être vérifiée, particulièrement en présence de plus de 2 modalités lorsque la régression est ajustée empiriquement. Chautru (1989) illustre cela dans le cas de nodules polymétalliques, dont la concentration sur le fond marin est liée à la pente topographique.

spatiale entre $M(x)$ et le résidu $Z(x) - M(x)$, on montre que $\gamma_Z(h) = \gamma_M(h) + \gamma_{Z-M}(h)$ et que les variogrammes croisés γ_{ZM} et $\gamma_{Z(Z-M)}$ sont proportionnels à γ_M et γ_{Z-M} respectivement. Ces trois critères, qui permettent de tester expérimentalement la validité de ce modèle, ne sont pas vérifiés ici. Le modèle[5], s'il est valide, permet d'obtenir le cokrigeage de Z avec M directement à partir des krigeages de M et $Z - M$.

4.3.2.2 Facteur auxiliaire issu de l'analyse des correspondances

Lorsque plusieurs variables qualitatives sont disponibles, une analyse des correspondances permet de dégager des facteurs éventuellement liés aux concentrations, comme par exemple le facteur auxiliaire distinguant sol remanié et sol en place. La structure croisée du modèle de la figure 4.17 résultant de son utilisation ne présente pas d'effet de pépite.

Un modèle multiplicatif utilisant la régression $\mathrm{E}[Z(x)|F(x)]$ entre la concentration $Z(x)$ et le facteur auxiliaire $F(x)$ est envisageable :

$$Z(x) = \mathrm{E}[Z(x)|F(x)]\, Y(x)$$

où $Y(x)$ est une variable indépendante, qui exprime les variations de Z non expliquées par le facteur auxiliaire [Rivoirard & Guiblin (1996)]. Un tel modèle signifie que l'écart-type de Z est proportionnel à la régression, ce qui est acceptable dans notre cas.

4.3.3 Synthèse

Pour tenir compte des corrélations entre HAP, un modèle linéaire de corégionalisation entre les HAP de même nombre de cycles est intéressant. Les variables qualitatives, dont les variabilités à petite distance ne sont pas corrélées à celles des concentrations et qui peuvent être combinées par analyse des correspondances, seront exploitées par un modèle linéaire de corégionalisation.

[5] de Fouquet & Mandallaz (1992) présentent une application de ce modèle en inventaire forestier.

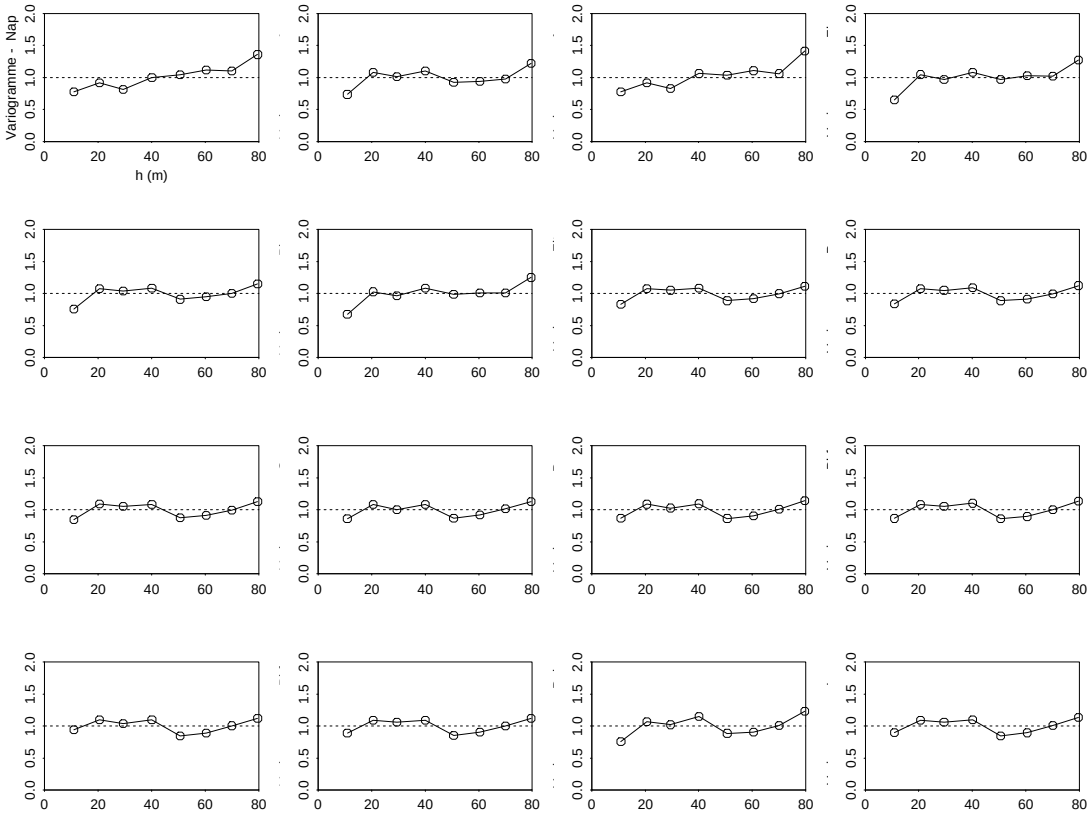


FIG. 4.11 – Site Y - Variogrammes moyens par échantillon des 16 HAP sur les **fosses**.

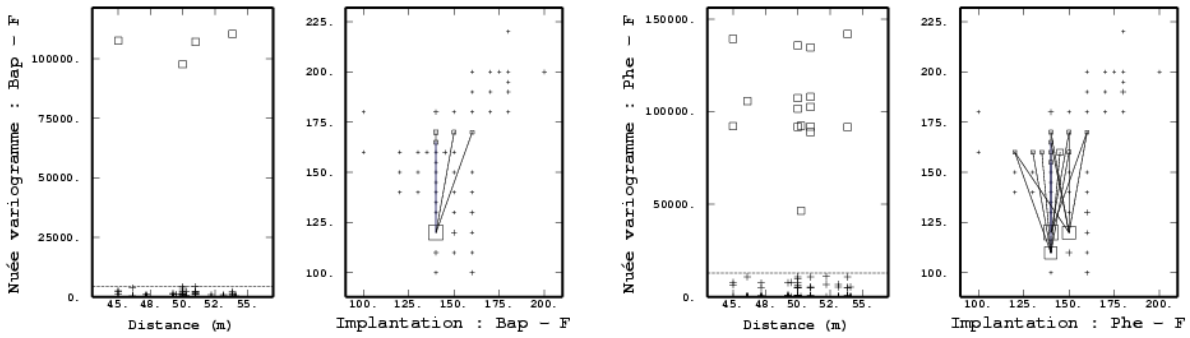


FIG. 4.12 – Site Y - Représentation des couples ayant une forte contribution à la nuée variographique entre 45 et 55 m, pour Bap et Phe.

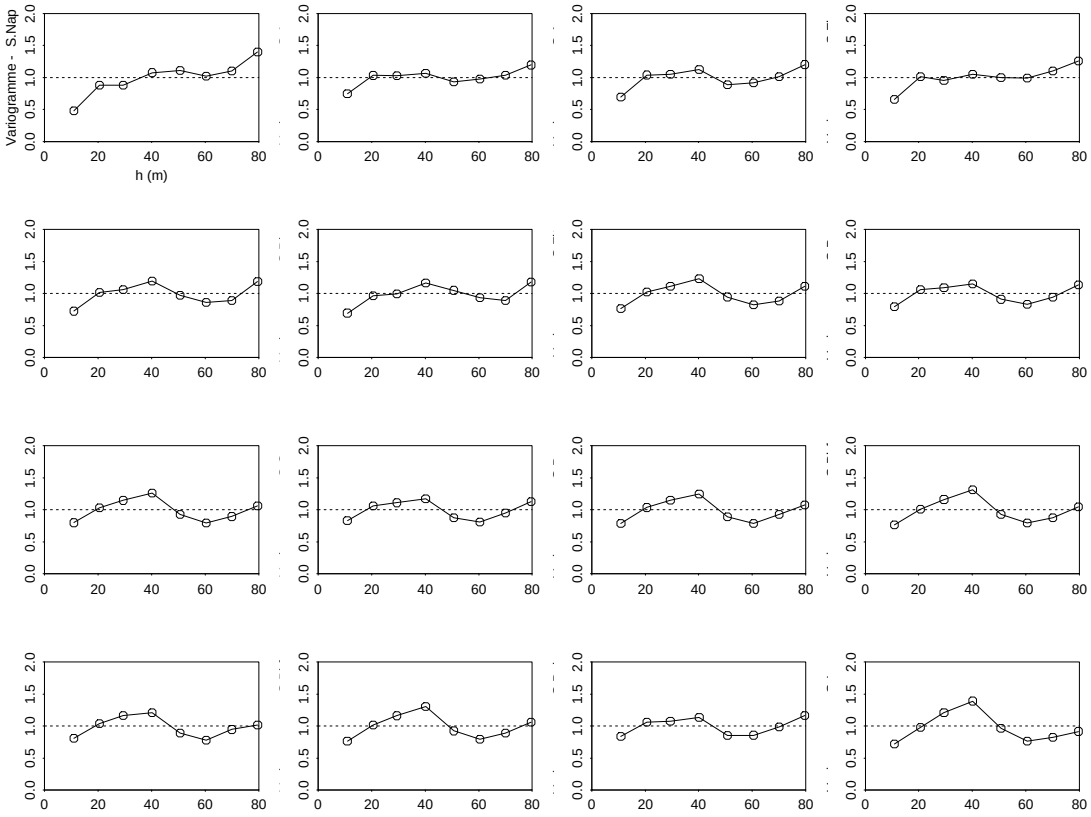


FIG. 4.13 – Site Y - Variogrammes moyens par échantillon des 16 HAP sur les **sondages**.

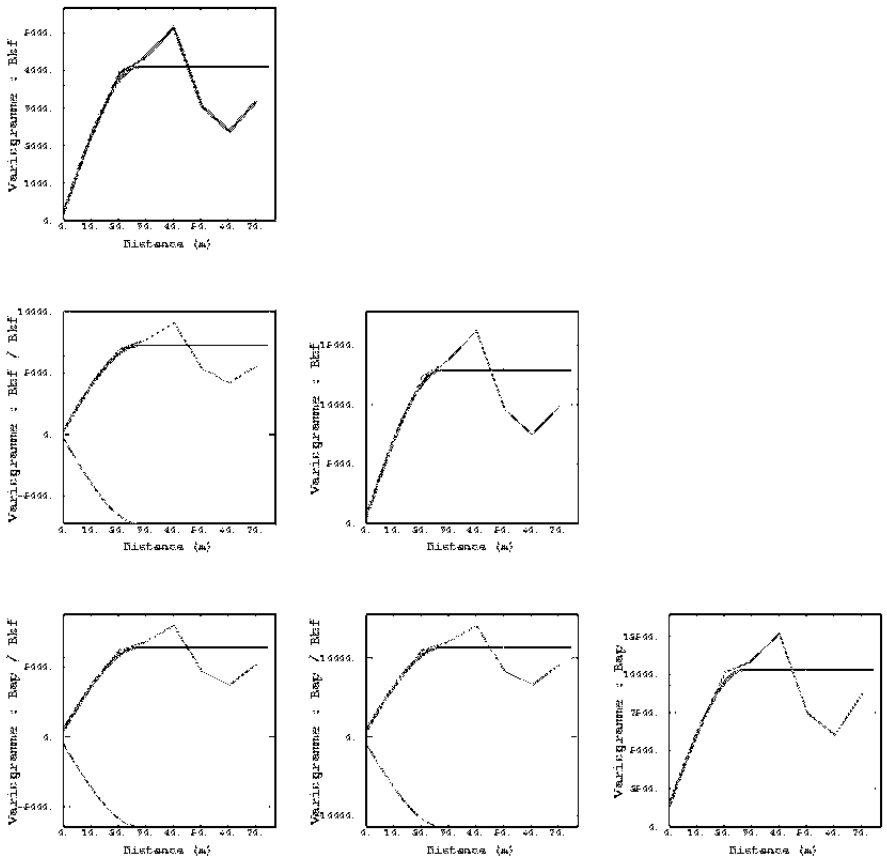


FIG. 4.14 – Site Y - Modèle linéaire de corégionalisation pour Bap, Bbf et Bkf (5 cycles).

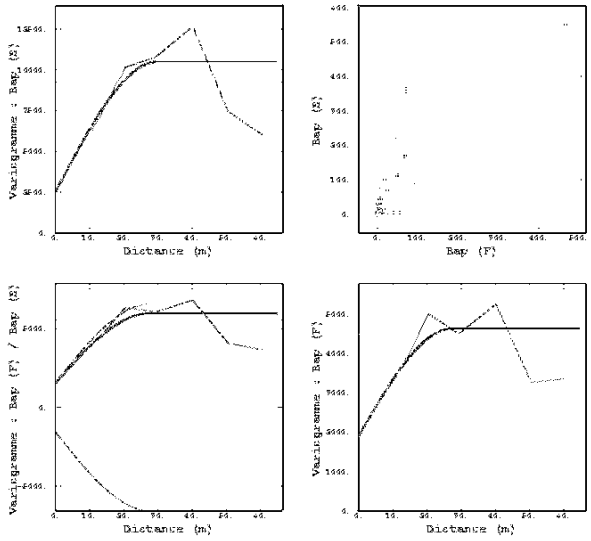


FIG. 4.15 – Site Y - Modèle linéaire de corégionalisation entre Bap sur les fosses (F) et sur les sondages (S). Nuage de corrélation de la concentration en Bap entre fosses et sondages.

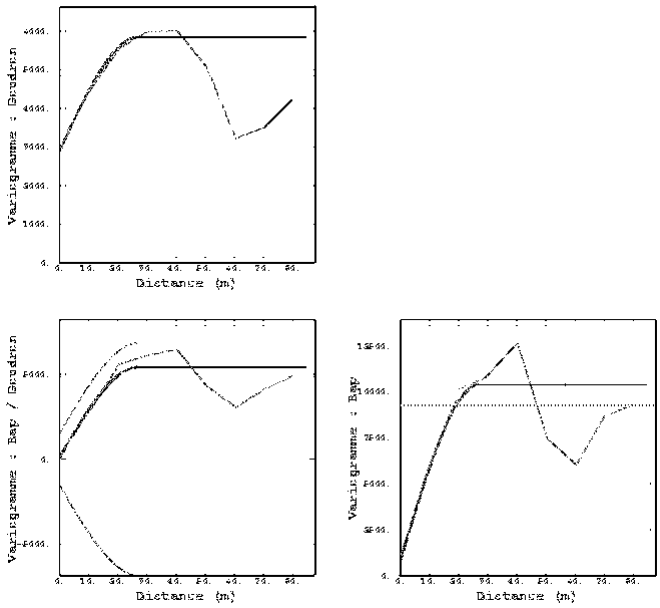


FIG. 4.16 – Site Y - Modèle ajusté entre Bap et la moyenne de Bap par classe de goudron.

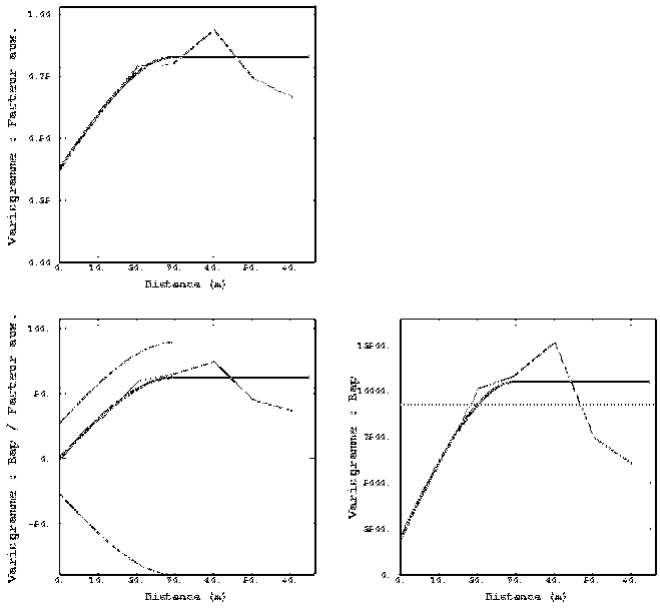


FIG. 4.17 – Site Y - Modèle ajusté entre Bap et le premier facteur auxiliaire.

Chapitre 5

Estimation des concentrations en place

Sommaire

Ce chapitre présente les méthodes d'estimation mises en œuvre pour les concentrations en HAP. L'estimation globale et l'apport du formalisme transitif sont discutés, mais l'accent est essentiellement placé sur l'estimation locale intrinsèque des HAP, menée en monovariable puis en multivariable. On montre comment les variables auxiliaires corrélées aux HAP peuvent améliorer leur estimation.

5.1 Estimation globale

Nous donnons ici un niveau de concentration moyen d'un polluant sur le site et une variance d'estimation, qui indique la précision de l'estimation ; nous nous intéressons donc à la totalité du champ, en utilisant toutes les données disponibles. Considérons la concentration en Bap sur les sondages. La statistique classique, en supposant l'absence de corrélation entre les différentes concentrations, fournit pour la concentration moyenne 37.83 ppm une variance d'estimation σ^2/n qui vaut 181.65, soit un coefficient de variation de 35.6 %. Prendre en compte la structure spatiale de ce HAP permet d'affiner ce résultat.

5.1.1 Hypothèse transitive

La *géostatistique transitive*, dont l'objectif est l'étude directe de la variable régionalisée hors de tout contexte probabiliste, a été développée très tôt par Matheron (1965). La variable régionalisée est supposée nulle en dehors de son champ d'étude D. Etudiant une pollution de sol qui n'existe qu'à l'intérieur des limites d'un site et ne migre pas, cette approche semble pertinente. La théorie

transitive est basée sur l'étude du *covariogramme transitif* $g(h)$, qui est le pendant de la covariance $C(h)$ d'un modèle de fonction aléatoire. En outre, les deux outils sont reliés par la relation

$$g(h) = |D \cap D_{-h}|\ C_R(h)$$

où $C_R(h)$ est la covariance régionale. Le terme $|D \cap D_{-h}|$ fait intervenir la géométrie du champ. Le covariogramme transitif de Bap sur les sondages montre un effet de pépite important et une structure dépendant de la direction (voir figure 5.1) ; nous ne pouvons en effet ici faire abstraction de l'anisotropie due à la géométrie du champ. Le pic du covariogramme dans la direction Nord-Sud pour des distances comprises entre 50 et 70 m est dû aux deux groupes de fortes concentrations.

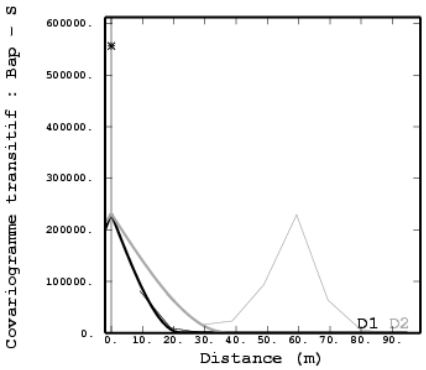


FIG. 5.1 – Site Y - Bap sur les sondages : covariogramme transitif, avec D1 la direction Est-Ouest, D2 la direction Nord-Sud. Traits fins (resp. épais) : covariogrammes expérimentaux (resp. modélisés).

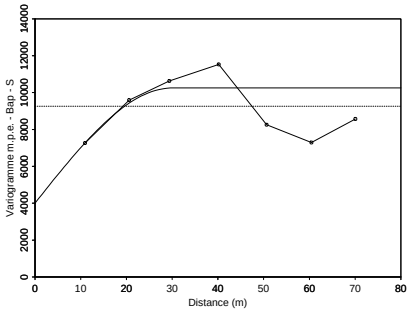


FIG. 5.2 – Site Y - Modèle structural ajusté sur le variogramme m.p.e. de la concentration en Bap sur les sondages.

Dans le cas d'échantillons implantés sur une grille régulière de maille a - ce qui est le cas de 44 données, pour une maille égale à 10 m -, la variance d'estimation de l'abondance totale s'écrit [Matheron (1970)]

$$\sigma^2_{ET} = a^2 \sum_{k_1=-\infty}^{+\infty} \sum_{k_2=-\infty}^{+\infty} g(k_1 a, k_2 a) \ - \int\int g(h)\, dh \tag{5.1}$$

Pour Bap sur les sondages, cette variance d'estimation globale vaut $2.8\,10^9$. La concentration moyenne en Bap étant de 37.83, ayant l'abondance totale $Q^\star = N\,.a^2\,.m$ il en découle un coefficient de variation σ_{ET}/Q de 31.8 %.

Ce formalisme ne permet pas la prise en compte des resserrements à 5 m des croix de sondages[1].

[1]Bez (1997) présente une méthode de pondération par les surfaces d'influence palliant cet inconvénient ; elle

5.1.2 Hypothèse intrinsèque

S'il est possible de faire la part entre géométrie du champ et régionalisation de la variable d'intérêt, et par conséquent de modéliser la structure spatiale de la fonction aléatoire, alors le cadre intrinsèque permet le calcul d'une variance d'estimation sur le champ V de la moyenne des concentrations $(Z_i)_{i=1,\dots,N}$, qui s'exprime comme [Matheron (1970), Petitgas & Rivoirard (1991)]

$$\sigma^2_{EI} = \overline{C}_{VV} + \overline{C}_{ij} - 2\,\overline{C}_{iV} \tag{5.2}$$

où $\overline{C}_{VV}$, $\overline{C}_{ij}$ et $\overline{C}_{iV}$ sont respectivement les covariances moyennes entre deux points décrivant indépendamment V, entre la donnée i et un point décrivant V et entre deux données i et j pour tous les couples (i, j). Par rapport à la formule transitive, il y a apparition du terme $\overline{C}_{iV}$ qui représente l'influence sur l'estimation de la géométrie des données, ici fixes, par rapport à celle du champ ; la grille n'est donc plus nécessairement régulière. Le modèle de variogramme $\gamma(h) = 4000 + 6200 \times \mathrm{Sph}(30\mathrm{m})$ illustré à la figure 5.2 conduit à un coefficient de variation pour l'estimation globale de 31.3 %, la concentration moyenne sur le champ étant estimée par krigeage à 38.2 ppm.

Les coefficients de variation et les variances d'estimation, comparables et cohérents pour les deux formalismes, sont inférieures au résultat de la statistique classique. La portée de la structure spatiale, grande par rapport à la taille du champ, explique cela. Ces variances indiquent la précision de l'estimation d'une concentration moyenne sur l'ensemble du champ. Dans le cas présent, cette connaissance du niveau de concentration moyen en polluant présente un intérêt pratique assez limité, et il est plus important de procéder à une estimation locale des concentrations en place lorsque ces dernières présentent une structuration spatiale.

5.2 Estimation locale monovariable des concentrations

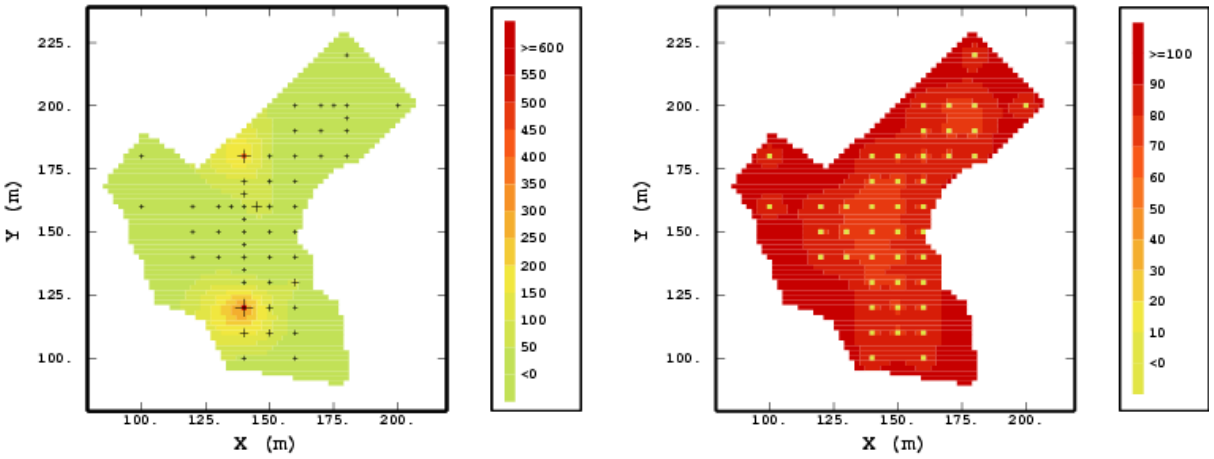
Les estimations sont réalisées ponctuellement sur une grille de 2 m sur 2 m dont les contours correspondent à ceux du site. Vu le peu de données disponibles, les estimations sont effectuées en voisinage unique. Nous déterminons les méthodes d'estimation les plus appropriées en analysant le cas du benzo(a)pyrène sur les sondages, avant de les appliquer aux autres HAP.

Deux critères permettront de comparer l'efficacité des différents modèles :
- une validation croisée ; l'erreur quadratique moyenne, donnée systématiquement, synthétise moyenne et variance des écarts entre les valeurs réelle et estimée[2]. Bien qu'étant conscient que l'utilisation d'une telle statistique est délicate vu le peu de données et la dissymétrie des variables, les conditions identiques des différentes estimations en font un critère expérimental de comparaison intéressant.
- les variances d'estimation associées aux estimations, pour chaque modèle.

n'a pas été approfondie ici, la prise en compte des irrégularités ne jouant que faiblement sur l'effet de pépite C_0 du covariogramme transitif, qui est le facteur déterminant dans le calcul de la variance d'estimation. En effet, celle-ci, qui est la différence entre l'intégrale du covariogramme et son approximation au pas de la maille, peut être correctement approchée par le produit $C_0\,.a^2$, cette approximation donnant dans le cas présent un coefficient de variation de 30 %.

[2] En effet, si Z est estimé par $Z^\star$, l'erreur quadratique moyenne $\mathrm{E}[(Z - Z^\star)^2]$ vaut $\mathrm{Var}[Z - Z^\star] + \mathrm{E}^2[Z - Z^\star]$.

Le modèle structural $\gamma(h) = 4000 + 6200 \times \mathrm{Sph}(30\ \mathrm{m})$ de la figure 5.2, ajusté sur le variogramme m.p.e., est utilisé pour le krigeage ordinaire (KO) de Bap sur les sondages. Les deux zones de valeurs fortes, au Sud et au Nord, ressortent sur la carte d'estimation (voir figure 5.3). La tache située au Nord se prolonge à l'Est de la croix centrale. Les écart-types d'estimation, nuls aux points expérimentaux où l'interpolation est exacte, deviennent très élevés dès que l'on s'en écarte, à cause de l'effet de pépite[3]. Ils montrent par ailleurs bien le risque que présente l'utilisation de cette estimation dès que l'on sort de la zone investiguée.



(a) Krigeage ordinaire (b) Ecart-types d'estimation

FIG. 5.3 – Site Y - Krigeage ordinaire de Bap sur les sondages, modèle ajusté sur le variogramme m.p.e.

Il est possible d'attribuer une part de l'effet de pépite à la variance de l'erreur de mesure et de la filtrer. Nous reviendrons sur ce point au chapitre 8 pour le second site. Le nuage de corrélation entre valeurs réelles aux points expérimentaux et estimées par validation croisée illustre l'effet de lissage du krigeage, les valeurs estimées étant nettement moins dispersées que les valeurs réelles (voir figure 5.4). L'erreur quadratique moyenne obtenue par validation croisée, égale à 8532, servira de référence dans la suite ; cette erreur est légèrement supérieure pour un modèle ajusté automatiquement sur le variogramme classique. Les deux concentrations les plus fortes sont mal ré-estimées.

Destiné essentiellement à l'estimation globale, le formalisme transitif peut également être utilisé pour l'estimation locale [Bez (1997), Rivoirard (1995)]. Le *krigeage transitif* consiste à estimer $z(x)$ par une combinaison linéaire des données situées dans un voisinage de x. L'optimalité est ici obtenue en choisissant les pondérateurs de la combinaison linéaire de sorte que, s'il était possible de translater partout dans l'espace la configuration formée par x et les points de son voisinage, la somme des erreurs quadratiques entre valeurs vraies et estimées soit minimale.

Le krigeage transitif obtenu à l'aide du covariogramme transitif de la figure 5.1 est illustré à

[3]Il est d'usage de procéder au krigeage en filtrant l'effet de pépite - erreur de mesure ou de positionnement ; cette technique, discutée à la page 144, conduit à des écart-types qui ne sont alors plus nuls aux points expérimentaux [Chilès & Delfiner (1999)]. Cela nécessite cependant une connaissance suffisante de ces erreurs de mesure, ce qui n'est pas le cas ici.

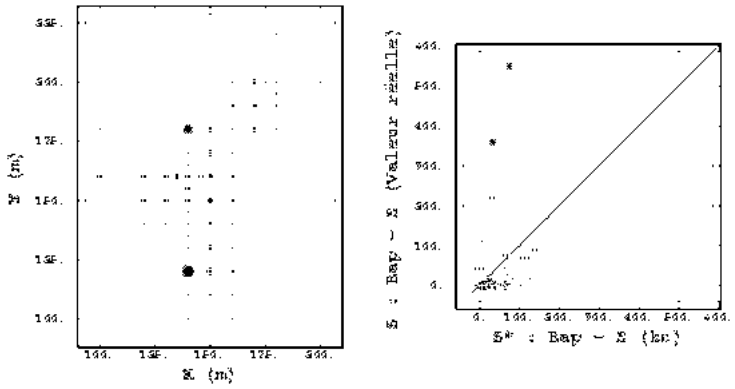


FIG. 5.4 – Site Y - Validation croisée du modèle monovariable ajusté sur le variogramme m.p.e. de Bap sur les sondages. La première bissectrice est représentée. Les points repérés par des ronds correspondent aux valeurs de $\frac{Z-Z^\star}{\sigma^\star}$ supérieures en valeur absolue à 2.5 ; cela permet de détecter les valeurs inférieures (resp. supérieures) au quantile à 5 % (resp. 95 %) d'une distribution normale, correspondant aux concentrations particulièrement mal ré-estimées par validation croisée.

la figure 5.5. L'anisotropie géométrique se retrouve sur la carte d'estimation, sans raison physique l'expliquant. Le nuage de corrélation de la figure 5.6 entre les krigeages transitif et ordinaire souligne bien le lissage plus important du krigeage transitif, dû à l'effet de pépite plus élevé.

Par ailleurs, le krigeage transitif ne permet pas le calcul de variances d'estimation locales, et ne donne donc pas d'indication sur la précision de l'estimation fournie. Ces inconvénients nous ont conduit à abandonner cette voie, qui ne permet par ailleurs pas de modélisation plus complète - par exemple multivariable lorsque les variables ne sont pas définies aux même points.

5.3 Estimation multivariable des concentrations

Les estimations sont réalisées à partir des modèles linéaires de corégionalisation discutés au paragraphe 4.3. Toutes les variables sont informées aux mêmes points - configuration isotopique.

5.3.1 Cokrigeage Fosse-Sondage

Le cokrigeage de la concentration en Bap sur les sondages à partir du modèle de la figure 4.15 n'améliore pas les résultats du krigeage ordinaire. Les variances d'estimation sont quasiment inchangées, exceptée une très légère diminution au voisinage des points expérimentaux, et l'erreur quadratique moyenne obtenue par validation croisée augmente légèrement.

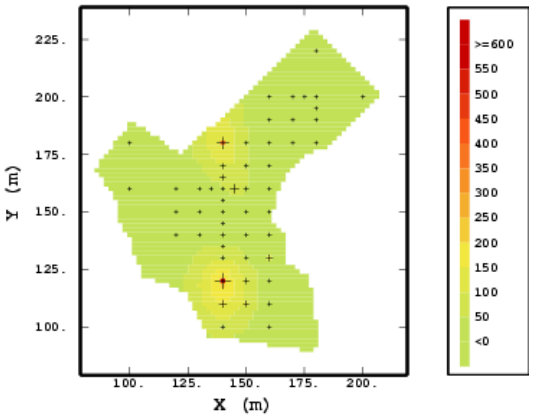


FIG. 5.5 – Site Y - Bap sur les sondages : krigeage transitif de la concentration en Bap.

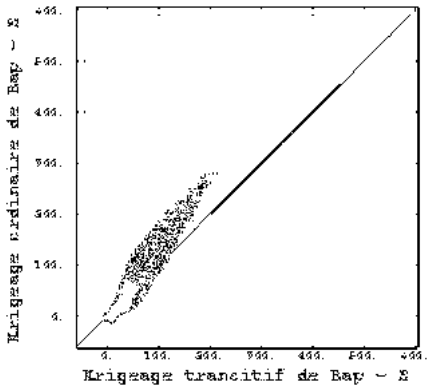


FIG. 5.6 – Site Y - Bap sur les sondages : nuage de corrélation entre krigeage transitif et krigeage ordinaire. La première bissectrice est représentée.

5.3.2 Utilisation des autres HAP

L'estimation de Bap (voir figure 5.7) à l'aide du modèle de corégionalisation de la figure 4.14 présente deux différences essentielles par rapport au krigeage ordinaire de Bap :

- L'estimation de la tache de pollution située au Nord est plus élevée qu'avec le krigeage ordinaire. Cela est dû à Bbf et Bkf, pour lesquels l'importance relative de cette tache est plus élevée que pour Bap.
- L'estimation de la concentration au Sud du site correspond mieux à l'implantation de l'ancienne mare à goudron sud. Par ailleurs, ce cokrigeage fait mieux ressortir les traces de pollution à l'Est du site et une tache de pollution située juste au Sud de la tache de pollution Nord.

En outre, les écart-types d'estimation obtenus par cokrigeage sont nettement plus faibles.

Nous avons vu que Dba est faiblement présent sur le site, avec 73 % de données inférieures au seuil de détection. L'inférence de la structure spatiale peut être encore plus ardue dans ce cas. Bien que peu intéressante en pratique dans le cas présent, vu les faibles concentrations de Dba, son estimation est améliorée en utilisant sa corrélation avec les autres HAP à 5 cycles.

Supposons que nous ne connaissions un HAP qu'en peu de points ; pour fixer les choses, prenons l'exemple de Bap, et supposons que nous ne connaissions sa concentration qu'en 26 points, choisis aléatoirement parmi les 52 (voir figure 5.8).

Une première estimation de Bap, menée uniquement à partir des 26 points, conduit à la disparition de la tache de pollution située au Nord, suite au retrait de la valeur forte qui en était à l'origine (voir figure 5.9(a)). Nous pouvons comparer cette estimation en chacun des 26 points non utilisés avec leurs concentrations réelles, connues (voir figure 5.9(b)). Les points se situent majoritairement en-dessous de la première bissectrice du nuage et ont donc tendance à être surestimée par rapport aux concentrations réelles. Cela provient du voisinage unique et du poids relativement élevé de la

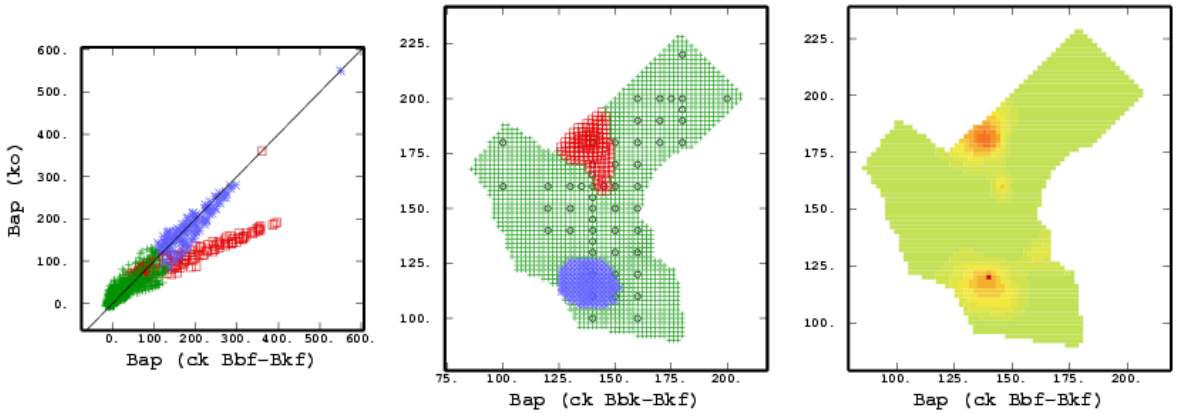


FIG. 5.7 – Site Y - Nuage de corrélation entre le krigeage ordinaire de Bap et son cokrigeage avec Bbf et Bkf. Résultat du cokrigeage pour Bap, sous forme de grille avec repérage de certaines zones puis sous forme de carte. Echelle de couleurs identique à celle du krigeage monovariable.

moyenne, cette dernière étant influencée par les valeurs fortes. On note deux exceptions : la valeur forte située au Nord, complètement sous-estimée, ainsi qu'une autre valeur, d'environ 50 ppm, dont l'estimation est même négative. L'erreur moyenne (voir tableau 5.1) montre la surestimation du krigeage.

Type d'estimation	Erreur moyenne (ppm)	Erreur quad. moyenne
Krigeage Bap	15.28	6253
CK avec Bbf+Bkf	5.85	858

TAB. 5.1 – Site Y - Erreur commise sur les concentrations des 26 sondages de Bap retirés, selon que l'on procède au krigeage de Bap seul ou à son cokrigeage avec Bbf+Bkf.

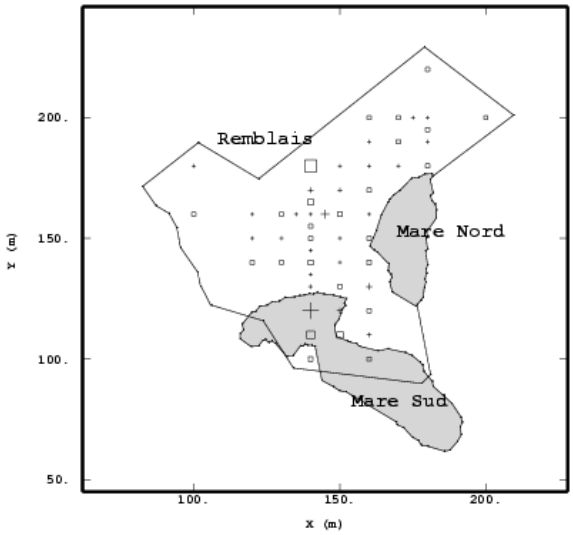
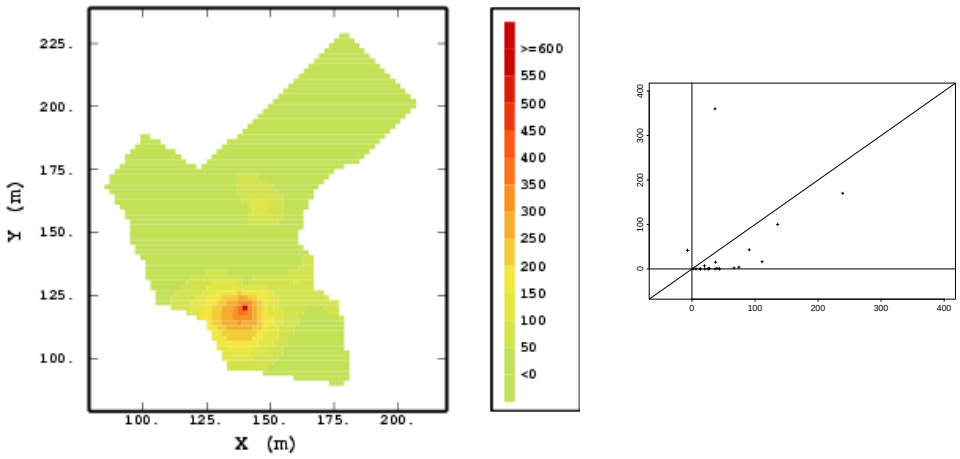


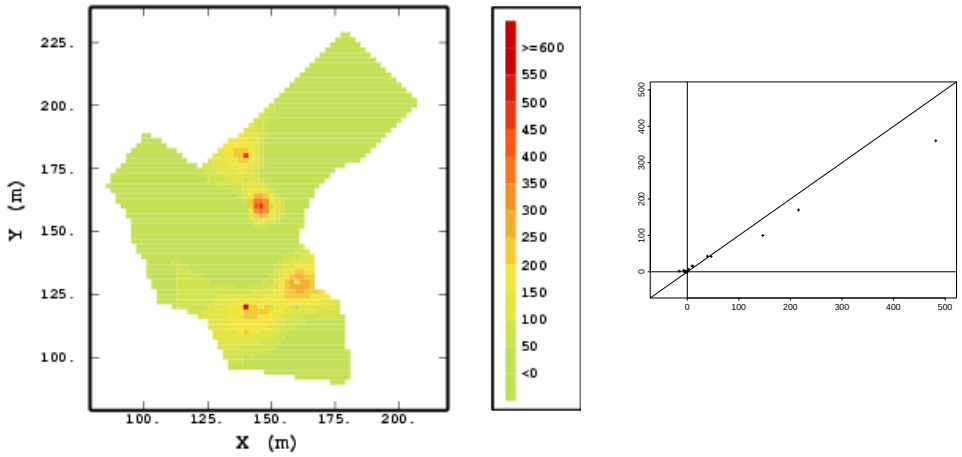
FIG. 5.8 – Site Y - Implantation des données de Bap, informations données par l'historique du site et indication par des carrés des points supposés non connus.



(a) (b)

FIG. 5.9 – Site Y - (a) Krigeage ordinaire de Bap sur la base de la connaissance de 26 points. (b) Nuage de corrélation entre l'estimation en chacun des 26 points non utilisés et leurs concentrations réelles.

D'autre part, nous pouvons effectuer un cokrigeage de Bap avec un HAP corrélé qui serait, quant à lui, connu en chacun des 52 points. Classiquement, l'analyse d'un spectrogramme donne pour un même coût les concentrations de chacun des 16 HAP, et sauf erreur de re-transcription des concentrations, la situation décrite n'a que peu de chance d'arriver. Ce qui n'est ici qu'un exercice peut cependant être transposé à d'autres polluants, par exemple métalliques pour lesquels une utilisation de corrélations entre métaux, en réduisant le nombre d'analyses nécessaires, générerait un gain financier important.



(a) (b)

FIG. 5.10 – Site Y - (a) Cokrigeage de Bap (26 points) avec Bbf+Bkf. (b) Nuage de corrélation entre l'estimation en chacun des 26 points non utilisés et leurs concentrations réelles.

On constate sur l'estimation de Bap par cokrigeage avec la somme de Bbf et Bkf que les deux taches de pollution ressortent à nouveau correctement (voir figure 5.10). D'autres taches plus locales apparaissent également. On note quelques sauts correspondant aux points uniquement informés en Bbf et Bkf. Les variances d'estimation obtenues sont plus faibles que pour le krigeage de Bap en 26 points seuls, particulièrement aux 26 points non utilisés. Le nuage entre les 26 valeurs non utilisées réelles de Bap et leur estimation par cokrigeage est plus satisfaisant que celui correspondant au krigeage (voir figure 5.9(b)). On ne "rate" plus ici de valeur forte, et seuls quelques points de concentration très faible sont estimés négativement. La légère surestimation des valeurs fortes, dont l'explication a été donnée plus haut, est à signaler. Erreur et erreur quadratique moyennes sont nettement plus faibles dans ce cas (voir tableau 5.1).

En conclusion, l'utilisation des corrélations entre HAP est surtout intéressante dans le cas hétérotopique, lorsque l'implantation d'une variable corrélée à la variable d'intérêt diffère de cette dernière et permet d'en préciser l'estimation.

5.3.3 Sommes de HAP et application

Plutôt qu'une estimation de la concentration de chacun des HAP, il est également possible de fournir une estimation de la somme des HAP par nombre de cycles, étant donné leur similarité et leurs caractéristiques structurales très proches[4]. La figure 5.11 donne, pour les sondages, la somme des 4 HAP à 5 cycles cokrigés. Ce cokrigeage fournit une estimation cohérente de la somme des HAP à 5 cycles, dans la mesure où $[\sum_i Z_i]^\star = \sum_i Z_i^\star$, $(Z_i)_{i=1,\dots,4}$ désignant les 4 HAP à 5 cycles[5].

En guise d'application, les estimations des figures 5.12 et 5.13 sont obtenues par cokrigeage des

[4] Une estimation de la concentration totale des 16 HAP est par contre à déconseiller pour les raisons mentionnées lors des chapitres précédents.

[5] L'estimation par krigeage de la somme des HAP à 5 cycles conduit à des résultats proches.

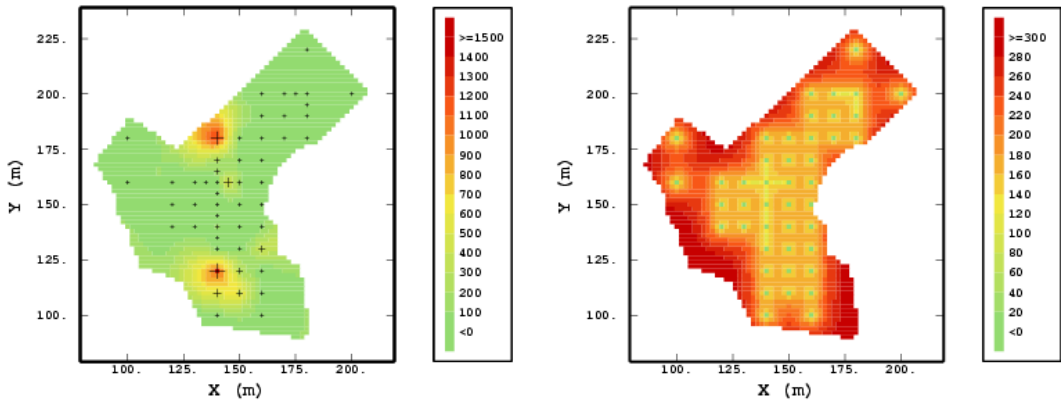


FIG. 5.11 – Site Y - Estimation par cokrigeage de la somme des HAP à 5 cycles sur les sondages, et écart-types d'estimation.

sommes de HAP par cycle. La non stationnarité modélisée pour Nap est visible sur la carte d'estimation. Concernant les HAP de 4, 5 et 6 cycles, les concentrations plus élevées en bordure, dans les zones où il n'y a plus de données, sont dues au poids de la moyenne. Par ailleurs, la zone située au Nord de la zone centrale présente des concentrations de plus en plus fortes, proportionnellement au niveau de concentration des sommes de HAP, lorsque le nombre de cycles des HAP augmente. Il en est de même pour la zone située à l'Est, à proximité de l'ancienne mare à goudron Nord. Ces remarques restent valables pour les sondages. La tache de pollution au Nord de la zone centrale tend à rejoindre la mare Nord, sans que cela semble être un artefact d'estimation. La présence prédominante des HAP à 3 et 4 cycles, déjà constatée lors de l'analyse exploratoire, est visible.

5.4 Utilisation des informations qualitatives

Quel serait l'apport d'une campagne d'échantillonnage dense visant le prélèvement d'informations qualitatives peu coûteuses ? Comment utiliser ces informations ?

Nous étudions tout d'abord l'apport d'une cartographie des informations qualitatives seules, par rapport à celle des concentrations en HAP. Ensuite, nous considérons le cas isotopique où les informations qualitatives sont renseignées aux mêmes points que les concentrations en HAP, avant de les envisager plus nombreuses ; le gain ainsi obtenu par rapport au krigeage de la concentration en HAP seule est discuté.

5.4.1 Krigeage du facteur auxiliaire

Le facteur auxiliaire discriminant sol remanié et sol en place, il est intéressant d'étudier l'apport de la cartographie de ce facteur seul. Le krigeage ordinaire de ce facteur présente une similarité notable avec les zones de fortes concentrations estimées en Bap (voir figure 5.14 (a)). Le nuage de corrélation avec le krigeage ordinaire de la concentration en Bap (voir figure 5.14 (b)) montre que,

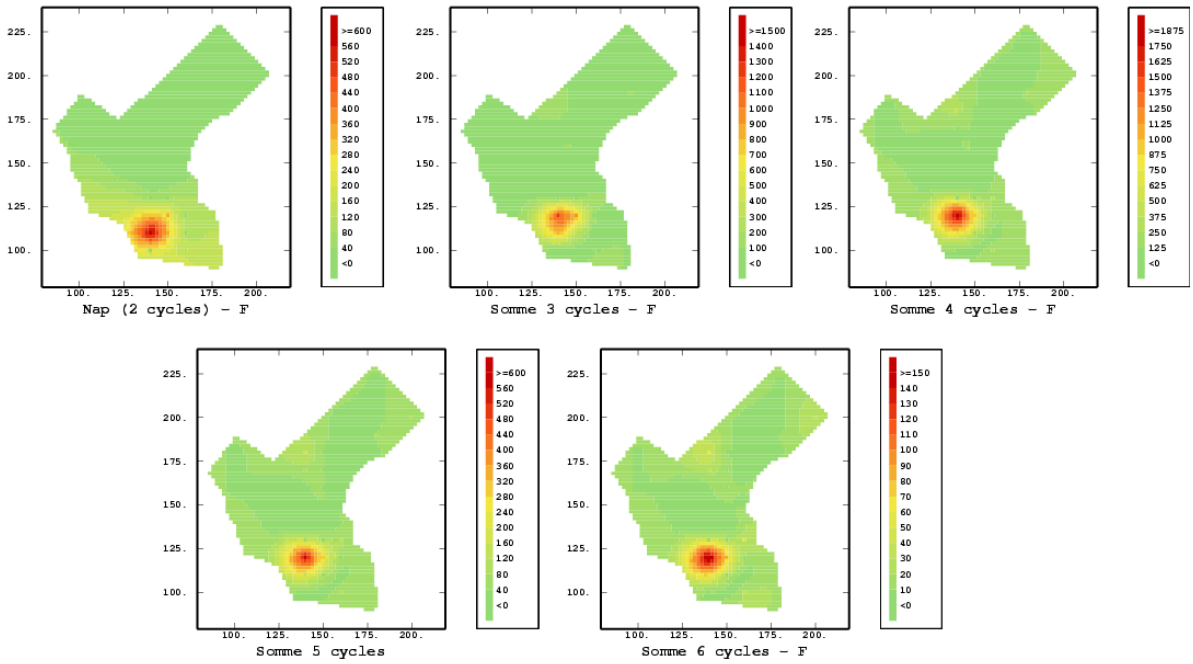


FIG. 5.12 – Site Y - Cokrigeage des sommes de HAP sur les fosses.

bien que cette similarité soit particulièrement marquée aux points expérimentaux, la cartographie du facteur auxiliaire fournit à moindre coût une première idée de la répartition des zones de forte concentration, même s'il existe des concentrations non négligeables situées en dehors.

5.4.2 Cokrigeage en configuration isotopique

Un cokrigeage en modèle linéaire de corégionalisation (voir figure 4.16) entre la concentration en Bap et sa moyenne par classe de goudron (présence/absence) améliore les résultats du krigeage ordinaire, en diminuant sensiblement les variances d'estimation. L'utilisation de cette variable qualitative entraîne par ailleurs des concentrations estimées plus fortes pour les deux taches de pollution principales. L'utilisation par cokrigeage du modèle entre concentration en Bap et le facteur auxiliaire issu de l'analyse des correspondances (voir figure 4.17) améliore plus nettement les résultats du krigeage ordinaire de la concentration en Bap, avec une erreur quadratique moyenne qui passe de 8532 à 8106 et surtout des variances d'estimation inférieures (voir figure 5.15) ; ce résultat est meilleur que ceux obtenus en combinant seulement les concentrations des différents HAP.

Rappelons l'absence d'effet de pépite sur le variogramme croisé entre concentration en Bap et facteur auxiliaire, indicatrice d'une non corrélation entre les erreurs de mesure ou les variabilités à petite distance des deux variables.

Pour l'estimation de Bap, utiliser à la fois les autres HAP à 5 cycles et le facteur auxiliaire conduit à des résultats qualitativement similaires à ceux obtenus par cokrigeage à l'aide du facteur auxiliaire seul. Ces conclusions sont analogues pour les HAP ayant un nombre de cycles différent.

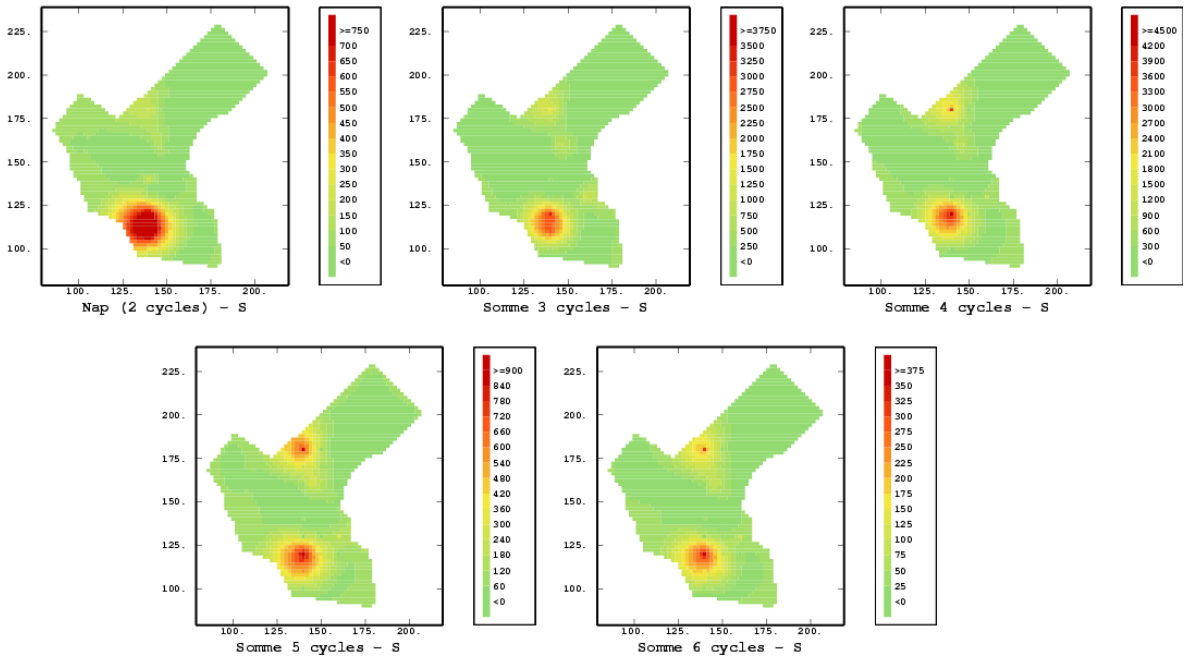


FIG. 5.13 – Site Y - Cokrigeage des sommes de HAP sur les sondages.

5.4.3 Informations qualitatives plus denses

L'estimation des concentrations est-elle améliorée par l'utilisation combinée d'informations qualitatives plus denses car peu coûteuses ? Bien que n'ayant pas à disposition un tel échantillonnage, nous nous sommes approché de ces conditions en ne retenant aléatoirement que la moitié des analyses de Bap sur les sondages, soit les 26 valeurs déjà présentées à la figure 5.8, tout en gardant les 52 données qualitatives.

Dans le cas présent, la structure spatiale des 26 valeurs restantes de Bap reste proche de celle pour les 52 données. En procédant à l'estimation à partir de ces 26 valeurs, nous pouvons évaluer l'erreur commise en chacun des autres points, non utilisés mais de concentrations connues. Ensuite, nous pouvons calculer ce que devient cette erreur si nous ajoutons par cokrigeage l'information apportée par le goudron - sous forme de moyennes de Bap par classes -, ou le facteur auxiliaire, variables supposées disponibles en chacun des 52 points et donc deux fois plus denses que les données de Bap. Les résultats sont repris au tableau 5.2.

Type d'estimation	Erreur moyenne (ppm)	Erreur quad. moyenne
KO Bap	15.28	6253
CK avec goudron	5.93	3309
CK avec facteur	8.66	4980

TAB. 5.2 – Site Y - Erreur commise sur les concentrations des 26 sondages de Bap retirés, selon que l'on procède au krigeage de Bap seul ou à son cokrigeage avec le goudron ou le facteur auxiliaire.

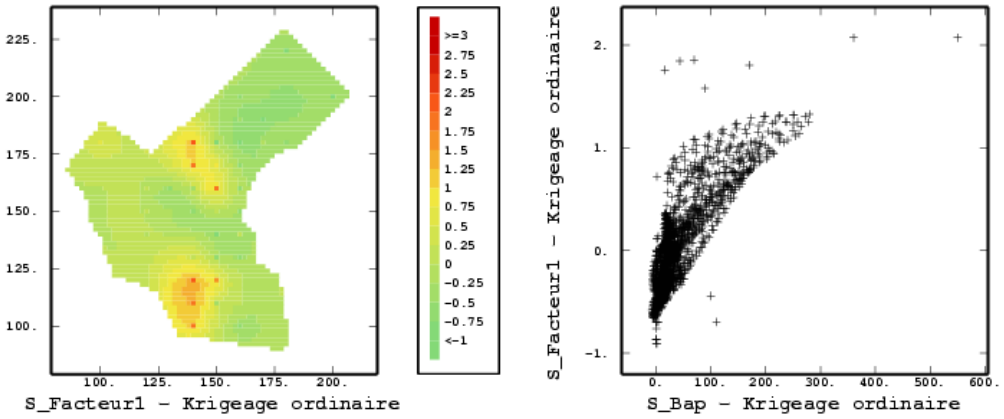


FIG. 5.14 – Site Y - Krigeage ordinaire du facteur auxiliaire (a) et nuage de corrélation avec le krigeage ordinaire de la concentration en Bap (b).

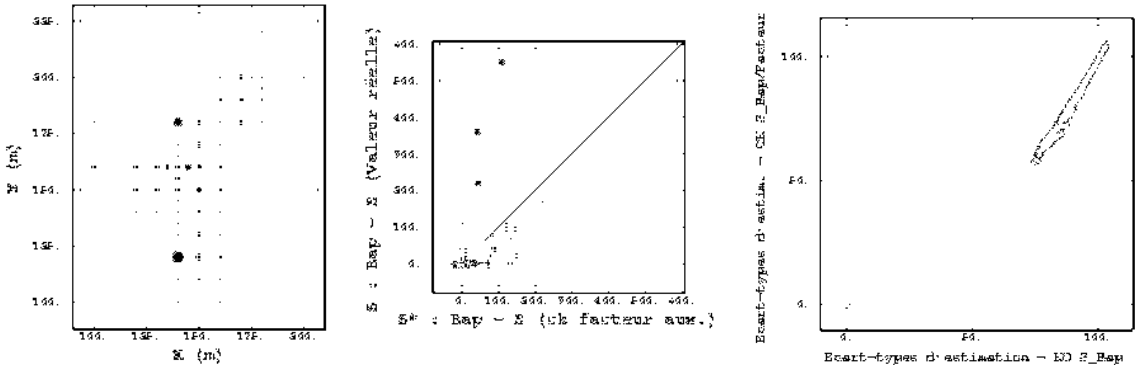


FIG. 5.15 – Site Y - Validation croisée pour Bap. Modèle linéaire de corégionalisation entre Bap et le premier facteur auxiliaire. Les points repérés par des ronds correspondent aux valeurs de $\frac{Z-Z^\star}{\sigma^\star}$ supérieures en valeur absolue à 2.5.

Que ce soit avec le goudron ou le facteur auxiliaire, l'apport du cokrigeage est décisif par rapport au krigeage ordinaire de la concentration seule ; de plus, l'utilisation du goudron conduit à de meilleurs résultats, la diminution de l'erreur moyenne étant de l'ordre de 60%.

Les nuages de la figure 5.16 montrent que le gain obtenu en cokrigeage provient en majeure partie de la concentration la plus élevée de Bap. Néanmoins, en la retirant les erreurs moyennes restent meilleures en cokrigeage. Certaines des concentrations faibles en Bap sont estimées négativement par cokrigeage. Le cokrigeage avec le facteur auxiliaire n'estime cependant négativement que des concentrations très faibles en Bap, tandis que le krigeage de Bap et son cokrigeage avec le goudron estiment négativement une concentration de 40 ppm en Bap. Le cokrigeage avec le goudron évite une surestimation des concentrations faibles par rapport au cokrigeage avec le facteur auxiliaire. Ces résultats ont été validés pour d'autres HAP sur les sondages.

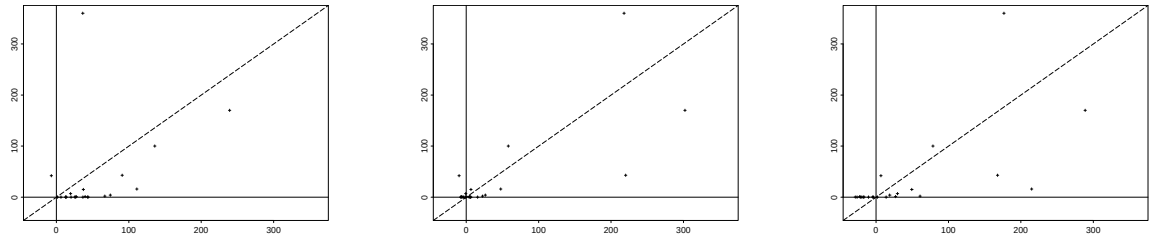


FIG. 5.16 – Site Y - Nuages de corrélation entre les concentrations des sondages retirés de Bap et leur estimation successivement par krigeage ordinaire, cokrigeage avec le goudron et cokrigeage avec le facteur auxiliaire.

5.5 Synthèse

Ne permettant pas le calcul de variances d'estimation locales, l'approche transitive a été abandonnée. Les deux zones de valeurs fortes, au Sud et au Nord, ressortent sur l'estimation par krigeage ordinaire ponctuel de Bap. L'effet de pépite important conduit à des écart-types d'estimation élevés dès que l'on s'écarte des données.

Les HAP étant particulièrement bien corrélés, il est logique d'envisager leur combinaison. Ce niveau de corrélation est tel que dans une configuration isotopique, où les différentes variables sont informées aux mêmes points, le cokrigeage apporte peu d'amélioration par rapport au krigeage monovariable, juste une légère diminution des variances d'estimation. Par contre, ces corrélations et la grande similarité structurale des HAP de poids proche améliorent sensiblement l'estimation d'un HAP en des points où l'on ne connaîtrait pas sa concentration mais bien celle d'un HAP proche ou de la somme des HAP de même nombre de cycles.

Estimer des sommes de HAP de même nombre de cycles plutôt que les HAP individuellement va dans le sens de la recherche d'un résultat synthétique, sans pour autant mélanger des variables présentant des propriétés différentes.

L'estimation des HAP gagne à utiliser des informations qualitatives corrélées, éventuellement sous forme de facteur auxiliaire. Même si les conditions expérimentales ne permettent pas d'aller très loin, l'utilisation d'informations qualitatives plus denses corrélées aux concentrations des HAP améliore systématiquement leur estimation.

Chapitre 6

Zones de dépassement de seuils de pollution

Sommaire

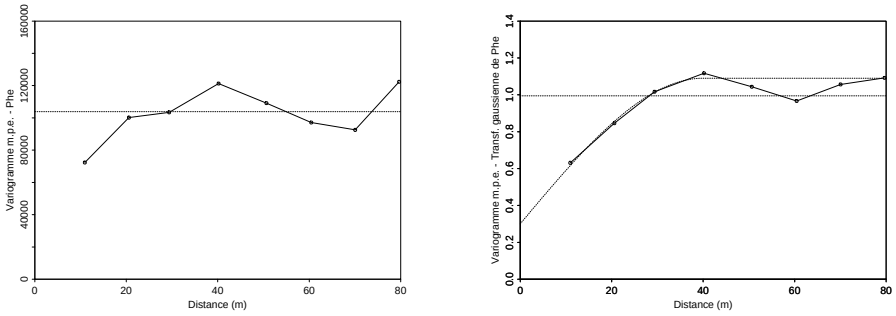
La méthodologie choisie pour le calcul de probabilités de dépassement de seuils de pollution, classique, est présentée à partir d'une variable sans valeur inférieure au seuil de détection ; l'influence de ce seuil est discutée. Des calculs par krigeage disjonctif et espérance conditionnelle sont comparés. La question du support est étudiée et un modèle de changement de support est appliqué à plusieurs HAP.

6.1 Principe

La méthodologie est présentée et validée à partir du phénanthrène Phe sur les sondages. Malgré la dissymétrie marquée de son histogramme, ce HAP à 3 cycles ne présente pas de valeurs inférieures au seuil de détection. Dans le cas contraire, ces valeurs inférieures au seuil auraient été ramenées à ce dernier, constituant un *atome* de valeurs identiques ; afin d'étudier son influence sur les résultats, nous créerons artificiellement un tel atome. Phe a pour moyenne et écart-type respectivement 139.94 ppm et 322.23 ppm, soit un coefficient de variation de 2.3. Son variogramme m.p.e. est donné à la figure 6.1(a).

6.1.1 Existence d'effets de bord

Le rapport entre variogrammes simples et croisés d'indicatrices permet de tester l'existence d'effets de bord. Pour 3 coupures de Phe - à 5, 70 et 400 ppm -, ces rapports sont croissants avec h aux premiers pas de calcul, montrant l'existence de tels effets de bord (voir figure 6.2). Un modèle de diffusion est donc adapté, contrairement au modèle mosaïque pour lequel il y a absence d'effets



(a) (b)

FIG. 6.1 – Site Y - Variogrammes moyens par échantillon de Phe (a) et de sa transformée gaussienne (b), avec l'ajustement choisi.

de bord, ou à d'autres modèles de type résidus d'indicatrices pour lesquels il n'existe d'effets de bord que dans un sens.

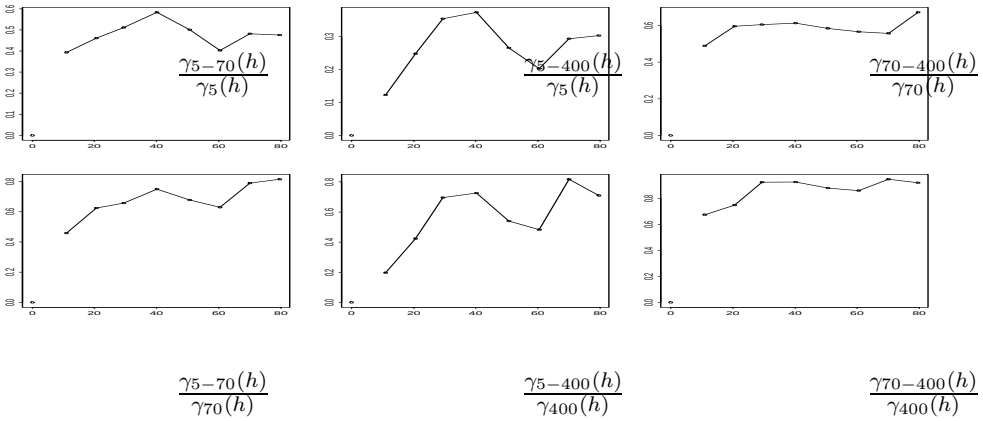


FIG. 6.2 – Site Y - Variogrammes simples et croisés d'indicatrices pour Phe ; coupures à 5 (46 % de données supérieures), 70 (23 %) et 400 ppm (13 %).

Parmi différents modèles de diffusion possibles pour l'estimation des probabilités de dépassement

de seuils, le modèle gaussien anamorphosé est tout d'abord envisagé.

6.1.2 Modèle gaussien anamorphosé

Le modèle gaussien suppose la fonction aléatoire de loi spatiale gaussienne. La teneur en HAP $Z(x)$ ne présentant pas de distribution statistique gaussienne, nous nous y ramenons par une anamorphose gaussienne empirique. La transformée gaussienne $Y(x)$ ainsi obtenue doit obéir à une hypothèse bigaussienne, *i.e.* les lois bivariables $(Y(x), Y(x+h))$ doivent être d'allures bigaussiennes. Différents tests permettent de s'en assurer [Lajaunie (1993)] :

– **Nuages de corrélation différée**
Les nuages de corrélation différée de la transformée gaussienne de Phe, qui représentent les couples de valeurs correspondants aux points séparés par une distance h choisie, doivent être d'allure elliptique (voir figure 6.3), ce qui est raisonnablement le cas.

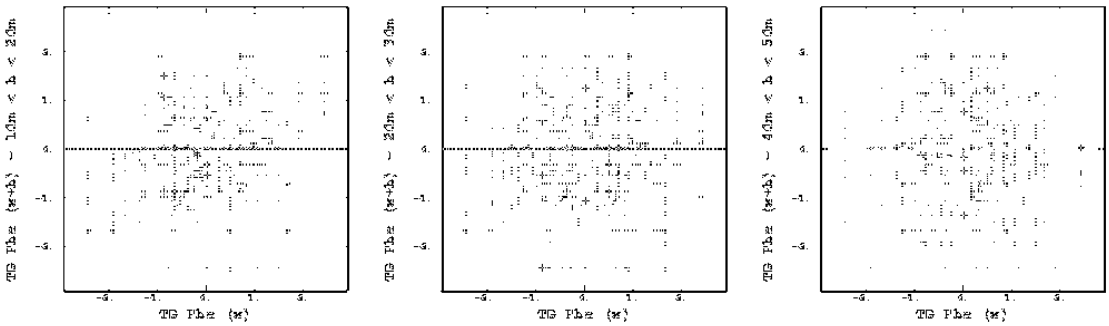


FIG. 6.3 – Site Y - Nuages de corrélation différée de la transformée gaussienne (TG) de Phe, pour des distances croissantes. Coefficients de corrélation respectifs : 0.24, 0.07 et 0.00.

– **Rapport entre variogrammes d'ordre 1 et 2**
Dans le cas d'un modèle bigaussien, les variogrammes d'ordre 1 $\gamma_1(h)$ et 2 $\gamma(h)$ sont liés par la relation [Matheron (1982)]

$$\gamma_1(h) = \sqrt{\frac{\gamma(h)}{\pi}}$$

Autrement dit, le rapport $\frac{\sqrt{\gamma(h)}}{\gamma_1(h)}$ entre les variogrammes d'ordre 2 et 1 est constant[1] et vaut $\sqrt{\pi}$. Ce test, moins lourd mais moins précis que le contrôle des nuages de corrélation différée aux différentes distances, est vérifié dans le cas de Phe (voir figure 6.4).

– **Ajustement de γ_Z à partir de γ^Y**
Nous choisissons comme modèle variographique de la transformée gaussienne de Phe la structure $\gamma^Y(h) = 0.3 + 0.8 \times \mathrm{Sph}(40\ \mathrm{m})$ (voir figure 6.1(b)) dont le palier est supérieur à 1, qui est la variance des données, c'est-à-dire la variance de dispersion d'un point dans le champ V $\sigma^2(0|V) = C_V(0)$.
Cela s'explique si nous considérons une variable régionalisée $z(x)$ connue sur un champ V : la variance des échantillons de V, qui caractérise leur dispersion autour de leur valeur moyenne,

[1] La proportionnalité entre $\gamma_1(h)$ et $\sqrt{\gamma(h)}$ est générale pour les modèles de diffusion purs ; seul le facteur de proportionnalité change [Matheron (1982), Chilès & Delfiner (1999)]. Par ailleurs, dans le cas d'un modèle mosaïque, $\gamma_1(h)$ est proportionnel à $\gamma(h)$.

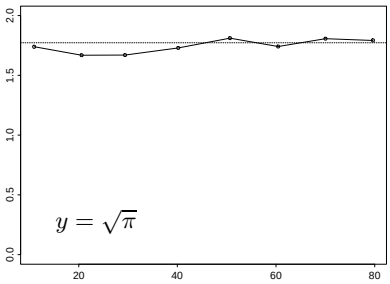


FIG. 6.4 – Site Y - Rapport $\frac{\sqrt{\gamma(h)}}{\gamma_1(h)}$ en fonction de h pour la transformée gaussienne de Phe.

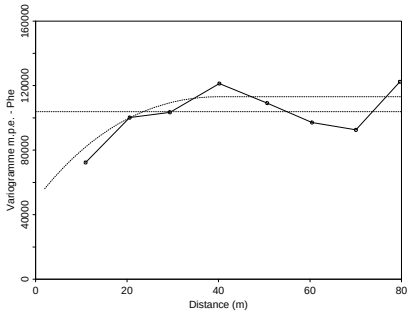


FIG. 6.5 – Site Y - Ajustement indirect du variogramme moyen par échantillon de Phe.

existe toujours et vaut $\sigma^2(0|V)$, le 0 faisant référence aux échantillons. La variance théorique de la FA $Z(x)$, dont z est une réalisation, vaut $C(0)$ si Z est stationnaire, et est infinie dans le cas contraire. Cette variance théorique est égale à la variance de dispersion $\sigma^2(0|\infty)$, d'où l'expression de la variance de dispersion d'un point dans V par la relation de Krige

$$\sigma^2(0|V) = C(0) - \sigma^2(V|\infty) \tag{6.1}$$

Par conséquent, la variance des données est toujours inférieure à la variance a priori. Exprimé en termes de variogramme, cela explique pourquoi le palier du variogramme, s'il existe, est supérieur à la variance des données, dont il est la limite ; l'écart entre les deux diminue au fur et à mesure que la taille du champ V augmente [Journel & Huijbregts (1978), Chilès & Delfiner (1999)]. Sans conséquence lorsqu'il s'agit de la structure de la variable brute, ajuster une structure de la gaussienne ayant un palier supérieur à 1 peut entraîner un problème de convergence du développement en polynômes d'Hermite[2]. La covariance dépend du champ V, contrairement au variogramme [Matheron (1970)]

$$\rho_V(h) = \underbrace{\overline{\gamma}^Y(V,V)}_{=C_V(0)} - \gamma^Y(h)$$

Une fois ajusté $\gamma^Y(h)$ sur le variogramme expérimental de la transformée gaussienne, il est possible de calculer $\overline{\gamma}^Y(V,V)$ et donc d'obtenir la *covariance à champ fixé* $\rho_V(h)$, qui est localement négative et en valeur absolue comprise entre 0 et 1. Connaissant le développement

[2] Un autre problème est mentionné par Chilès & Delfiner (1999) : une simulation non conditionnelle réalisée dans ces conditions aura une variance supérieure à 1, ce qui se répercutera lors du retour en variable brute par anamorphose. Les tests de l'hypothèse bigaussienne, qui portent uniquement sur la transformée gaussienne, ne prennent pas en compte le fait que le champ soit fini.

de l'anamorphose en polynômes d'Hermite, on déduit la covariance $C_V(h)$ de Z par la relation

$$C_V(h) = \sum_{n\geq 1} \phi_n^2 \, [\rho_V(h)]^n$$

en évitant le problème de convergence.
Dans le cas bigaussien, le variogramme $C_V(0) - C_V(h)$ ainsi obtenu doit fournir un bon ajustement du variogramme expérimental de Z. Dans notre cas, l'ajustement indirect de $C_V(h)$ est satisfaisant[3] (voir figure 6.5).

Ces différents tests montrent que l'on peut raisonnablement accepter les hypothèses du modèle gaussien anamorphosé, que nous utiliserons dans la suite[4].

6.1.3 Krigeage disjonctif

Considérons un seuil z_c sur $Z(x) = \Phi[Y(x)]$, et le seuil équivalent $y_c = \Phi^{-1}(z_c)$ sur $Y(x)$. A titre d'exemple, prenons pour Phe la valeur guide $z_c = 100$ ppm, d'où il découle $y_c = 0.8$.

Le krigeage disjonctif (KD), premier estimateur possible, ne requiert qu'une hypothèse bigaussienne et sa mise en œuvre seulement le krigeage des polynômes d'Hermite. Lorsque la concentration de Phe estimée par krigeage est égale à 100 ppm, la probabilité de dépassement de cette valeur estimée par KD est comprise entre 0.01 et 0.27 (voir figure 6.6(b)), la probabilité a priori p de dépasser cette valeur étant de 0.21. Le modèle utilisé pour le krigeage ordinaire, $\gamma(h) = 50000 + 63000 \times \mathrm{Sph}(40\ \mathrm{m})$, approche l'ajustement indirect du variogramme m.p.e. de Phe (voir figure 6.5). Certaines "probabilités" estimées par KD sont inférieures à 0 ou supérieures à 1. Bien que cela soit un inconvénient, il a été vérifié que ces valeurs surviennent en des points où la concentration estimée est largement inférieure au seuil, ou aux points d'échantillonnage.

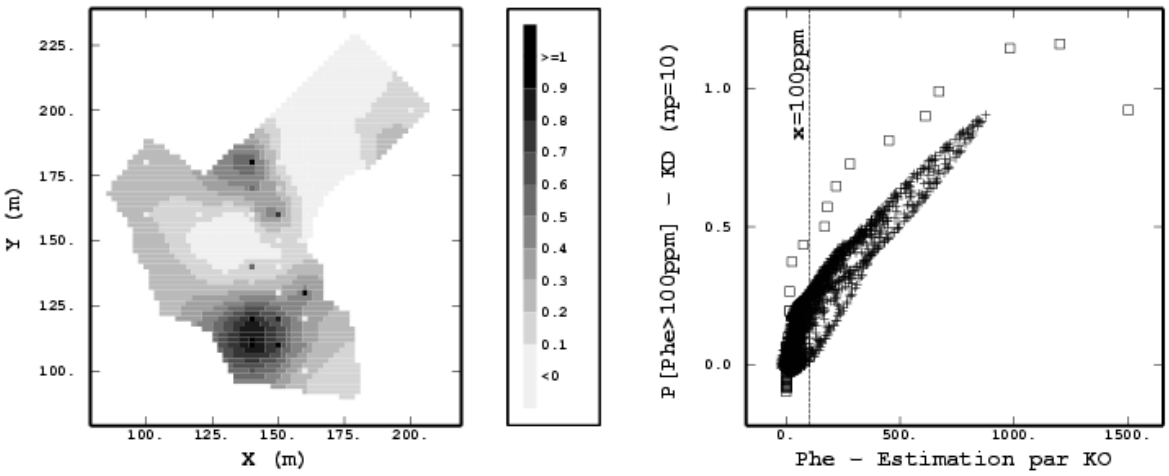
A partir de 10 polynômes, les résultats du KD ne sont plus modifiés, excepté aux points expérimentaux pour lesquels la convergence est beaucoup plus lente, nécessitant plus de 100 polynômes. En chaque point expérimental, l'estimation par KD devrait tendre vers 1 lorsque la concentration est supérieure au seuil de 100 ppm, et vers 0 lorsqu'elle est inférieure à ce seuil. Cependant, le krigeage étant un interpolateur exact, nous avons pour chaque donnée y_i l'estimation par krigeage disjonctif

$$[\mathbb{1}_{y_i \geq y_c}]^{KD} = 1 - G(y_c) - \sum_{n\geq 1} \frac{1}{\sqrt{n}} \, H_{n-1}(y_c) \, g(y_c) \, H_n[y_i]$$

Or, Rivoirard (1985) montre que la convergence du développement tronqué d'une indicatrice en polynômes d'Hermite est très lente, ce qui est illustré à la figure 6.7 pour le développement avec 10 polynômes d'Hermite de $\mathbb{1}_{Y(x)\geq 0.8}$; cela explique que les probabilités de dépassement aux points

[3] Il est possible de procéder en sens inverse, en déterminant numériquement le variogramme $\gamma^Y(h)$ qui découle du meilleur ajustement du variogramme brut. Mais il y a alors perte de cohérence, car si pour toute covariance C il existe une fonction aléatoire de distribution gaussienne admettant C pour covariance, cela n'est pas le cas pour une fonction aléatoire suivant une distribution quelconque, par exemple lognormale [Matheron (1988)].

[4] D'autres modèles existent, adaptés aux distributions dissymétriques - modèle bigamma [Hu (1988)] - et à l'existence d'atomes importants - modèle binomial négatif [Demange et al. (1987)]. Ayant peu de données, l'adéquation du modèle gaussien nous a poussé à privilégier ce choix simple.



(a) (b)

FIG. 6.6 – Site Y - (a) Probabilité de dépassement de 100 ppm estimée par KD (10 polynômes), et (b) nuage de corrélation entre cette probabilité et le krigeage ordinaire de Phe. Les carrés représentent les points expérimentaux.

expérimentaux fluctuent autour de 1 (resp. 0) lorsqu'elles sont bien supérieures (resp. inférieures) au seuil de l'indicatrice, et sont comprises entre 0 et 1 lorsqu'elles sont proches du seuil[5].

6.1.4 Traitement des données égales au seuil de détection

La présence de données inférieures au seuil de détection, inconnues, est source d'indétermination lors de l'anamorphose. Il est important de savoir si cette indétermination ne risque pas de biaiser les estimations, auquel cas une attention particulière doit être portée à ces données dans le modèle.

Pour étudier cela, construisons à partir de Phe un atome artificiel constitué de toutes les concentrations inférieures à 5 ppm. La variable résultante, Phe.t5, possède 28 valeurs ramenées à 5 ppm, soit 54 % des données. On ne constate aucune différence sensible entre les variogrammes moyens par échantillon de Phe et Phe.t5. Par ailleurs, les développements en polynômes d'Hermite de ces deux variables sont très proches. Nous allons comparer les probabilités de dépassement d'une valeur guide égale à 100 ppm estimées pour différentes prises en compte possibles de cet atome. Le facteur entre seuil de détection et valeur guide de dépollution n'étant ici que de 20, cela fait de cet exemple un cas particulièrement pessimiste.

Lors de l'anamorphose, par défaut toutes les valeurs inférieures au seuil de détection sont transformées en une valeur $y_e = \mathrm{E}[Y(x) \,|\, Y(x) < y_p]$, où $y_p = G^{-1}(p)$, p étant le pourcentage de valeurs inférieures au seuil. Afin de prendre en compte cet atome de façon rigoureuse, il serait nécessaire de conditionner le calcul de probabilité par les valeurs de la gaussienne situées dans l'atome[6]. Celles-ci

[5] Ce problème peut être contourné par l'introduction d'un coefficient de changement de support proche de 1 ; nous y reviendrons au paragraphe 6.2 lors de l'application de modèles de changement de support.

[6] Ce conditionnement nécessite cependant une hypothèse multigaussienne plus forte que celle requise par le KD.

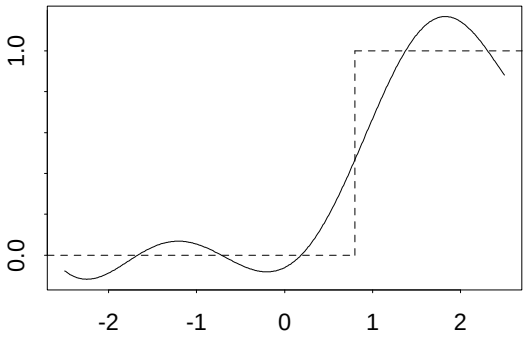


FIG. 6.7 – Site Y - Approximation de l'indicatrice $\mathbb{1}_{Y(x)\geq 0.8}$ par son développement avec 10 polynômes d'Hermite.

étant inconnues, il faudrait
- simuler un grand nombre de réalisations possibles pour les valeurs appartenant à cet atome,
- calculer l'estimation par KD pour chaque réalisation de l'atome,
- en prendre l'espérance.

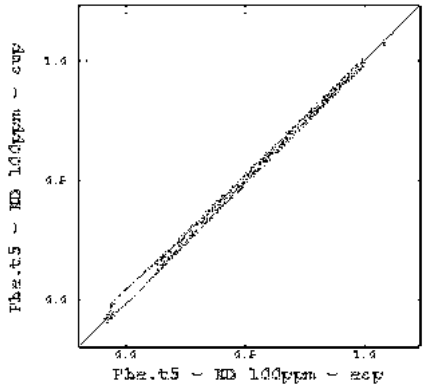
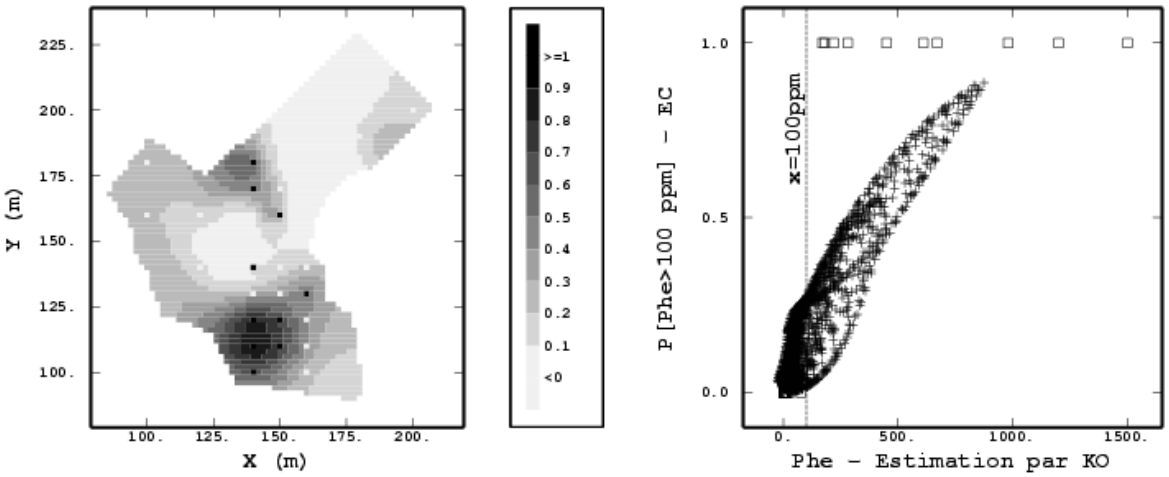


FIG. 6.8 – Site Y - Nuage de corrélation entre probabilités de dépassement de 100 ppm en Phe par KD selon que les valeurs inférieures au seuil de détection sont ramenées sur la gaussienne à $y_e = \mathrm{E}[Y(x)|Y(x) < y_p] = -0.8$ (esp) ou à $y_p = 0.01$ (sup).

Avant d'envisager cette solution, nous avons ramené toutes les valeurs de l'atome à y_p, qui est la borne supérieure des réalisations possibles de la gaussienne sur l'atome. Les probabilités estimées sont similaires à celles estimées en affectant à y_e ces valeurs (voir figure 6.8). Cela s'explique par le fait que dans notre cas le seuil de détection y_p est bien inférieur à la coupure y_c. Il n'est par conséquent pas nécessaire, *dans ce contexte*, de prendre l'atome en compte de manière spécifique.

6.1.5 Espérance conditionnelle

Pour estimer $\mathbb{1}_{Y(x)\geq y_c}$, le meilleur estimateur est l'espérance conditionnelle (EC), calculable dans le cadre d'un modèle multigaussien. Si l'hypothèse multigaussienne est admissible, ce calcul ne nécessite que le krigeage simple de la transformée gaussienne et fournit des résultats cohérents, les probabilités estimées étant comprises entre 0 et 1. La validation de l'hypothèse multigaussienne se réduit en pratique à celle de l'hypothèse bigaussienne ; il est par conséquent délicat d'évaluer dans quelle mesure cette hypothèse multigaussienne est plus contraignante que la seule hypothèse bigaussienne.



(a) (b)

FIG. 6.9 – Site Y - (a) Probabilité de dépassement de 100 ppm estimée par EC, et (b) nuage de corrélation entre cette probabilité et le krigeage ordinaire de Phe. Les carrés représentent les points expérimentaux.

Le résultat obtenu par EC pour le seuil 100 ppm et le nuage de corrélation entre celui-ci et le krigeage ordinaire de Phe sont donnés à la figure 6.9. Les valeurs 0 et 1 correspondent uniquement aux points d'échantillonnage. Pour une valeur estimée à 100 ppm, la probabilité de dépasser ce seuil est comprise entre 0.02 et 0.27, intervalle comparable à celui obtenu par KD.

Le passage au multivariable est aisé en espérance conditionnelle[7], nécessitant seulement dans le calcul de $\mathbb{1}_{Y(x)\geq y_c}$ le cokrigeage simple de Y avec une variable auxiliaire au lieu de son krigeage simple. Un modèle de corégionalisation linéaire de facteurs gaussiens[8] [Freulon (1992)] prenant en compte le facteur auxiliaire ou le goudron modifie peu les probabilités de dépassement estimées par espérance conditionnelle, à cause de l'isotopie de l'échantillonnage déjà discutée au chapitre précédent.

[7] Un krigeage disjonctif multivariable est également envisageable, sous réserve de stationnarité et que les couples de transformées des variables impliquées soient bigaussiens [Maréchal (1982)].

[8] Les ajustements sont réalisés à partir des transformées gaussiennes, en vérifiant que les covariances brutes ajustent bien les structures expérimentales simples et croisées, afin d'assurer la cohérence interne du modèle.

6.1.6 Comparaison

Les différences entre les probabilités de dépassement de 100 ppm estimées par KD et EC sont faibles, si nous excluons les points expérimentaux pour lesquels le KD pose des problèmes de convergence (voir figure 6.10). Les plus élevées, comprises entre 0.05 et 0.1, sont peu nombreuses et les probabilités estimées par KD sont alors supérieures à celles estimées par EC. Les zones où les différences entre KD et EC sont les plus marquées sont regroupées géographiquement et non dispersées sur le site.

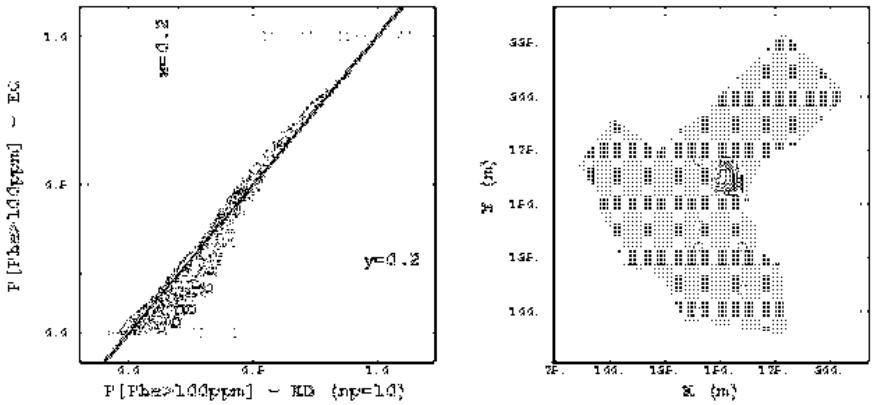


FIG. 6.10 – Site Y - Nuage de corrélations entre KD (abscisse) et EC (ordonnée) pour le seuil 100 ppm ; points pour lesquels la différence entre KD et EC est supérieure à 0.05 repérés par des carrés, avec implantation correspondante sur la grille d'estimation.

On compare à la figure 6.12 les zones où la probabilité de dépassement de 100 ppm estimée par KD et EC est supérieure à 0.2 ou 0.8 avec la sélection que l'on obtiendrait par seuillage à 100 ppm de la carte d'estimation de la concentration. Cette comparaison est menée pour les zones où l'estimation peut se faire dans de bonnes conditions. Une sélection est donc faite sur les écart-types du krigeage ordinaire de Phe (voir figure 6.11), dont la bimodalité de l'histogramme des écart-types est aisément identifiable à la distinction entre zone investiguée, reconnue par la maille, et zone du site non investiguée.

Les résultats montrent tout d'abord des similitudes : les deux zones de valeurs fortes au Nord et au Sud sont bien sélectionnées, que ce soit par seuillage sur la carte estimée, par KD ou par EC. Le seuillage de la carte d'estimation ne permet bien entendu pas d'apprécier le risque pourtant existant de ne pas sélectionner des zones de concentration supérieures au seuil de 100 ppm. Le KD a tendance à sélectionner plus de points que l'espérance conditionnelle. Cependant, pour les zones où la probabilité estimée est supérieure ou égale à 0.2, la zone Ouest est sélectionnée par EC et par contre pas du tout par KD. Les concentrations analysées en Phe y sont égales à 6.9 et 8.3 ppm et donc bien inférieures à 100 ppm. Néanmoins, ces valeurs n'étant pas négligeables au regard de la pollution et ayant à l'esprit l'importante variabilité à petite distance, cette sélection peut se justifier. Concernant les points où la probabilité estimée est supérieure à 0.8, le KD sélectionne légèrement plus de points sans que cela ne soit attribuable à un problème de convergence du développement en polynômes d'Hermite.

A ce stade, nous pouvons conclure à

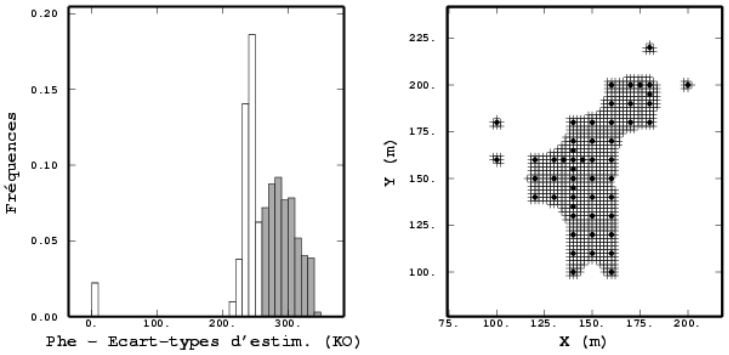


FIG. 6.11 – Site Y - Histogramme des écart-types d'estimation de Phe et implantation correspondante. Points de la grille d'estimation restant après suppression des points ayant un écart-type supérieur à 260 ppm, représentés en grisé sur l'histogramme. Points expérimentaux représentés par des ronds noirs.

- l'existence de problèmes de convergence aux points expérimentaux en KD,
- une plus grande rapidité du calcul par EC,
- une bonne similitude des résultats entre krigeage disjonctif et espérance conditionnelle.

Cette estimation de la probabilité de dépassement de seuil constitue un apport crucial par rapport au seuillage direct sur la carte d'estimation, car il permet de faire un choix cohérent avec le risque que l'on s'autorise de laisser en place un point où la concentration serait supérieure au seuil fixé. Prenons sur la figure 6.12 la carte correspondant à la sélection par KD des zones où la probabilité de dépassement de 100 ppm est supérieure ou égale à 0.2. Elle nous apprend que, si la zone sélectionnée - en rouge - est dépolluée, alors en échantillonnant dans le matériau laissé en place, nous trouverons avec une probabilité au plus égale à 0.2 - soit 1 chance sur 5 - une concentration supérieure à 100 ppm. C'est bien cela qui est en jeu en cas de contrôle ultérieur de la dépollution.

Néanmoins, ce calcul, mené en ponctuel, ne prend pas en compte le support de dépollution : en effet, lors de la dépollution, on ne traitera par les terres à l'échelle de l'échantillonnage - une fosse, voire une carotte de sondage - mais à une échelle beaucoup plus large, par exemple des blocs de 5 m par 5 m. Ce support de dépollution dépend des possibilités d'intervention des engins et de l'objectif de la dépollution en terme de risque sanitaire. En effet, selon que le site est destiné à devenir une zone industrielle ou un jardin d'enfants, le risque que l'on s'autorise varie : tandis que dans le premier cas il peut être suffisant d'assurer une concentration moyenne inférieure au seuil de dépollution sur des blocs relativement larges, on imposera au contraire que ce seuil ne soit pas dépassé sur des supports de taille beaucoup plus réduite, voire ponctuelle, dans le second cas ! Prendre en compte ce support lors de la sélection des zones à traiter est donc primordial.

6.2 Changement de support en modèle gaussien discret

Le modèle gaussien discret permet le calcul de probabilités de dépassement de seuils de pollution sur des blocs v de taille donnée. Ce modèle repose sur l'hypothèse que les lois des transformées

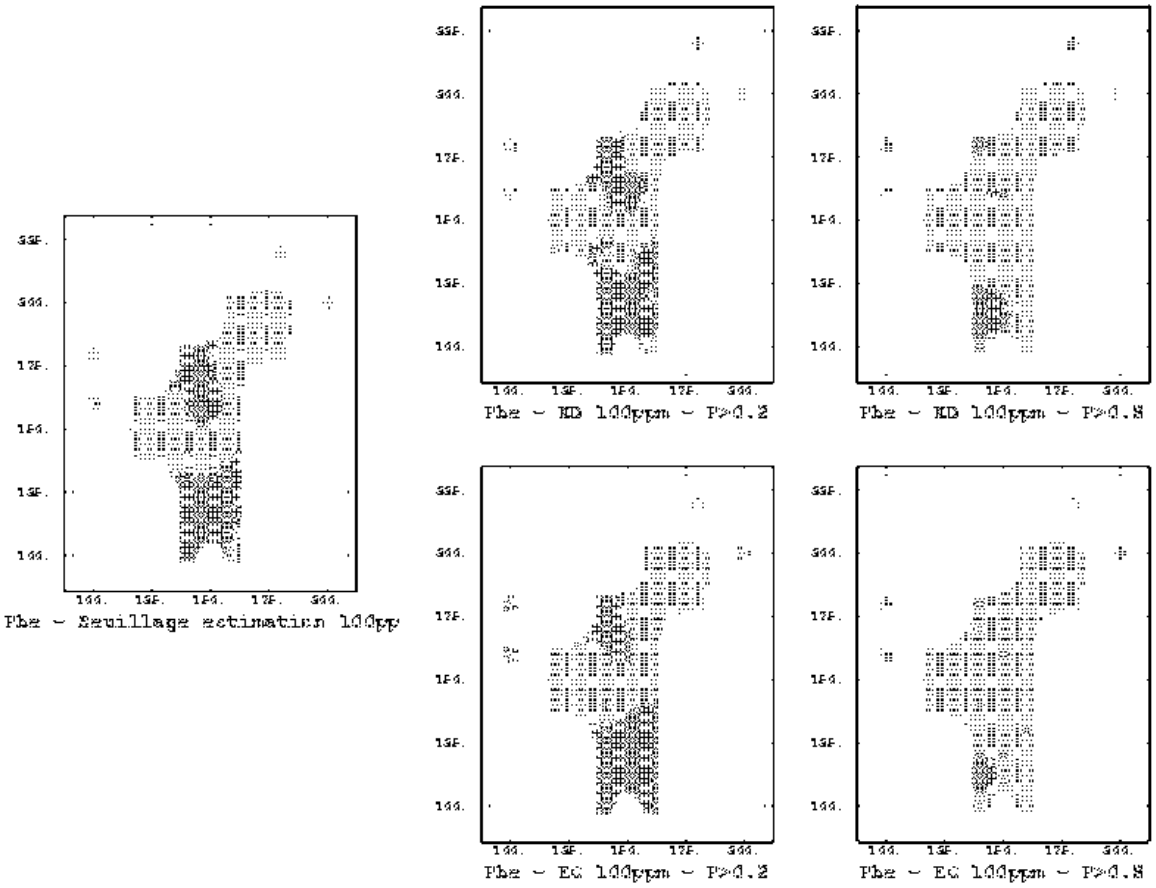


FIG. 6.12 – Site Y - Seuillage à 100 ppm de l'estimation de Phe par krigeage ordinaire, zones où la probabilité de dépassement de 100 ppm en Phe estimée est supérieure à 0.2 et 0.8 pour le krigeage disjonctif et l'espérance conditionnelle. Les carrés indiquent les zones sélectionnées.

gaussiennes des points et des blocs sont bigaussiennes pour un calcul par krigeage disjonctif, ou multigaussiennes pour un calcul par espérance conditionnelle.

Ces deux calculs sont comparés pour un seuil de 100 ppm de Phe sur les sondages. Nous ne nous intéressons plus au seuil $y_c = \Phi^{-1}(z_c) = 0.8$ sur la gaussienne, mais au seuil correspondant sur la gaussienne des blocs $y_{v_c} = \Phi_{(v)}^{-1}(z_c) = 0.59$. Le coefficient de changement de support est égal à 0.93. Les résultats par KD et EC en modèle gaussien discret pour des blocs de 5 × 5 m concordent (voir figure 6.13), même si l'estimation de la probabilité par EC est légèrement plus contrastée. L'estimation par KD tend à diluer au milieu du champ les deux zones de probabilité estimées élevées, ce que la figure 6.14 confirme ; en effet, les différences les plus marquées entre les estimations par KD et EC se situent :

- dans les zones où les concentrations en place sont estimées faibles - milieu du champ -, où les probabilités estimées par KD sont supérieures à celles par EC ;
- à la périphérie des zones de concentrations estimées fortes, où les probabilités estimées par EC sont cette fois plus élevées.

Ces constatations, auxquelles s'ajoutent la plus grande rapidité du calcul par espérance conditionnelle et sa cohérence, nous poussent à préconiser ce calcul dans le cas présent. La figure 6.15(a) illustre l'effet du modèle gaussien discret sur les probabilités estimées par EC en ponctuel et sur des blocs ; le modèle de dispersion tend à augmenter les probabilités estimées de dépasser le seuil, excepté pour les blocs contenant les points expérimentaux. Par ailleurs, la figure 6.15(b) montre néanmoins que dans le cas présent la sélection reste délicate - certains blocs aux concentrations estimées faibles ayant une probabilité estimée non négligeable de dépasser 100 ppm -, même pour un seuil aussi élevé que 100 ppm.

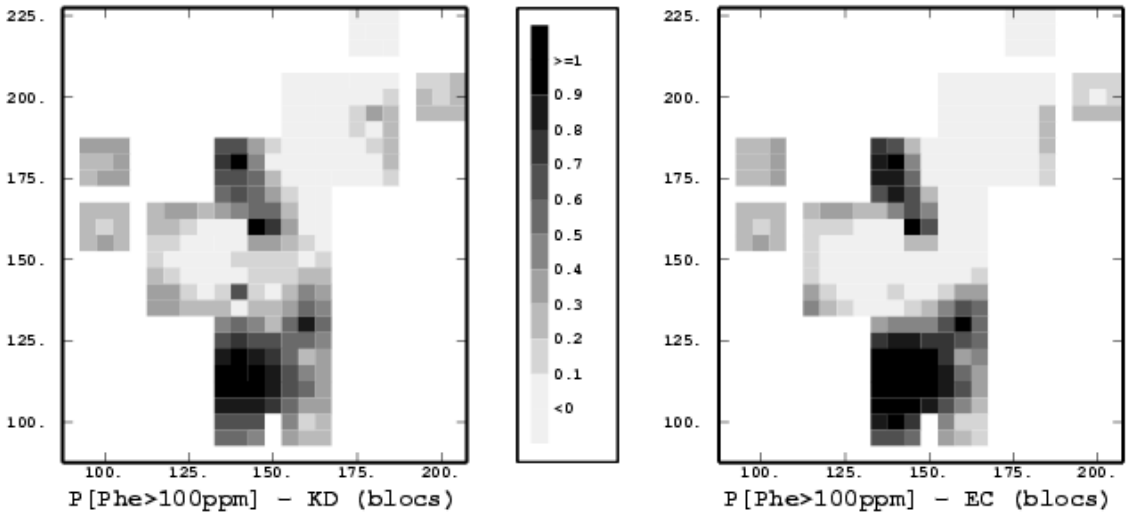


FIG. 6.13 – Site Y - Probabilité de dépassement de 100 ppm estimée par KD et EC en modèle gaussien discret.

Avant d'appliquer la méthode choisie à d'autres HAP, il a été vérifié que les résultats ne sont pas sensiblement modifiés par un travail en accumulation plutôt qu'en concentration (voir paragraphe 3.2.4). Ainsi, les différences les plus importantes sont de l'ordre de 0.2 entre les estimations des probabilités de dépassement de 100 ppm en Phe et du produit entre 100 ppm et la masse moyenne de la fraction inférieure à 2 mm pour l'accumulation. Elles surviennent uniquement pour les deux zones de fortes concentrations Nord et Sud et sont plus élevées en concentration qu'en accumulation. Cela ne semble pas justifier le calcul en accumulation.

6.3 Application

Les probabilités de dépassement des seuils 10 et 100 ppm pour Nap et Bap sur les sondages sont estimées par espérance conditionnelle en modèle gaussien discret. Le seuil de 10 ppm correspond à une norme utilisée pour les HAP pris individuellement. Les sélections des zones de dépassement des seuils sont présentées pour les probabilités 0.2, 0.5 et 0.8. Le cas 0.2 constitue un choix prudent, conduisant à la sélection de toutes les zones ayant au moins une chance sur 5 de dépasser le seuil. Inversement, la probabilité 0.8 représente un choix risqué, seules les zones pour lesquelles on a plus

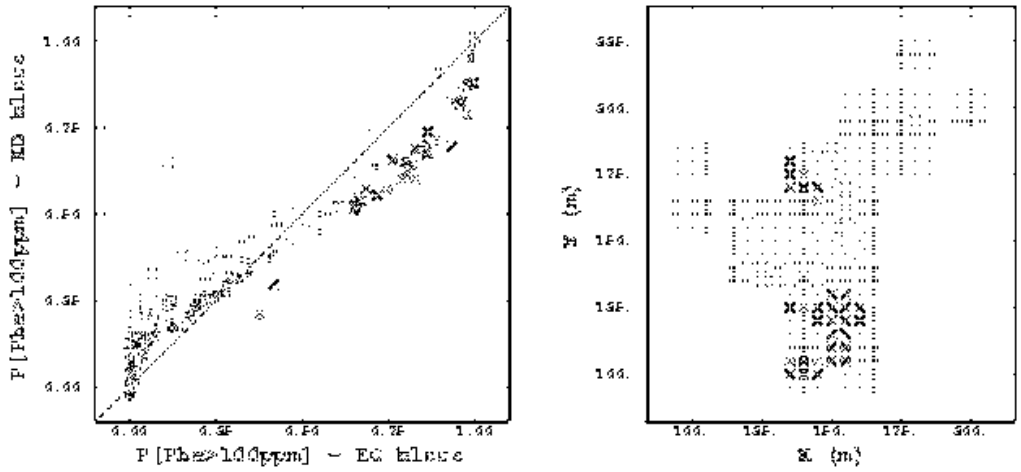


FIG. 6.14 – Site Y - Nuage de corrélation entre les probabilités de dépassement de 100 ppm estimées par KD (P^*_{KD}) et EC (P^*_{EC}) en modèle gaussien discret, avec indication de certains points sur la grille d'estimation : les carrés représentent les blocs pour lesquels $P^*_{KD} - P^*_{EC} \geq 0.2$, les étoiles ceux pour lesquels $P^*_{EC} - P^*_{KD} \geq 0.2$. Les tirets indiquent la première bissectrice.

de 80 % de chance d'avoir une concentration supérieure au seuil étant sélectionnées.

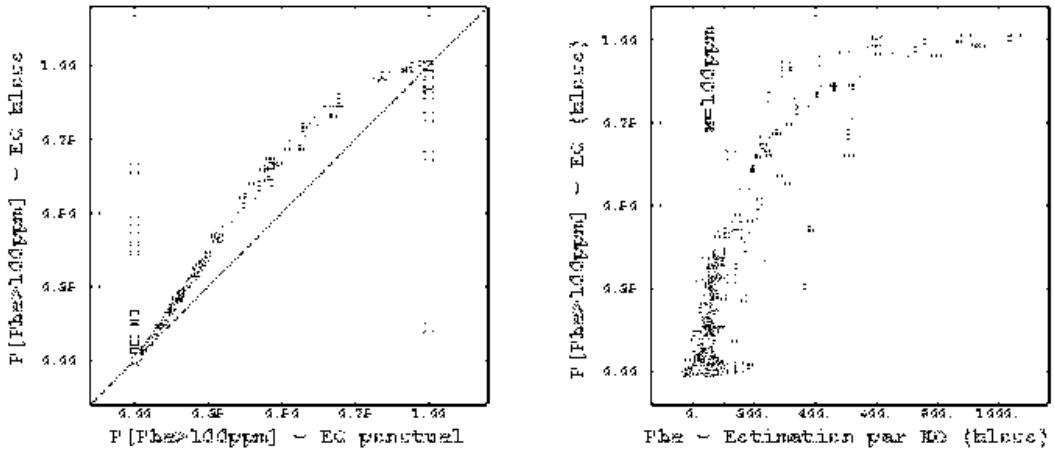
Malgré la stationnarité discutable de Nap, la méthode a été appliquée telle quelle, ce qui peut poser problème car tout modèle de changement de support nécessite une hypothèse de stationnarité stricte[9]. Les résultats montrent que pour le seuil 10 ppm, l'entièreté du site est quasiment sélectionnée pour la probabilité 0.2, excepté 6 blocs de la zone Nord-Est (voir figure 6.16). Ensuite, si l'on s'accorde un risque plus important, on laisse en place l'ensemble de la zone Nord-Est, et l'Est de la partie centrale. Pour un risque élevé on ne sélectionne plus que la zone Sud, la tache de pollution située au Nord de la partie centrale du site et quelques blocs à l'Ouest du site. Les zones sélectionnées pour le seuil 100 ppm sont logiquement plus restreintes que pour le seuil 10 ppm ; 21 % des concentrations de Nap sur les sondages sont supérieures à 100 ppm. Le risque de 0.5 ne conduit à sélectionner que la mare Sud. Une prise de risque élevée laisse en place l'ensemble des blocs.

Bap ne présente pas de problème de stationnarité et nous avons montré que les 26 % de données inférieures au seuil de détection de 0.1 ne posent pas de problème pratique. Le seuil à 10 ppm conduit à la sélection de zones plus réduites que pour Nap, le Bap étant quantitativement moins présent sur le site[10] (voir figure 6.17). Aucune zone n'est sélectionnée pour une probabilité de dépasser 100 ppm supérieure à 0.5, et par conséquent supérieure à 0.8.

En conclusion, une fois choisis le seuil d'action et le support sur lequel on considère ce seuil, la méthode permet l'estimation de probabilités de dépassement de ce seuil sur des blocs de taille égale au support fixé. La sélection des zones où la probabilité estimée de dépasser le seuil est supérieure

[9] Il est possible de contourner ce problème par un changement de support aux points expérimentaux [Rivoirard (1991)] ; cela n'a pas été approfondi ici.

[10] Il est néanmoins important de garder à l'esprit que Bap est nettement plus toxique que Nap.



(a) (b)

FIG. 6.15 – Site Y - (a) Nuage de corrélation entre les probabilités de dépassement de 100 ppm estimées par EC en ponctuel et en modèle gaussien discret. Les carrés représentent les blocs contenant les points expérimentaux, les tirets indiquent la première bissectrice. (b) Nuage de corrélation entre probabilité de dépassement de 100 ppm en Phe (sondages) estimée par EC en modèle gaussien discret et estimation par KO.

à une valeur fixée constitue alors un outil d'aide à la décision intéressant.

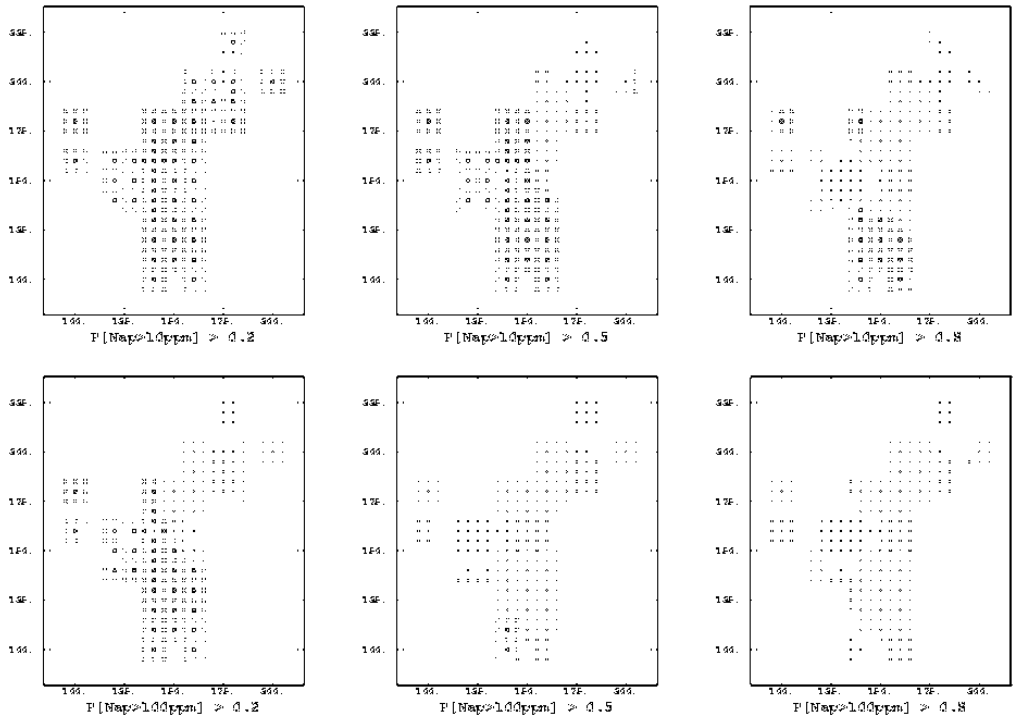


FIG. 6.16 – Site Y - Probabilités de dépassement de 10 et 100 ppm de Nap (sondages) estimées par EC en modèle gaussien discret ; sélection (□) des blocs où l'estimation est supérieure à 0.2, 0.5 et 0.8.

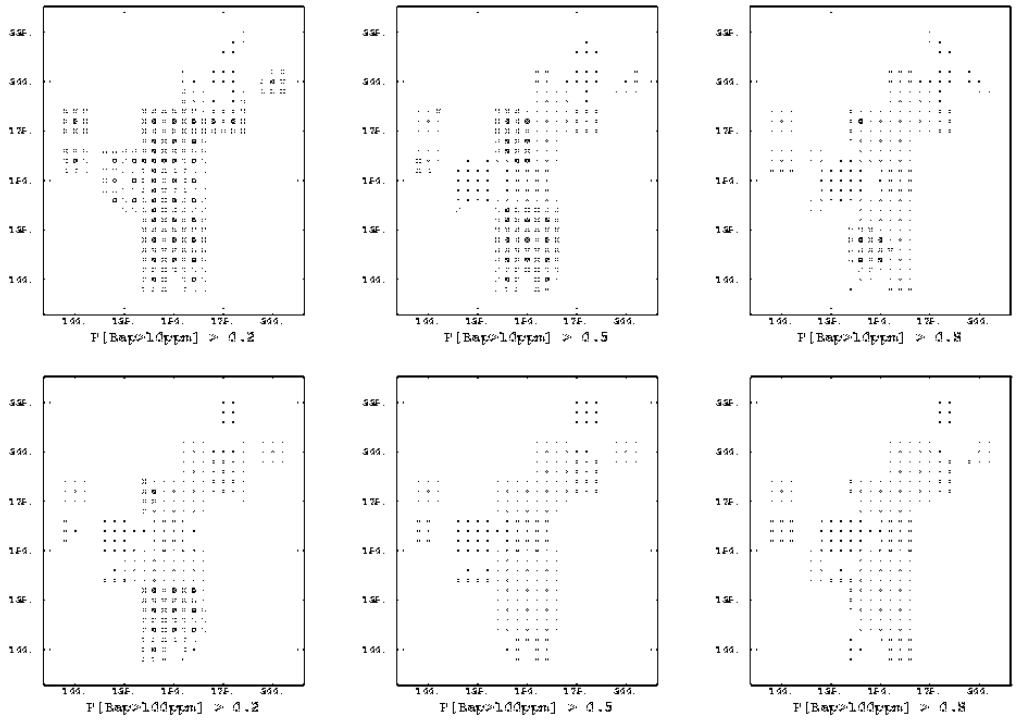


FIG. 6.17 – Site Y - Probabilités de dépassement de 10 et 100 ppm de Bap (sondages) estimées par EC en modèle gaussien discret ; sélection (□) des blocs où l'estimation est supérieure à 0.2, 0.5 et 0.8.

Quatrième partie

Site X

Chapitre 7

Echantillonnage, analyse exploratoire et variographique, estimations

Sommaire

Après avoir présenté le second site et la campagne d'échantillonnage, les analogies et les différences entre les deux sites sont décrites lors de l'analyse exploratoire. Plusieurs variables qualitatives et semi-quantitives sont comparées aux analyses chimiques classiques. Les concentrations en HAP apparaissent très peu structurées ; on discute le sens d'une estimation globale puis locale dans pareil cas.

7.1 Site et stratégie d'échantillonnage

7.1.1 Description et historique

L'étude historique de cette friche minière du Nord de la France a été réalisée lors de campagnes préalables des organismes A et B, en 1994 et 1996. Trois cokeries s'y sont succédées entre 1925 à 1973. Le site a atteint une production maximale de 287 000 tonnes de coke par an en 1936 et traitait également ses sous-produits : sulfates, benzols, goudrons. La figure 7.1 présente une vue générale du site et des anciennes infrastructures. On y distingue notamment la cokerie à proprement parler avec ses batteries et la tour à charbon, au sud, une aire de stockage avec les voies ferrées à l'est, une zone de bassins de décantation et de réfrigérants, une fosse à brai, une usine à benzol, une usine à sulfate ainsi que des gazomètres et une station d'épuration des gaz à l'ouest. Le démantèlement complet des infrastructures, sur lequel nous n'avons pas d'information, s'est achevé en 1980. A ce jour, la partie sud a été réaménagée et reverdie. Aucun aménagement n'a été entrepris dans la partie nord, et une végétation parfois abondante a recouvert le site laissé à l'abandon. De par son activité ce site présente également en certains endroits des concentrations plus ou moins importantes en phénols, cyanures totaux, sulfates et métaux - chrome total, zinc, cadmium, cuivre.

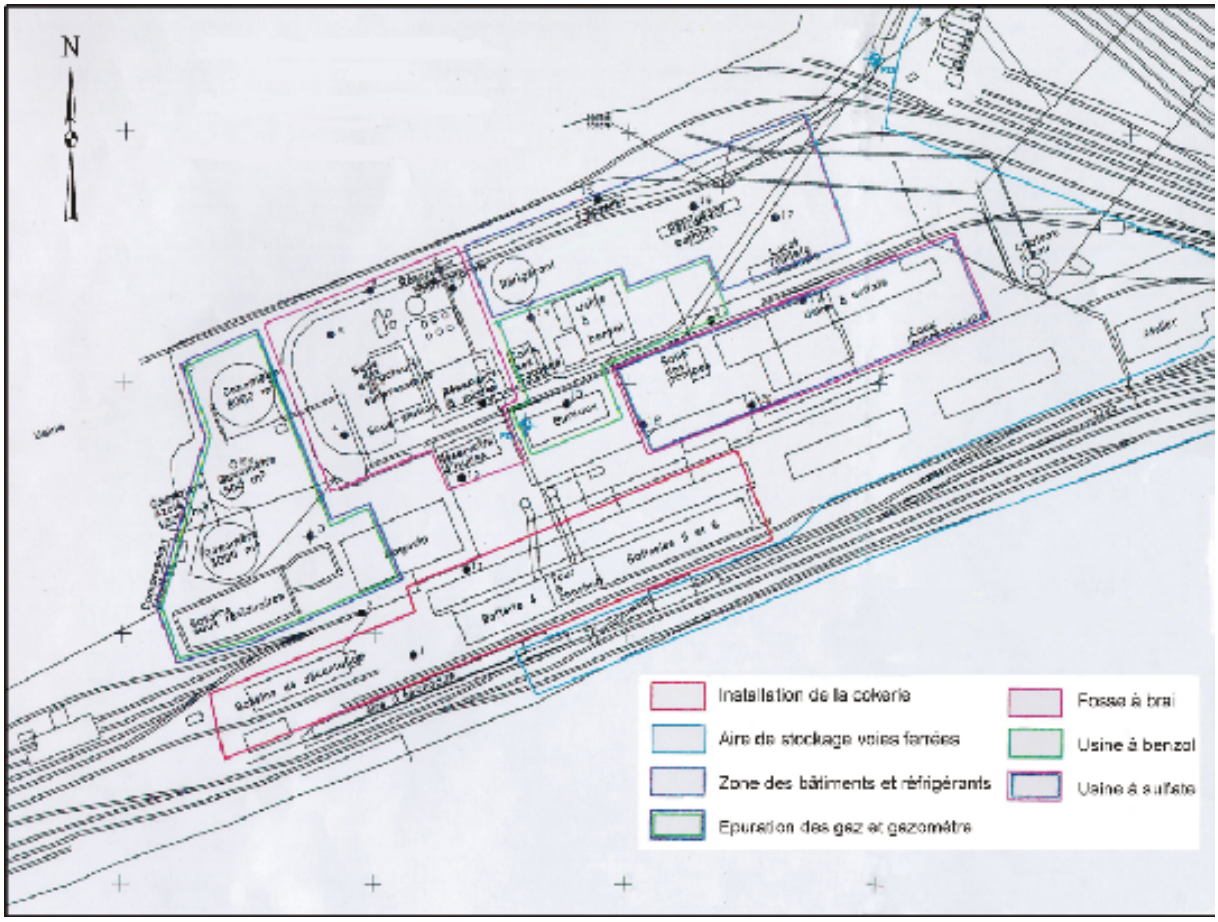


FIG. 7.1 – Site X - Vue générale du site, implantation des infrastructures.

Du point de vue géologique, on trouve à l'aplomb du site, de la surface vers la profondeur :
- quelques mètres de formations superficielles essentiellement composées de produits de démantèlement des installations de la cokerie. On y trouve également, mélangés à des silts ou des fines schisto-charbonneuses, de nombreuses briques entières ou en fragments, des blocs de maçonnerie atteignant parfois plusieurs mètres, de la ferraille, des câbles, etc. Il est par conséquent difficile d'apprécier l'épaisseur de ces remblais, constitués également de terre végétale et d'argile rapportée.
- des limons quaternaires beiges silto-argileux (environ 6 m),
- la craie séno-turonienne (plus de 60 m),
- les marnes bleues du Turonien moyen.

7.1.2 Campagne d'échantillonnage

L'échantillonnage de la zone sélectionnée a été réalisé en 1997 suivant un maillage carré de 20 m de côté (voir figure 7.2). La partie Nord-Est de la friche n'a pu être échantillonnée en raison d'une végétation trop abondante. 51 points ont été retenus [Wavrer (1997a)].

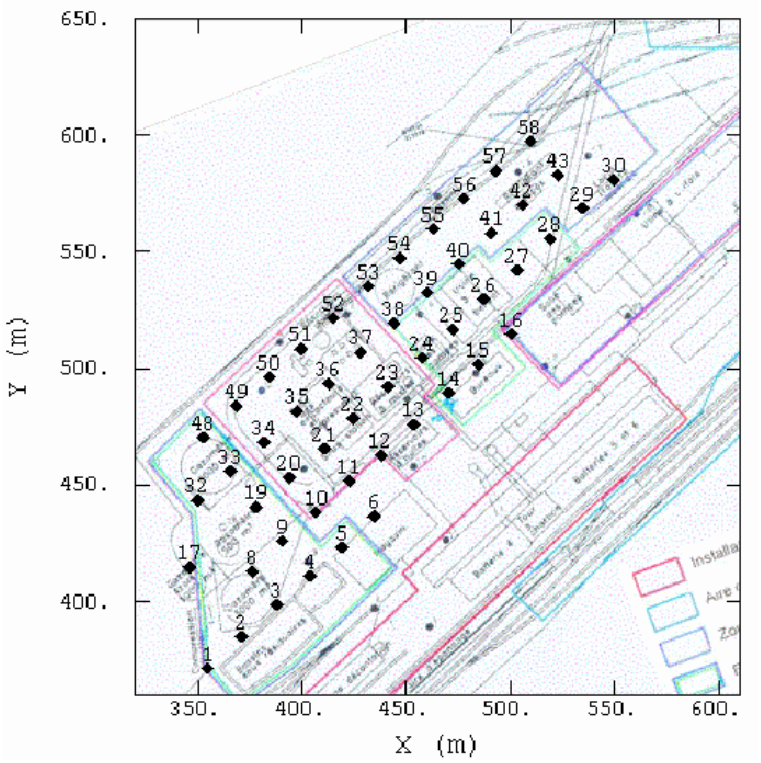


FIG. 7.2 – Site X - Implantation des données superposée au plan des installations.

Comme sur le site Y, deux prélèvements ont été réalisés en chaque point : par rainurage vertical le long de la paroi d'une fosse d'environ 0.5 m de profondeur, ainsi que par prélèvement du dernier mètre d'un sondage de 1.50 m réalisé en un coin de la fosse. En pratique, il est rare que l'on ait

réussi à récupérer un mètre complet, à cause du tassement du matériau et des obstacles rencontrés. L'histogramme des hauteurs récupérées est relativement symétrique avec un mode situé à 0.7 m, ce qui correspond à une perte de 30 % du sondage. Les 16 HAP de l'US EPA (en $mg.kg^{-1}$ sec) ainsi que l'indice Phénol (en $mg.kg^{-1}$) ont été analysés.

7.2 Analyse exploratoire

7.2.1 Fosses et Sondages

7.2.1.1 Statistiques élémentaires

Même si cela est peu significatif compte tenu des écarts-types élevés, les concentrations moyennes sont plus élevées sur les fosses (voir tableau 7.1). Les coefficients de variation (σ/m) sont plus élevés sur les fosses pour les HAP légers, cette tendance s'inversant à partir des HAP à 4 cycles. Les pourcentages de valeurs inférieures aux seuils de détection sont relativement faibles, excepté pour Acy (plus de 70 %), Dba (plus de 10 %), Ace et Fle sur les fosses (33 et 18 %).

HAP	Nbre de cycles	Données de fosses				Données de sondages			
		m	σ	σ/m	DIS	m	σ	σ/m	DIS
Nap	2	43.43	204.23	4.70	0	17.39	25.60	1.47	0
Phe	3	53.22	119.97	2.25	0	48.14	101.35	2.11	0
Baa	4	30.11	37.92	1.26	2	28.06	66.54	2.37	4
Bap	5	24.61	28.29	1.15	0	17.60	29.80	1.69	0
Inp	6	12.95	15.01	1.16	6	10.62	16.65	1.57	8
Phl	-	4.73	15.82	3.34	10	1.85	5.41	2.92	16

TAB. 7.1 – Site X - Statistiques des données de fosses et sondages. Les pourcentages de données inférieures aux seuils de détection (DIS) sont donnés. Le nombre de cycles est indiqué pour les HAP.

Les médianes sont systématiquement bien inférieures aux moyennes, ce qui est caractéristique de distributions dissymétriques (voir tableau 7.2). L'écart entre le quantile à 75% et la valeur maximale caractérise l'ampleur de la queue de la distribution, constituée par quelques valeurs parfois extrêmement fortes.

On observe entre les fosses et les sondages quelques différences notables. Pour les HAP à 2 et 3 cycles, les quantiles à 25, 50 et 75 % sont plus faibles sur les fosses que sur les sondages ; cela s'inverse à partir de 4 cycles. Les valeurs maximales étant par ailleurs dans l'ensemble supérieures sur les sondages pour ces HAP, on en conclut la plus grande dissymétrie des données de sondages pour les HAP d'au moins 4 cycles, chose que l'on ne percevait pas de manière systématique avec les écarts-types.

Par rapport au site Y, ce site présente un niveau de pollution moyen sensiblement moins élevé en profondeur, alors qu'en surface les concentrations moyennes sont relativement proches sur les

HAP	Données de fosses					Données de sondages				
	Min	25 %	50 %	75 %	Max	Min	25 %	50 %	75 %	Max
Nap	0.34	1.48	4.00	10.80	1425.00	0.30	2.50	6.89	18.70	144.00
Phe	0.10	1.90	9.70	35.00	628.00	0.38	7.80	13.00	28.00	525.00
Baa	0.10	3.30	12.00	39.00	150.00	0.10	0.80	6.60	18.00	450.00
Bap	0.18	5.10	8.70	39.00	110.00	0.04	0.94	5.70	17.00	160.00
Inp	0.01	2.00	5.10	20.00	57.00	0.10	0.68	4.00	9.70	74.00
Phl	0.10	0.32	1.00	2.40	88.00	0.10	0.16	0.38	1.60	39.00

TAB. 7.2 – Site X - Valeurs extrêmes et quartiles.

deux sites. Les coefficients de variation, moins importants, reflètent le caractère moins variable des données du site X. Cela est d'ailleurs confirmé par les pourcentages de données inférieures aux seuils de détection, bien moins importants ici que sur le site Y. Le site Y présentait en outre des concentrations plus fortes en certains points. Les deux premiers quartiles (25 % et 50 %) sont largement inférieurs sur le site Y. Pour les fosses, c'est encore le cas pour le troisième quartile. En résumé, tandis que le site Y présentait une proportion importante de concentrations très faibles, voire mal détectées, et quelques concentrations très fortes, ce site X se distingue par une variabilité moins importante des concentrations.

En ramenant les concentrations des différents HAP à la concentration totale en HAP, on remarque la prédominance de Phe, Flt et Nap sur les fosses ainsi que de Pyr, Phe et Flt sur les sondages (voir figure 7.3). Acy est très peu présent, tout comme Ant et les HAP les plus lourds : Dba, Bkf, Bgh et Inp.

Par rapport au site Y, la prédominance de Flt et Nap par rapport aux autres HAP est moins sensible ; à l'inverse, Ace, Inp, Bbf, Bkf sont plus présents. Cela peut provenir de l'utilisation de goudrons de signatures - compositions - différentes sur cette cokerie. On constate par ailleurs une différenciation entre fosses et sondages plus élevée, qui pourrait indiquer une moins bonne corrélation entre les deux modes de prélèvements.

7.2.1.2 Histogrammes

Les histogrammes de la figure 7.4 illustrent la dissymétrie des HAP, systématique et plus marquée sur les sondages pour les HAP à partir de 4 cycles.

Pour les fosses, on distingue des traces importantes de pollution au Nord du site près d'un bassin de composés phénoliques et d'une usine à benzol (fosse 55), au centre du site à l'emplacement d'anciens réservoirs (fosse 37 pour tous les HAP, également les fosses voisines pour les HAP lourds) ainsi qu'à l'extrême Ouest pour les HAP les plus lourds (fosses 3, 9, 33), près de gazomètres. L'implantation des valeurs fortes pour les sondages fait surtout ressortir la fosse 42 (Nord du site), les fosses 33 et 48 (Ouest) près d'un gazomètre pour les HAP les plus lourds ainsi que la fosse 22 (au centre, près de réservoirs) pour ces mêmes HAP, ce dernier point ne ressortant absolument pas sur les fosses. L'implantation des valeurs fortes diffère donc notablement entre les deux modes de prélèvement !

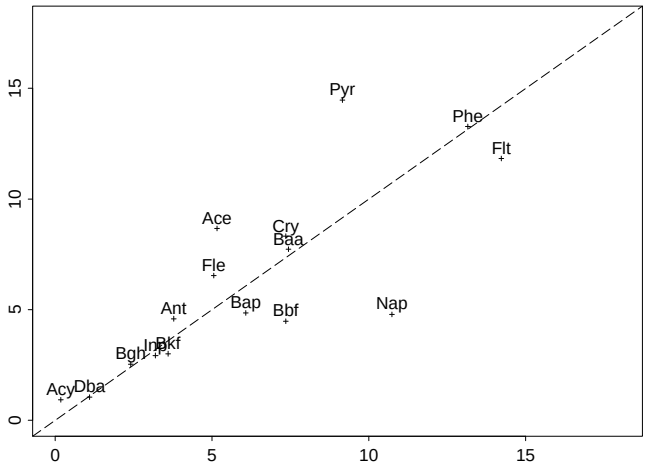


FIG. 7.3 – Site X - Pourcentages comparés en HAP entre fosses et sondages, par rapport à la teneur totale en HAP. Echelle identique à celle de la figure 3.4, pour le site Y.

Par rapport au site Y, qui présentait d'une part des zones de faibles concentrations et localement quelques groupes de fortes concentrations, les fortes concentrations sont plus dispersées sur la zone investiguée.

7.2.1.3 Corrélations

Tout comme pour le site Y, pour chaque type de prélèvement le niveau de corrélation est très bon pour les HAP de poids atomique et de nombre de cycles voisins, en particulier pour les HAP lourds (voir tableau 7.3). Les variables provenant des fosses sont légèrement mieux corrélées.

Les nuages de corrélation (voir figure 7.5) montrent que pour le couple Nap-Inp deux points se détachent de l'axe de corrélation ; ils correspondent aux fosses 37 et 55 déjà mentionnées. Une présence aussi forte de napthalène peut résulter d'une pollution locale par un goudron de signature différente des autres points, ou par un mélange de plusieurs goudrons.

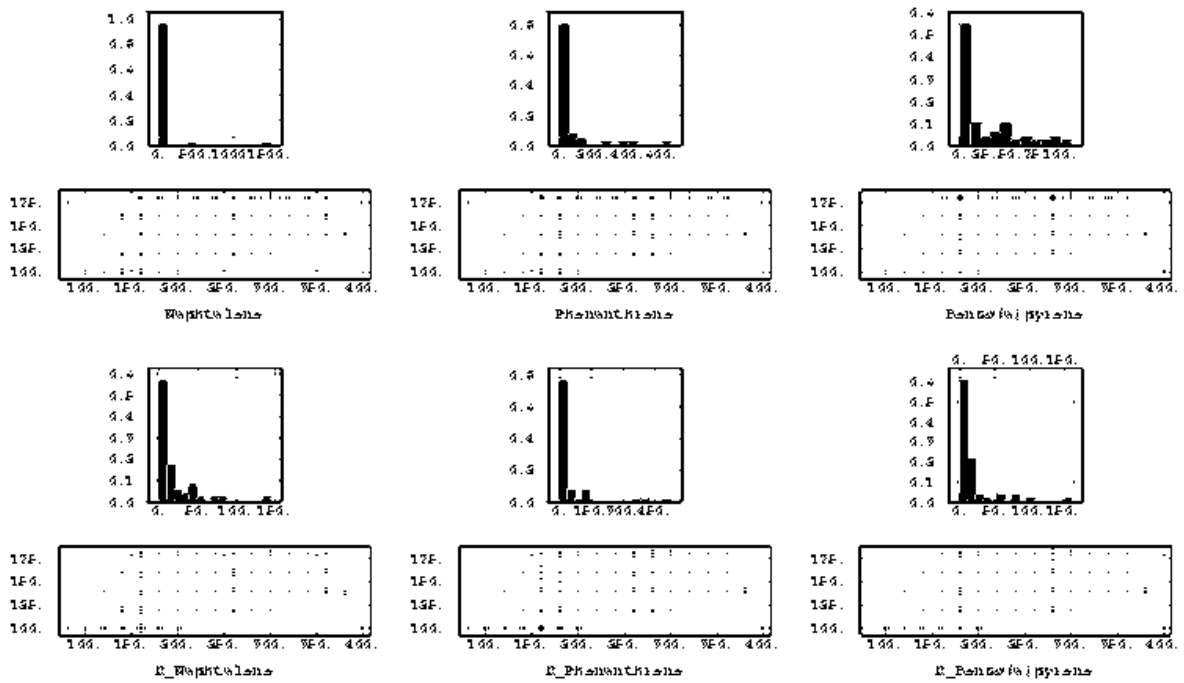


FIG. 7.4 – Site X - Histogrammes des données de fosses et de sondages. Les noms des variables correspondant aux sondages sont précédés de "*S_*". Les échelles de représentation des croix diffèrent pour chaque figure.

7.2.1.4 Corrélation fosses/sondages

Les coefficients de corrélation entre les concentrations en HAP sur les fosses et les sondages sont très faibles (voir tableau 7.4) - ce coefficient est en outre égal à 0.06 pour l'indice phénol. L'importance du remaniement sur le site peut expliquer cela. On note juste une légère amélioration de ces coefficients avec l'augmentation du nombre de cycles des HAP, explicable par la moindre mobilité des HAP les plus lourds.

Un HAP peut présenter en certains points une concentration très élevée sur le sondage et très faible pour la fosse, et réciproquement (voir figure 7.6). Les fosses, qui prélèvent en surface, ne sont donc absolument pas indicatrices d'une pollution du niveau inférieur.

7.2.2 Indices qualitatifs

La présence d'un certain nombre d'indices organoleptiques et qualitatifs a également été décrite sur ce site lors du prélèvement de chaque échantillon : présence d'odeur, de goudron, de débris de maçonnerie, de laitier, de craie, de charbon, etc. On observe notamment la présence de débris de maçonnerie sur la plupart des fosses et des sondages. Une odeur de goudron a été relevée essentiellement sur la fosse à brai, à proximité des réservoirs, et également près du bassin au Nord et en certains points situés près des usines de benzol et sulfates. La présence de goudron est

HAP	Nap	Phe	Baa	Bap	Inp
Nap	1.00				
Phe	0.79	1.00			
Baa	0.49	0.53	1.00		
Bap	0.52	0.50	0.91	1.00	
Inp	0.44	0.47	0.93	0.96	1.00
Phl	0.84	0.82	0.62	0.59	0.57

Fosses

HAP	Nap	Phe	Baa	Bap	Inp
Nap	1.00				
Phe	0.70	1.00			
Baa	0.26	0.58	1.00		
Bap	0.26	0.50	0.92	1.00	
Inp	0.29	0.44	0.73	0.92	1.00
Phl	0.78	0.56	0.13	0.14	0.21

Sondages

TAB. 7.3 – Site X - Coefficients de corrélation entre HAP, pour les fosses et les sondages.

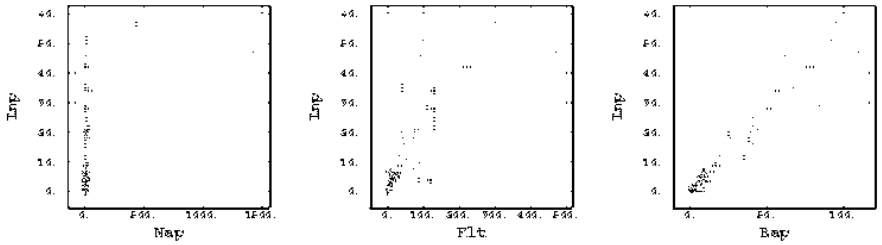


FIG. 7.5 – Site X - Exemples de nuages de corrélation pour les fosses. Echelles en ppm.

qualitativement comparable à celle d'odeur, mais plus marquée sur les fosses que sur les sondages. Deux types lithologiques sont rencontrés : des silts sableux et des silts argileux.

En suivant la même méthodologie que pour le site Y, on montre qu'aucun indice n'est correctement lié aux concentrations ; en particulier, le goudron n'est pas utilisable comme variable guide des zones de forte concentration, les concentrations les plus fortes n'étant pas détectées pour les HAP lourds. L'analyse des correspondances n'a conduit à aucun résultat probant. Par ailleurs, même si l'essentiel des fortes concentrations est situé dans les matériaux rapportés, quelques concentrations en HAP non négligeables sont situées dans les silts.

7.2.3 Mesures de gaz

Afin d'établir un premier diagnostic de l'état réel du terrain, d'orienter les analyses de laboratoire en permettant un tri préalable des échantillons récoltés et donc d'optimiser le coût analytique des investigations sur site, diverses méthodes rapides de terrain se sont développées ces dernières années. Sur le site X, lors de l'échantillonnage régulier des fosses et des sondages, une mesure de gaz a été effectuée en chaque point. Pour cela, un capot en tôle muni d'un tuyau en caoutchouc a été systématiquement posé sur les fosses immédiatement après leur creusement. Un détecteur à photo-ionisation relié au tuyau en caoutchouc a permis d'évaluer la teneur en polluants organiques volatils totaux à cet endroit ; une mesure de gaz a été effectuée toutes les minutes. Afin de laisser un temps de stabilisation, nous considérons ici la réponse en gaz après 5 minutes.

Bien que le naphtalène soit le seul HAP réellement volatil, les corrélations entre les concentrations sur les fosses et la réponse gaz restent sensiblement identiques avec ce HAP ou les HAP de

HAP	Nap	Acy	Ace	Fle	Ant	Phe	Flt	Pyr	Baa	Cry	Bap	Bbf	Dba	Bkf	Inp	Bgh
ρ	.05	.11	.23	.22	.09	.13	.30	.21	.30	.26	.31	.21	.23	.35	.18	.40

TAB. 7.4 – Site X - Coefficients de corrélation ρ entre fosses et sondages.

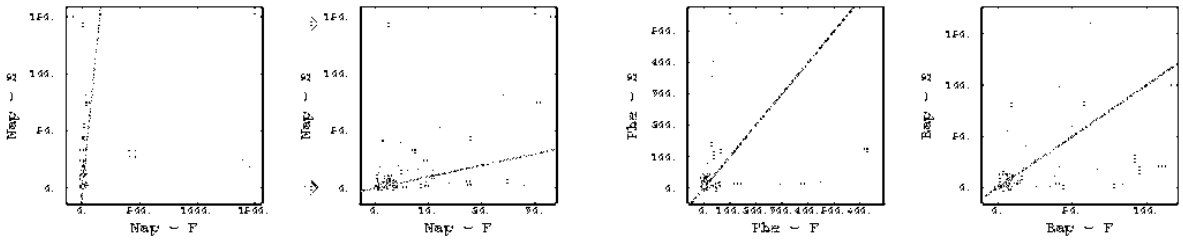


FIG. 7.6 – Site X - Nuages de corrélation entre fosses (abscisses) et sondages (ordonnées). La première bissectrice est indiquée. Pour les deux nuages correspondant au naphtalène, les deux valeurs extrêmes de fosses (carrés) sont supprimées sur le nuage de droite. Echelles en ppm.

nombre de cycles plus élevés - à cause du bon niveau de corrélation entre les concentrations en HAP. Le nuage de corrélation de la figure 7.7, entre la concentration de la somme des 16 HAP sur les fosses et la réponse en gaz après 5 minutes, montre que la réponse en gaz restitue uniquement les signaux de concentration les plus forts.

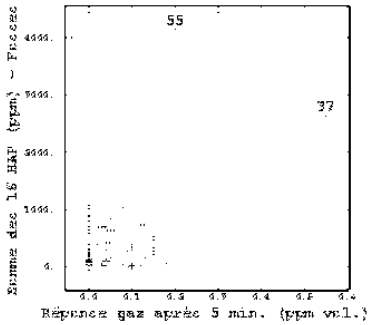


FIG. 7.7 – Site X - Nuage de corrélation entre la réponse en gaz après 5 minutes (en ppm vol.) et la concentration de la somme des 16 HAP sur les fosses.

Aucune réponse en gaz n'est associée à certains points dont l'analyse classique montre pourtant des concentrations en HAP de l'ordre de 1000 ppm - le seuil d'intervention étant fréquemment de 500 ppm pour la somme des 16 HAP. Plusieurs facteurs expliquent cela :

- Les réponses dépendent étroitement des conditions climatiques. En particulier, le taux d'humidité de l'air et la température, qui ont sensiblement varié durant la campagne, semblent être des facteurs contraignants.
- Des problèmes d'étanchéité du système de mesure, essentiellement au niveau du capot recouvrant les fosses, peuvent être à l'origine de pertes de gaz.
- La cokerie ayant cessé toute activité depuis plus de vingt ans, il est probable qu'une proportion considérable des polluants volatils se soit déjà échappée.

Ces mesures de gaz, également envisagées sur le site Y [Wavrer & Jeannée (1998)], ont été rapidement abandonnées, les conditions climatiques et la trop grande taille des fosses conduisant à une absence systématique de réponse malgré une pollution visuellement décelable.

En conclusion, la méthode semble pouvoir fournir une image qualitative des points de forte concentration mais dépend fortement des conditions climatiques et pose des problèmes de mise en œuvre.

7.2.4 Kits chimiques

Deux croix ont été réalisées afin de comparer l'efficacité d'une seconde méthode de terrain, un kit chimique[1] - spectrophotomètre Pastel UV de SECOMAM -, par rapport à une analyse chimique classique [Steyer (2000a, 2000b)].

Bien qu'aucun géoréférencement n'ait été effectué lors de l'échantillonnage des fosses et des sondages, la position des croix par rapport aux fosses et sondages est relativement fiable, grâce aux piézomètres et aux traces encore visibles des premières campagnes (voir figure 8.10).

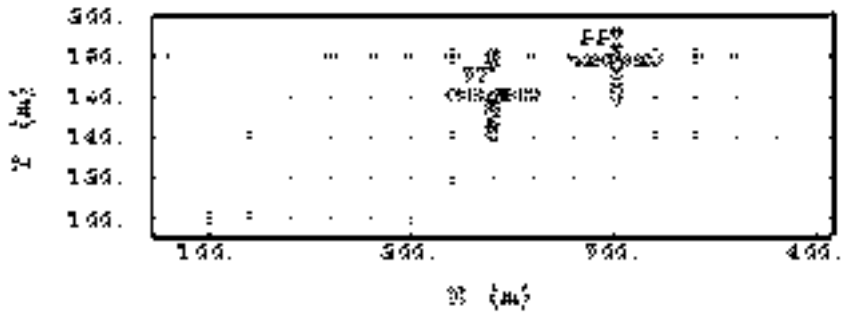
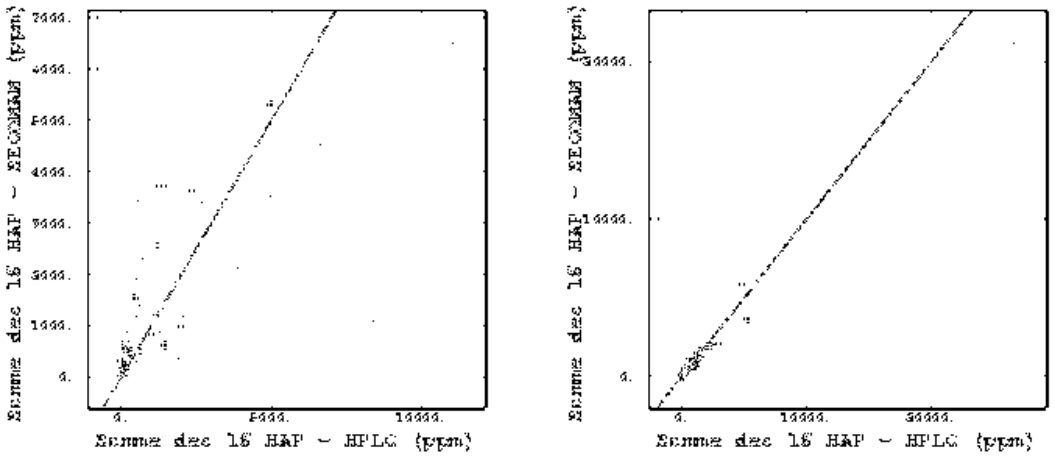


FIG. 7.8 – Site X - Implantation des deux croix de sondages (losanges) sur le site, par rapport aux données régulières.

Ces croix ayant également pour objectif de réduire le pas d'échantillonnage entre des points de la maille régulière à 20 m, nous les étudierons de façon détaillée au chapitre suivant. L'analyse des 16 HAP de l'US EPA porte sur des échantillons provenant de trois profondeurs comprises entre 0.5 et 2 m : 0.5-1 m, 1-1.5 m, 1.5-2 m. A partir des centres des croix, localisés à proximité des deux fosses 37 et 55 présentant les concentrations les plus importantes en somme des 16 HAP, les directions de la grille régulière ont été échantillonnées à 2.5, 5, 7.5, 10, 15 et 20 m dans les quatre directions cardinales ; la présence de dalles en béton et autres obstacles a empêché la réalisation de certains points. Dans la suite, nous appellerons ces deux croix 37 et 55.

[1] Après extraction des HAP par agitation manuelle à l'aide d'un solvant, l'extrait est filtré puis recueilli pour être analysé. La méthode d'extraction et d'analyse de la concentration totale des 16 HAP de l'US EPA varie en fonction du kit.



(a) Croix 37 (b) Croix 55

FIG. 7.9 – Site X - Nuages de corrélation entre la concentration totale en 16 HAP obtenue par HPLC et par kit SECOMAM pour les deux croix investiguées. La première bissectrice est indiquée.

Pour la croix 37 le kit de terrain a tendance à surestimer les concentrations faibles analysées par HPLC et surtout sous-estimer les concentrations les plus fortes (voir figure 7.9(a)). Les techniques de préparation des échantillons, qui diffèrent en fonction de la méthode, expliquent cela : tandis que l'échantillon destiné à l'analyse classique est séché à l'air libre, homogénéisé puis broyé manuellement pour ne conserver que la fraction inférieure à 2 mm, l'échantillon prélevé pour le kit de terrain est séché chimiquement, homogénéisé puis broyé manuellement, avec récupération de la fraction inférieure à 500 μm seule. Il y a donc "perte" d'une partie de l'échantillon. Afin d'en tenir compte, une attention particulière à été portée lors de l'échantillonnage de la croix 55 à l'homogénéisation et au broyage, afin de minimiser le plus possible la quantité de refus. Il en découle une corrélation entre les deux méthodes nettement meilleure pour la croix 55 (voir figure 7.9(b)).

Le kit SECOMAM semble donc donner des résultats sur la concentration totale en 16 HAP tout à fait comparables à ceux obtenus par une méthode classique de laboratoire, ce qui lui confère un intérêt important : il fournit une image précise de la concentration totale en 16 HAP, utilisable pour améliorer ensuite l'estimation des différents HAP analysés classiquement de façon moins intensive, comme cela a été présenté au paragraphe 5.3.2.

7.3 Comparaison avec les campagnes précédentes

Les campagnes antérieures A et B ont mis en évidence des teneurs en HAP élevées dans certaines zones (voir figure 7.10), et qui décroissent avec la profondeur.

L'échantillonnage orienté de l'organisme A n'a pas permis la détection de concentrations fortes au voisinage des réservoirs de la fosse à brai, ce que montrent pourtant bien certains échantillons de B et les fosses et sondages réalisés par le CNRSSP : points 22, 23, 36, 37. Les zones de forte

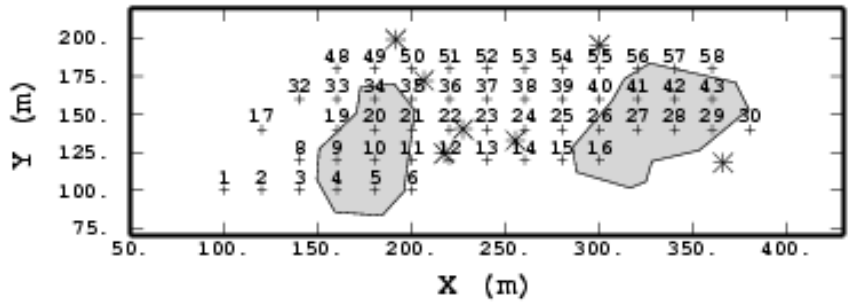


FIG. 7.10 – Site X - Implantation des données régulières de la campagne CNRSSP ; indication des zones reconnues comme étant fortement chargées en polluants organiques par l'organisme A (grisées) et des points échantillonnés par B qui présentent de fortes concentrations en HAP (étoiles).

concentration détectées par le CNRSSP correspondent bien à celles de la campagne systématique de B.

Tout comme sur le site Y, cela souligne le risque de non détection de taches de pollution lié à une reconnaissance guidée par les informations historiques.

7.4 Analyse variographique

Les figures 7.11 et 7.12 reprennent, respectivement pour les fosses et les sondages, les variogrammes expérimentaux moyens par échantillon de 5 HAP et de l'indice Phénol.

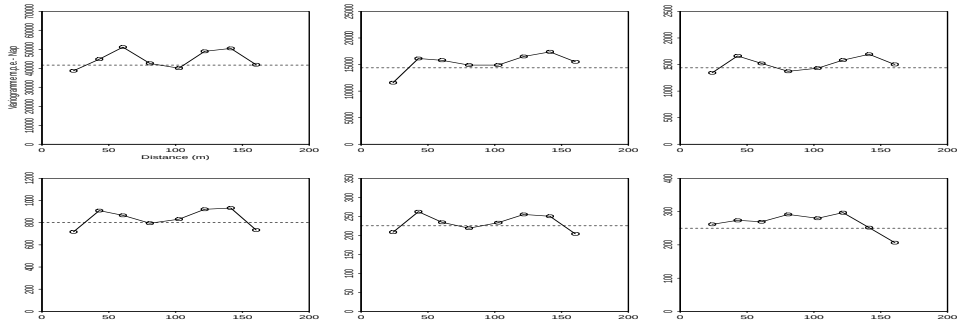


FIG. 7.11 – Site X - Variogrammes m.p.e. de 5 HAP et de Phl sur les **fosses**.

Les HAP les plus légers sont les seuls à présenter une légère structuration. A partir de Phe, les structures deviennent très semblables, reposant uniquement sur le premier point du variogramme expérimental, à 20 m. Ces observations sont compatibles avec celles concernant les fosses du site Y. L'absence de structure sur les sondages est par contre systématique, ce qui différencie sensiblement les deux sites. On note même une décroissance des variogrammes expérimentaux correspondant aux HAP jusqu'à 4 cycles et à Phl.

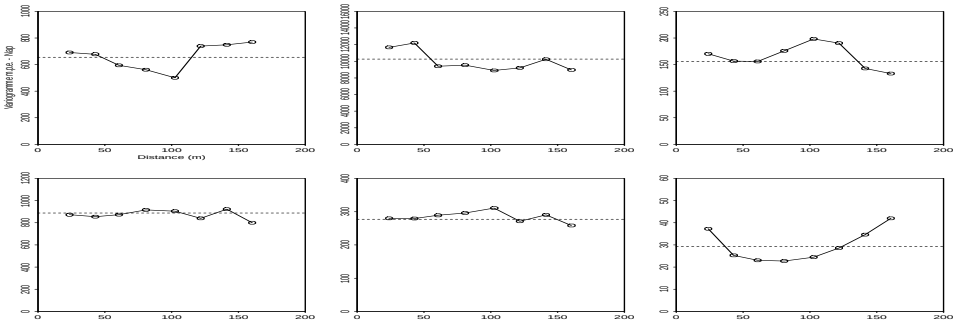


FIG. 7.12 – Site X - Variogrammes m.p.e. de 5 HAP et de Phl sur les **sondages**.

Par ailleurs, l'indice phénol présente une structure pépitique, sur les fosses et les sondages.

Ces résultats ont été confortés par l'analyse des structures sur les transformées de la variable brute. Insistons une nouvelle fois sur le manque de robustesse de ces outils structuraux face à des variables aussi dissymétriques. Par ailleurs, il n'est pas exclu qu'un échantillonnage à 20 m du site Y conduise à des résultats similaires au site X : en prenant un pas de calcul de 20 m pour les variogrammes expérimentaux, seul le premier point est révélateur d'une structure, ce qui est cohérent avec les structures que nous avions observées, de l'ordre de 40 m.

7.5 Modèles

Le modèle du tableau 7.5 est ajusté sur le variogramme moyen par échantillon de Nap sur les fosses (voir figure 7.13).

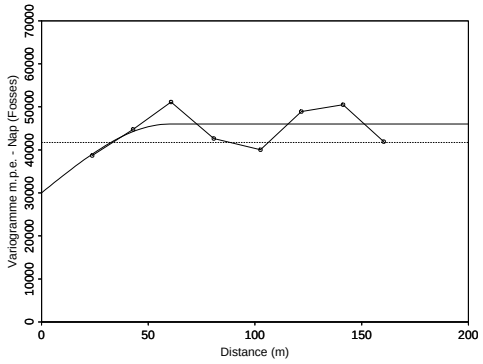


FIG. 7.13 – Site X - Ajustement du variogramme m.p.e. pour Nap sur les fosses.

7.6 Estimations

Nous discutons dans cette dernière section les résultats d'une estimation globale des HAP ainsi qu'un exemple d'estimation locale pour ces variables.

7.6.1 Estimation globale

La grille étant régulière, les valeurs estimées (voir tableau 7.5) à l'aide des modèles ajustés au paragraphe 7.5 sont très proches de la moyenne expérimentale des données ; seuls les points en bordure de champ se voient attribuer des poids de krigeage plus élevés, à l'origine des écarts entre les moyennes.

Les variances d'estimation globale géostatistique sont inférieures à celles obtenues par la formule classique, qui correspond au cas pépitique, et ce malgré une structuration non négligeable. Ces valeurs inférieures s'expliquent par la portée de 60 m, relativement grande par rapport à la taille du champ dans la direction Nord-Sud, qui est de 80 m.

7.6.2 Estimation locale et niveau d'incertitude

Un krigeage ordinaire de blocs est réalisé en voisinage unique (voir figure 7.14) pour la concentration en Nap sur les fosses, à l'aide du modèle de la figure 7.13. On distingue l'influence des deux

HAP	Statistiques			Estimation globale				
				classique		géostatistique		
	m	σ^2	$\frac{\sigma}{m}$	$\frac{\sigma^2}{n}$	$\frac{\sigma}{m\sqrt{n}}$	m_E	σ_E^2	$\frac{\sigma_E}{m_E}$
Nap						$30000 + 15000 * \text{sph}(60\text{m})$		
	43.43	41714	4.70	817.91	0.66	41.73	639.61	0.61
Bap						$520 + 350 * \text{sph}(60\text{m})$		
	24.61	801	1.15	15.70	0.16	24.46	11.50	0.14

TAB. 7.5 – Site X - Statistiques des données de fosses pour 2 HAP : moyenne, variance, coefficient de variation. Estimation globale classique - variance et coefficient de variation - et géostatistique - moyenne, variance et coefficient de variation -, avec indication du modèle utilisé.

fortes concentrations, et l'effet de la moyenne en bordure de champ, qui accroît les valeurs estimées. Les écart-types d'estimation atteignent rapidement des valeurs importantes.

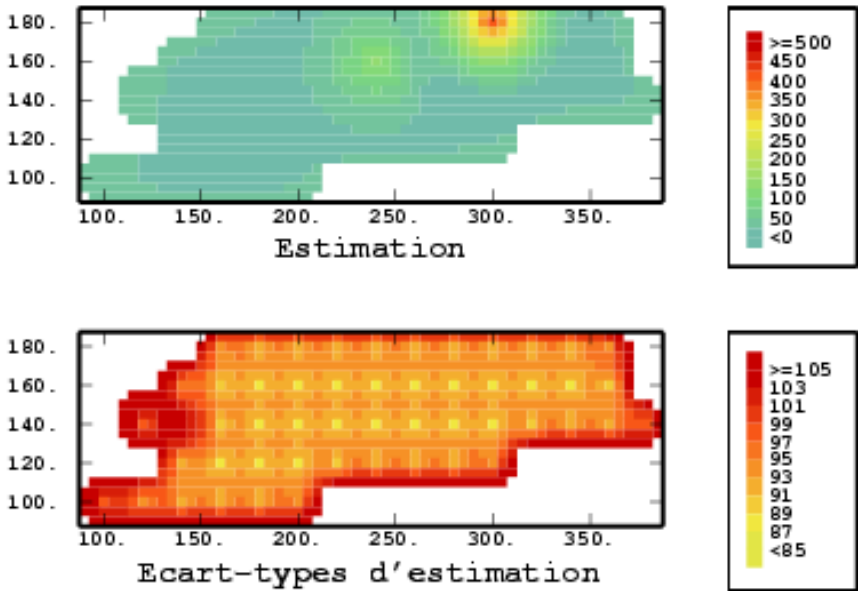


FIG. 7.14 – Site X - Krigeage de blocs et écart-types d'estimation pour Nap sur les fosses.

Forts des enseignements du site Y, nous constatons qu'il est difficile d'aller plus loin pour ces estimations. Par ailleurs, aucune variable qualitative n'étant corrélée aux concentrations, il n'est pas pertinent de les envisager pour en améliorer l'estimation.

7.7 Synthèse

Cette seconde cokerie se distingue du site Y par l'importance des remaniements de sol qui y ont été effectués, ce qu'indique l'étude historique et que confirment les indices qualitatifs. La dispersion des fortes concentrations est plus importante sur ce site, bien que le niveau moyen de concentration en HAP soit, par rapport au site Y, similaire pour les fosses et inférieur pour les sondages. On note l'importante différence d'implantation des fortes concentrations entre fosses et sondages. Par contre, ici encore les HAP présentent entre eux de bonnes corrélations, ces dernières étant d'autant meilleures que les poids atomiques des HAP sont proches.

Les structures expérimentales observées sur les fosses sont comparables à ce qu'elles étaient sur le site Y en ce sens que, exception faite des HAP les plus légers Nap et Ace pour lesquels les résultats sont légèrement meilleurs, la structuration repose entièrement sur le premier point du variogramme expérimental. Il est possible dans ces conditions de mener une estimation globale ainsi qu'une estimation locale, en gardant toutefois à l'esprit l'incertitude liée à l'ajustement d'un modèle de variogramme sur des structures expérimentales aussi peu robustes. Par ailleurs, les cartes d'écart-types d'estimation indiquent la mauvaise précision des estimations dès que l'on s'écarte des données.

Par contre, les structures spatiales des concentrations en HAP sur les sondages sont systématiquement pépitiques. Aucune estimation locale ne peut raisonnablement être menée dans ces conditions, le meilleur estimateur accessible de la concentration restant la moyenne statistique des données.

Chapitre 8

Représentativité des échantillons

Sommaire

Plusieurs campagnes complémentaires menées sur le site X nous permettent de préciser les différentes sources de variabilité des concentrations en HAP : analyses multiples d'un même prélèvement, échantillonnages à différentes échelles inférieures à la maille régulière de 20 m mise en œuvre lors de la première campagne.

Pour un HAP, la concentration analysée à partir d'un prélèvement ponctuel est-elle représentative de la concentration de ce HAP au voisinage du point de prélèvement ? Avant d'aborder de façon détaillée trois problèmes pouvant nuire à la représentativité d'un échantillon, revenons sur deux sources de variabilité plus amont : le remaniement du site et la stratégie d'échantillonnage. Tout d'abord, comme cela a été vu au chapitre précédent, il y a eu un important remaniement du terrain lors du démantèlement des infrastructures, décelable par la présence importante de débris de maçonnerie sur l'ensemble du site et de matériaux rapportés. L'hétérogénéité accrue du sol qui en découle ne peut que nuire à la représentativité des échantillons.

Par ailleurs, le choix de la stratégie d'échantillonnage est fondamental. Plusieurs exercices ont été menés sur ce site afin de comparer différentes stratégies : échantillonnage au hasard, orienté, systématique [Wavrer (1998), Dubourguier (1999)]. L'échantillonnage aléatoire est celui qui prête le plus à caution, ne permettant absolument pas de restituer les "points chauds" détectés par l'échantillonnage systématique. La sélection de points sur la base des plans d'infrastructures ne rend absolument pas compte des points chauds observés avec la maille régulière. Cela est également illustré par les résultats de la campagne antérieure A qui, par un échantillonnage orienté, ne détecte pas une partie des points présentant les concentrations les plus fortes (voir page 133).

Une reconnaissance systématique qui ne se focalise pas *a priori* sur l'information historique du site est donc fondamentale.

8.1 Fraction granulométrique

Tout comme pour le site Y, l'analyse n'est pas menée sur tout l'échantillon mais seulement sur un extrait de la fraction granulométrique inférieure à 2 mm, de masse non constante, que ce soit pour les fosses ou les sondages. Cela reste le cas si nous considérons non plus les masses absolues d'échantillon, mais les pourcentages de la fraction inférieure à 2 mm par rapport à la masse totale de l'échantillon (voir figure 8.1).

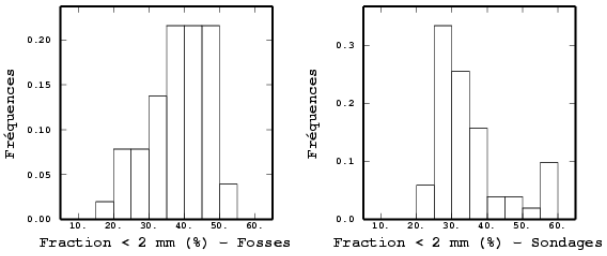


FIG. 8.1 – Site X - Histogrammes pour les fosses et les sondages de la fraction granulométrique inférieure à 2mm, en pourcentage de la masse totale de l'échantillon.

Les figures 8.2 et 8.3 représentent pour 3 HAP les corrélations entre concentration et fraction inférieure à 2 mm (relative).

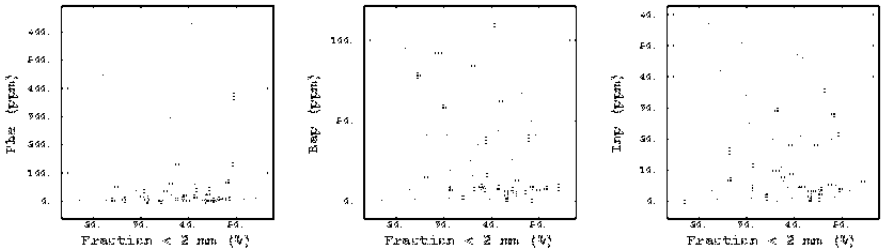


FIG. 8.2 – Site X - Corrélations pour les **fosses** entre la fraction granulométrique inférieure à 2mm, en pourcentage de la masse totale de l'échantillon, et les concentrations de 3 HAP.

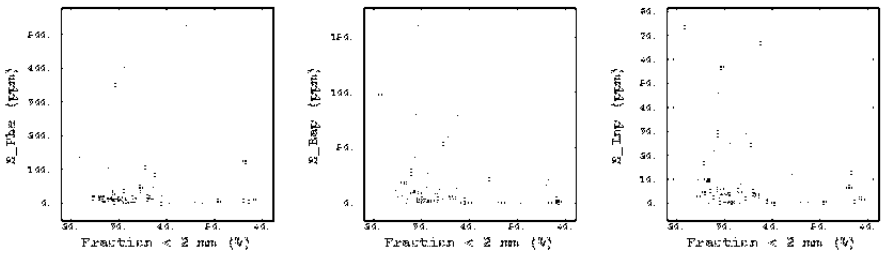


FIG. 8.3 – Site X - Corrélations pour les **sondages** entre la fraction granulométrique inférieure à 2mm, en pourcentage de la masse totale de l'échantillon, et les concentrations de 3 HAP.

Les concentrations les plus élevées ne correspondent pas préférentiellement à certains pourcen-

tages de la fraction inférieure à 2 mm pour les fosses ; pour les sondages par contre, les concentrations les plus importantes correspondent à des fractions inférieures à 2 mm situées entre 25 et 45 %. Cela peut également s'expliquer par le fait que ces pourcentages soient les plus fréquents.

Les histogrammes des accumulations sont semblables à ceux des concentrations, et les corrélations entre concentration et accumulation sont bonnes (voir figures 8.4 et 8.5). Les valeurs des coefficients correspondants sont tous supérieurs à 0.90. Ces corrélations nous poussent à privilégier une poursuite du travail en concentration plutôt qu'en accumulation.

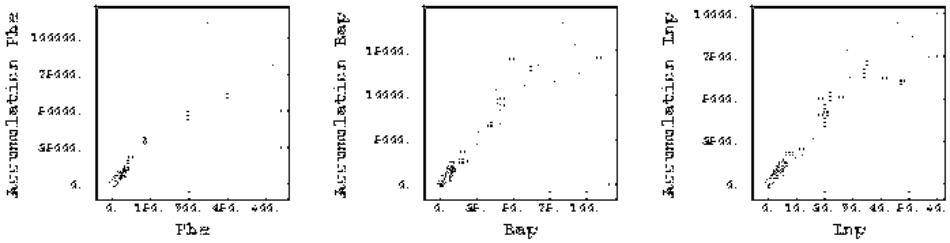


FIG. 8.4 – Site X - Corrélations pour les **fosses** entre accumulations et concentrations pour 3 HAP.

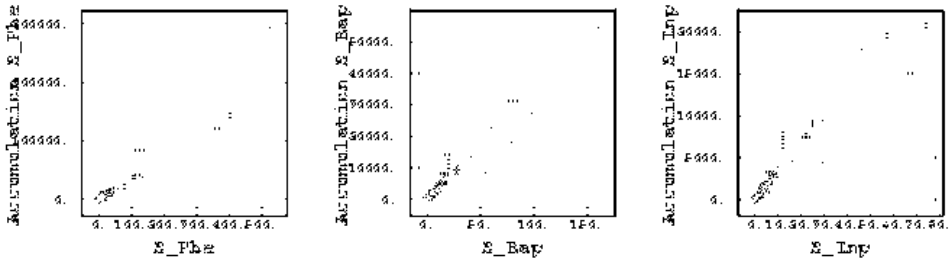


FIG. 8.5 – Site X - Corrélations pour les **sondages** entre accumulations et concentrations pour 3 HAP.

8.2 Préparation des échantillons

Quelques kilogrammes de sol ont été prélevés en *un* point du site, à 0.25 m de profondeur ; après homogénéisation, deux bocaux de 0.5 l ont été constitués. Dans chaque bocal, cinq échantillons ont été prélevés pour analyse. Celles-ci vont permettre d'améliorer la connaissance de la variabilité de concentration issue de la préparation des échantillons. Par ailleurs, on tentera de déterminer sur les structures spatiales observées la part d'effet de pépite imputable à cette "erreur" d'échantillonnage.

8.2.1 Statistiques élémentaires

La gamme des concentrations analysées est très étendue, même au sein d'un même bocal (voir tableau 8.1). La variabilité intra-bocal reste donc très importante. Le bocal 2 présente des concen-

trations en moyenne plus fortes. Comparées aux autres HAP, les concentrations en Nap sont très faibles. Cela peut provenir d'un problème d'extraction, ou plus vraisemblablement d'une volatilisation de ce composé lors de l'homogénéisation du sol.

Bocal	HAP	Min	Max	m	σ	σ/m
1	Nap	8.10	17.30	12.06	3.03	0.25
	Phe	264.00	546.00	355.20	102.89	0.29
	Bap	30.10	73.30	45.80	15.05	0.33
2	Nap	10.90	26.20	18.40	5.39	0.29
	Phe	334.00	674.00	511.00	118.06	0.23
	Bap	43.00	89.00	68.52	15.85	0.23

TAB. 8.1 – Site X - Statistiques par bocal de 3 HAP.

8.2.2 Histogrammes

Sur les histogrammes des deux bocaux (voir figure 8.6), on constate que le bocal 1 présente des valeurs plutôt inférieures à celles du bocal 2, et que les valeurs par bocal sont moins dispersées que l'ensemble. Bien qu'il soit délicat de tirer des conclusions à partir de 5 concentrations, il semble qu'il y ait eu au sein de chacun des deux bocaux une nouvelle homogénéisation, mais que celle-ci ne soit pas suffisante pour que les différentes analyses donnent des concentrations similaires.

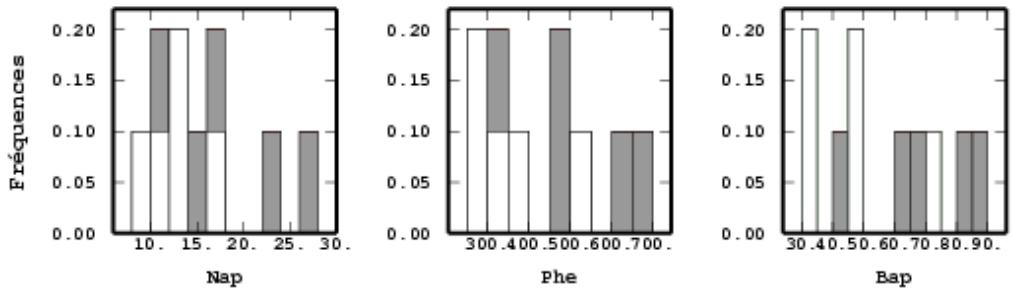


FIG. 8.6 – Site X - Histogrammes des données des deux bocaux. Bocal 1 en blanc.

Regrouper sur un même histogramme les données des fosses échantillonnées à maille régulière et celles des bocaux (voir figure 8.7) confirme la remarque précédente sur la variabilité des données des bocaux. En effet, pour certains HAP ces données de bocaux se répartissent sur la majeure partie de l'histogramme. Donc, l'analyse répétée d'un prélèvement ponctuel permet de restituer une bonne partie de la gamme des concentrations rencontrées sur le site. L'erreur due à l'analyse à proprement parler étant faible pour ce type de produits, cette variabilité est due à la préparation des échantillons, c'est-à-dire essentiellement à l'homogénéisation.

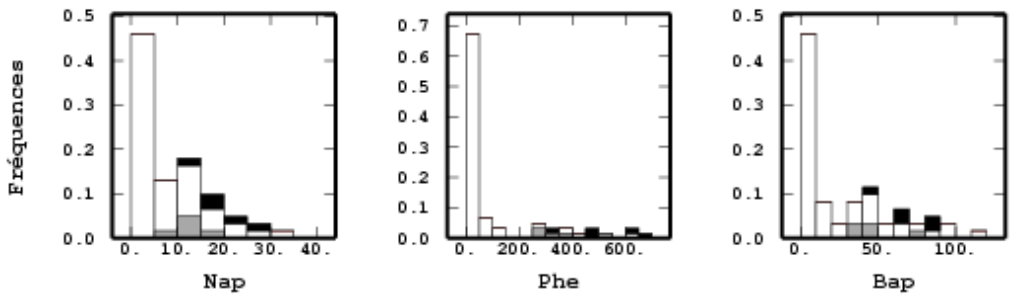
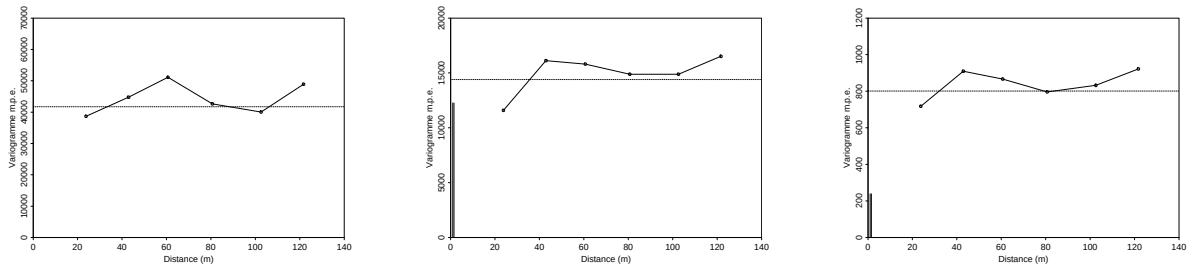


FIG. 8.7 – Site X - Histogrammes des données du site X : données des fosses (blanc) et des deux bocaux (gris et noir). Pour Nap, les deux valeurs fortes de 433 et 1425 ppm (fosses 37 et 55 du site) sont masquées.

8.2.3 Erreur d'échantillonnage et effet de pépite

L'idée est ici de regarder quelle part d'effet de pépite est due à l'erreur d'échantillonnage. Pour cela, on considère les variogrammes moyens par échantillon des données de fosses, et on reporte en ordonnée la variance intra-bocal, i.e. la moyenne sur les deux bocaux des variances des cinq analyses (voir figure 8.8).



Nap Phe Bap

FIG. 8.8 – Site X - Variogrammes moyens par échantillon des concentrations en 3 HAP sur les fosses. Variance intra-bocal reportée en double trait en ordonnée.

Pour Phe et Bap cette variance intra-bocal restitue une part de l'effet de pépite appréciable graphiquement ; l'essentiel de la variabilité proviendrait dans ce cas de l'erreur d'échantillonnage. Ce n'est pas le cas de Nap pour lequel la variance intra-bocal est trop faible pour être visible sur la figure 8.8 ; le faible nombre d'échantillons par bocal et de bocaux analysés, non suffisamment représentatif de l'ensemble du site - on "rate" en particulier les valeurs extrêmes -, peut en être la

cause.

Il est possible de prendre en compte explicitement la variance des erreurs de mesure lorsque celle-ci est connue ou calculable. Tout d'abord, considérons le cas d'une erreur de mesure additive : au lieu d'observer $Z(x_i)$ en x_i nous observons $Z_e(x_i) = Z(x_i) + e_i$, où e_i est une erreur de mesure. Ne connaissant pas cette erreur de mesure, nous considérons e_i comme une variable aléatoire ; si les erreurs sont indépendantes de Z, non-systématiques, mutuellement indépendantes et de même variance, alors le variogramme de Z_e est $\gamma_e(h) = \gamma(h) + \sigma_e^2$, avec $\gamma(h)$ le variogramme de Z et σ_e^2 la variance des erreurs de mesure, qui a donc le sens d'un effet de pépite "artificiel" qui s'ajoute à la structure de Z. Si σ_e^2 est connu, le variogramme de Z peut donc être retrouvé et l'erreur de mesure filtrée [Deverly (1984), Chilès & Delfiner (1999)].

Lorsque cette variance est inconnue, elle est englobée dans l'effet de pépite. Ne possédant que 10 répétitions d'analyse d'un prélèvement provenant d'un seul point du site, il n'est pas raisonnable de déduire de ces données plus qu'une idée de la variance de mesure, et nous nous en tiendrons donc à cela[1].

8.2.4 Calcul théorique de l'erreur fondamentale d'échantillonnage

En procédant à des ré-échantillonnages, un calcul expérimental de la variance des erreurs de mesure peut être effectué. Il est également possible de calculer théoriquement l'erreur fondamentale d'échantillonnage (voir annexe page 169). Cette erreur découle de l'hétérogénéité de constitution de la matière, qui résulte des propriétés physico-chimiques - taille, masse, etc - des particules qui la composent et ne peut être annulée sans modification de son état physique. Cette erreur fondamentale, dont la contribution à la variabilité totale diminue lorsque la taille de l'échantillon prélevé augmente, fournit un minimum incompressible de l'erreur totale d'échantillonnage au-delà duquel il est illusoire de vouloir énoncer une quelconque précision sur une détermination effectuée sur un échantillon.

Afin de calculer cette erreur, un échantillon "gros volume" de plusieurs kilogrammes a été prélevé sur le site [Wavrer (1998)]. Les principaux faciès présents ont été décrits et leur distribution massique au sein de la friche estimée. La figure 8.9 montre par exemple qu'il subsiste avec un échantillon de 10 tonnes 35 % d'erreur fondamentale relative ; les prélèvements réalisés sur les fosses et les sondages ne dépassant pas le kilogramme, il faut être conscient que l'erreur fondamentale relative s'élève à plus de 400 % ! Cela pose de façon assez sévère les limites à donner à l'interprétation d'une analyse portant sur quelques milligrammes provenant d'échantillons de ce type.

8.3 Variabilité à petite distance

Afin d'améliorer la connaissance de la variabilité des HAP à petite échelle, deux campagnes complémentaires ont été réalisées sur le site. Tout d'abord, deux fosses - appelées *stations* dans la suite, pour éviter tout risque de confusion avec les fosses échantillonnées à 20 m - d'environ 2 m x

[1] Le cas plus complexe d'une erreur corrélée à la variable peut être résolu dans un contexte multivariable [Chilès & Delfiner (1999)].

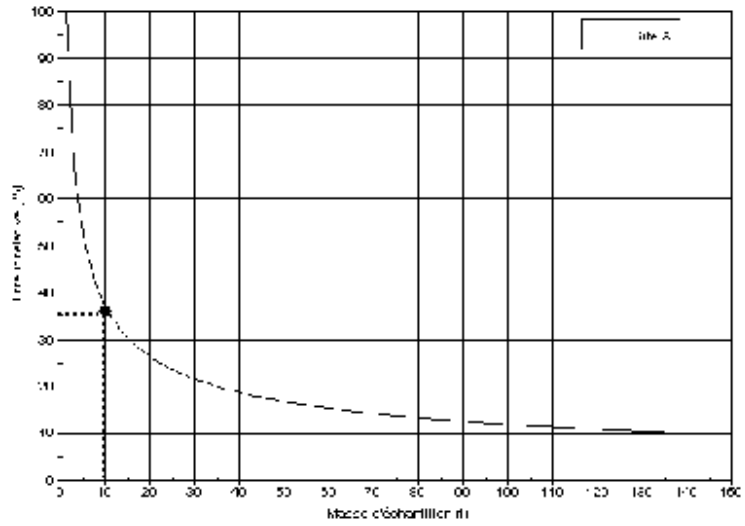


FIG. 8.9 – Site X - Erreur fondamentale relative d'échantillonnage.

2 m x 1 m (L x l x p) ont été ouvertes à la pelle mécanique, afin d'étudier les hétérogénéités à une échelle inférieure au mètre. Par ailleurs, deux croix ont été réalisées à proximité de points "chauds" pour réduire le pas d'échantillonnage entre des points de la maille régulière à 20 m. Les points de prélèvement de ces campagnes complémentaires sont indiqués à la figure 8.10.

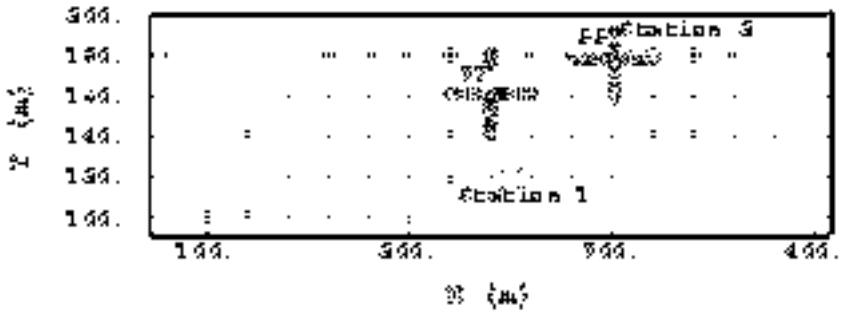


FIG. 8.10 – Site X - Implantation des deux stations (carrés) et des croix de sondages (losanges) sur le site, par rapport aux données régulières.

8.3.1 Stations

Les deux stations, de 2 m et 2.40 m de côté (voir figure 8.11) et d'1 m de profondeur, sont situées à proximité des fosses 14 et 55. Douze échantillons devaient initialement être prélevés par rainurage régulier le long des parois. Sur la station 1, un mur de briques dans un coin a empêché la prise de l'échantillon 7.

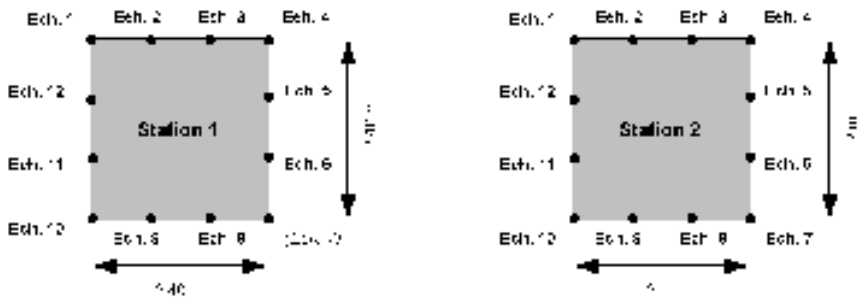


FIG. 8.11 – Site X - Description des deux stations.

8.3.1.1 Statistiques élémentaires

Le niveau de pollution diffère entre les deux stations, ce qui était intentionnel (voir figure 8.2). La première station est globalement très peu polluée, avec comme HAP les plus présents Phe et Flt, ce qui est classique. Seuls les HAP légers présentent des concentrations inférieures aux seuils de détection. La seconde station est bien plus polluée. Aucun HAP n'y présente de données inférieures au seuil de détection.

HAP	Station 1						Station 2				
	Min	Max	m	σ	σ/m	DIS	Min	Max	m	σ	σ/m
Nap	0.01	0.18	0.04	0.05	1.25	27	1.54	16.80	6.38	5.01	0.79
Phe	0.04	2.11	0.90	0.67	0.74	0	3.82	149.31	63.15	48.56	0.77
Bap	0.04	1.05	0.44	0.31	0.70	0	1.84	48.69	11.09	11.75	1.06
Inp	0.01	0.44	0.17	0.12	0.71	0	1.34	27.69	6.56	6.61	1.01
Σ16	0.36	12.90	5.82	3.99	0.69	0	49.22	741.40	331.33	225.57	0.68

TAB. 8.2 – Site X - Statistiques élémentaires des concentrations de quelques HAP sur les deux stations. Pourcentages de données inférieures au seuil de détection (DIS) pour la station 1.

Pour les deux stations, les écarts-types restent relativement élevés, et les coefficients de variation sont plus importants sur la station la plus polluée, excepté pour les HAP les plus légers, très faiblement détectés sur les fosses, d'où des moyennes très faibles. Pour la station peu polluée, ces écart-types doivent être mis en perspective avec l'erreur fondamentale d'échantillonnage, très élevée. La gamme des concentrations obtenues en des points séparés par d'aussi petites distances est très importante : rapport de 1 à 100 parfois entre les valeurs fortes et faibles. Cette variabilité est cohérente avec les effets de pépite observés. Les corrélations restent très bonnes sur ces stations entre les HAP de nombres de cycles proches.

8.3.1.2 Implantation

En ordonnant les échantillons de chacune des stations en fonction de leur numéro (voir figure 8.12), ce qui permet d'observer les variations de concentrations lorsque l'on se déplace le long des parois de la fosse, on constate l'existence non systématique de transitions entre valeurs fortes et

valeurs faibles. Par exemple, pour Bap sur la station 1, le passage de la valeur forte de l'échantillon 3 à la valeur faible de l'échantillon 6 se fait avec une transition. Au contraire, pour le Bap sur la station 2, les échantillons 1 et 3 qui entourent la concentration extrêmement forte de l'échantillon 2 présentent tous les deux des concentrations faibles. En considérant un seuil à 10 ppm souvent utilisé pour les HAP, on note que les concentrations analysées sont supérieures à ce seuil en certains points, et inférieures en des points voisins. Comment dans de telles conditions sélectionner les zones où la concentration est supérieure au seuil à partir d'un échantillonnage ponctuel effectué à une maille de 20 m ?

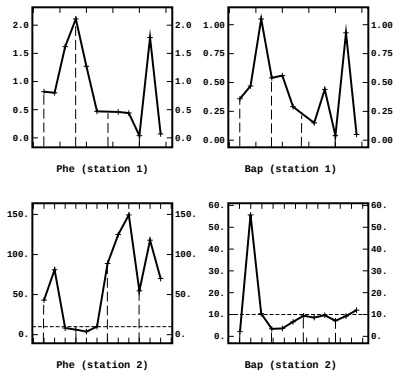


FIG. 8.12 – Site X - Concentrations en Bap et Phe en fonction de l'échantillon. Les échantillons (de 1 à 6 et de 8 à 12 pour la station 1) sont en abscisse. Les traits verticaux indiquent les coins des fosses, les traits horizontaux le niveau de concentration égal à 10 ppm.

8.3.1.3 Histogrammes

Regroupons toutes les données du site X provenant de l'horizon d'échantillonnage de surface : les données de fosses et celles des deux stations (voir figure 8.13). La station 1 est représentative de la partie la plus faible des histogrammes alors que la station 2 s'étale beaucoup plus, recouvrant la plus grande partie des histogrammes, exceptés les queues de distribution.

Il subsiste donc à une échelle inférieure au mètre une variabilité très forte des concentrations en HAP.

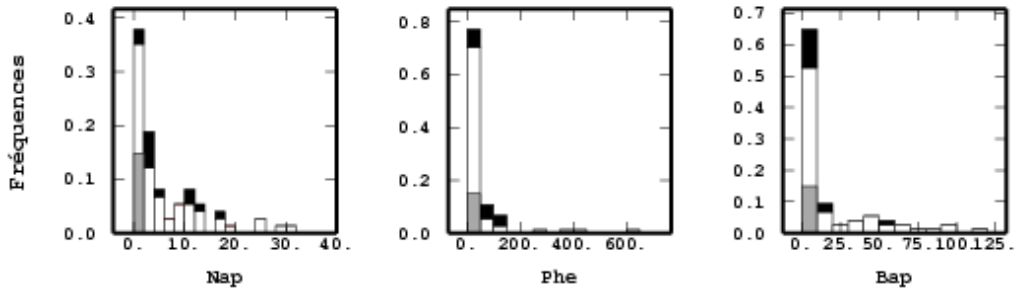


FIG. 8.13 – Site X - Histogrammes de toutes les données : régulières (en blanc) et provenant des stations (station 1 en gris, station 2 en noir).

8.3.2 Croix

8.3.2.1 Statistiques élémentaires

Le tableau 8.3 reprend quelques statistiques sur chacune des croix de l'accumulation en Bap et en somme des 16 HAP entre 0.5 et 1.5 m. Cette accumulation permet d'être cohérent avec le support des sondages réguliers à 20 m, échantillonnés entre 0.5 et 1.5 m. La gamme de variation des concentrations, explicable par l'implantation dans des zones de fortes concentrations, est importante.

Elément	Croix 37				Croix 55			
	Min	Max	m	σ	Min	Max	m	σ
Benzo(a)pyrène	0.0	648.0	119.5	190.8	0.6	676.5	44.8	138.6
Somme des 16 HAP	5.0	11029.0	1435.1	2529.5	35.5	15773.9	1377.0	3203.4

TAB. 8.3 – Site X - Statistiques des concentrations en benzo(a)pyrène et somme des 16 HAP entre 0.5 et 1.5 m sur les croix situées à proximité des points 37 et 55 (en mg/kg sec).

8.3.2.2 Implantation

La carte d'implantation des données, en illustrant la répartition des valeurs fortes sur chacune des deux croix, montre l'hétérogénéité de la distribution des concentrations (voir figure 8.14).

Verticalement, aucune tendance systématique des concentrations n'a pu être observée : elles augmentent avec la profondeur en certains points, décroissent ou stagnent en d'autres points, soulignant encore l'hétérogénéité de distribution de la matière, même à cette échelle.

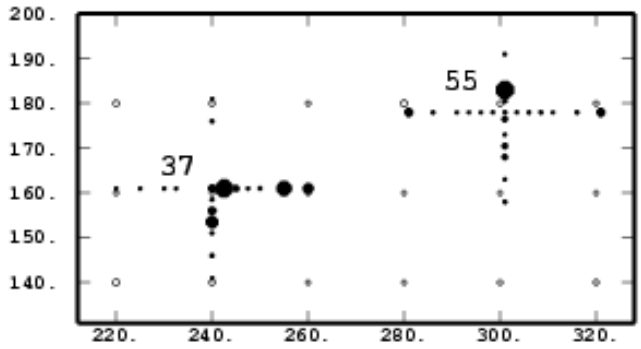


FIG. 8.14 – Site X - Représentation proportionnelle des concentrations en somme des 16 HAP des croix (cercles pleins) et des points de la grille situés à proximité.

8.3.2.3 Comparaison des différentes échelles

La seconde station finement échantillonnée, présentée au paragraphe 8.3.1, est localisée comme la seconde croix de sondages à proximité du point 55. Il est intéressant de comparer les variations des concentrations à différentes échelles dans cette zone : le sondage du point 55 lui-même, la station 2, la croix centrée en ce point et finalement les données sur l'ensemble du site. Les concentrations - ponctuelle dans le cas du sondage 55 et moyennes ensuite - varient en fonction du support (voir tableau 8.4). Par ailleurs, la variance augmente lorsque l'on passe de la station à la croix ; cela est logique, car c'est une variance de dispersion, dont on attend qu'elle augmente avec la taille du champ servant à son calcul avant de se stabiliser éventuellement. Pourtant, cette variance décroît lorsque l'on passe aux données régulières à 20 m sur l'ensemble du site, ce qui indique bien que l'on a manqué une partie du phénomène en échantillonnant à cette échelle.

Elément	Sondage 55	Station 2	Croix 55	Données à 20 m
Benzo(a)pyrène	21.0	11.5 (13.6)	44.8 (138.6)	17.6 (29.8)
Somme des 16 HAP	487.4	331.5 (225.5)	1377.0 (3203.4)	362.8 (580.2)

TAB. 8.4 – Site X - Concentrations en benzo(a)pyrène et somme des 16 HAP pour le sondage 55 en ppm, moyennes (écart-types) pour différentes échelles d'échantillonnage.

8.3.2.4 Structures spatiales

Aucune structure spatiale n'a pu être mise en évidence pour la croix 37, que ce soit pour les différents HAP ou leur somme. Il n'en est pas de même pour la croix 55. L'absence de distinction géologique nette entre les différents niveaux d'une part et de relation entre concentration et profondeur d'autre part a poussé à un travail en accumulation. En travaillant ainsi sur la somme pour chaque niveau des produits entre concentration et taille du support d'échantillon, nous constatons sur la figure 8.15 la présence d'une structure spatiale associée à un effet de pépite qui est de l'ordre de 50 % de la variance pour la somme des 16 HAP - ces résultats sont similaires à ceux HAP

par HAP. L'existence de cette structure est confirmée par un calcul en logarithme translaté et en transformée gaussienne, qui conduisent à des structures pour lesquelles l'effet de pépite descend en dessous de 10 %. Il existe donc pour cette croix une structure de portée comprise entre 10 et 20 m. Le nombre de points utilisés doit cependant inciter à considérer ces résultats avec prudence.

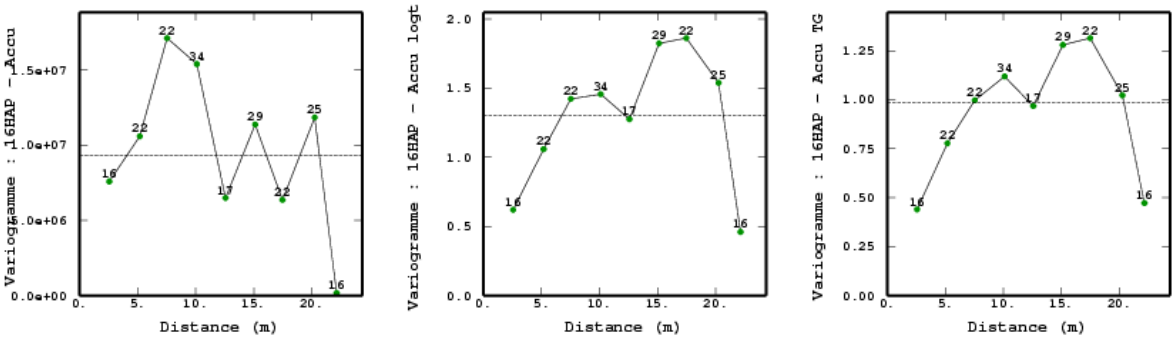


FIG. 8.15 – Site X - Variogrammes pour la croix 55 de l'accumulation des concentrations en somme des 16 HAP : variable brute, logarithme translaté et transformée gaussienne.

8.4 Synthèse

Ce site illustre tout d'abord l'importance d'un échantillonnage à maille régulière pour l'investigation de ce type de site, les risques inhérents aux autres stratégies pouvant s'avérer importants.

La représentativité de la concentration analysée en HAP est discutable ; l'analyse multiple d'un prélèvement ponctuel a montré que l'homogénéisation d'un prélèvement ponctuel potentiellement hétérogène est source d'une variabilité importante des concentrations en HAP. Ce point, qui constitue un des problèmes majeurs de l'investigation de tels sites, explique en outre l'intérêt suscité par des méthodes le contournant en analysant par exemple les gaz intersticiels du sol plutôt que le sol lui-même. Ces méthodes ne sont néanmoins pas exemptes de difficulté, leur calibrage et une corrélation correcte avec la concentration en HAP du sol correspondant étant nécessaires.

Finalement, deux campagnes ont montré l'importance de la variabilité des concentrations même à petite échelle. Les stations illustrent que la variabilité des teneurs est accrue dans les zones de fortes concentrations, ce qui remet en question les résultats d'un échantillonnage ponctuel dans de telles conditions. Un échantillonnage composite pourrait être envisagé, mais celui-ci s'exposerait d'autant plus aux difficultés d'homogénéisation discutées ci-dessus. Les conclusions de l'échantillonnage à 2.5 m sur deux croix de sondages se sont révélées étroitement liées à l'implantation de ces resserrements d'échantillons : une structuration spatiale non négligeable des concentrations a été constatée pour l'une des croix, tandis que l'hétérogénéité de distribution des concentrations sur la seconde croix a systématiquement conduit à des effets de pépite purs. Face à un site aussi hétérogène, ce dernier point montre la difficulté de l'implantation de croix de sondages et surtout le risque lié à la sélection de zones de dépassement de seuils de pollution !

Cinquième partie

Conclusion

Conclusion

"Qu'apporte la géostatistique à la caractérisation d'une pollution industrielle de sols ?" Partant de cette question formulée lors du démarrage du travail, des recommandations peuvent être apportées à la question plus générale : **"comment s'y prendre face à un site industriel pollué ?"**

Concernant les développements géostatistiques, pour l'étude de variables très dissymétriques et relativement peu informées, le variogramme déduit de la covariance non centrée a été écarté en raison de l'hypothèse de stationnarité d'ordre deux requise, malgré une meilleure robustesse par rapport au variogramme classique. Le variogramme moyen par échantillon s'est avéré être l'outil le plus approprié. On a privilégié dans ce travail des outils simples, permettant d'alimenter la réflexion du praticien et de lui fournir des éléments de compréhension et d'action utilisables.

Echantillonnage d'un site pollué

L'efficacité d'une caractérisation de pollution dépend en premier lieu de l'ensemble des étapes situées entre la conception d'un plan d'échantillonnage et l'obtention des analyses des concentrations. Seul un échantillonnage régulier, non préférentiel, peut fournir une reconnaissance homogène du site. Il gagne à être si possible effectué en plusieurs étapes. La maille d'échantillonnage peut être choisie assez large, en fonction du budget alloué à l'investigation, afin de permettre un échantillonnage complémentaire tirant parti du premier. Bien que l'objectif principal soit l'analyse chimique classique des échantillons de sol, le relevé d'indices organoleptiques et qualitatifs est à recommander, ainsi que l'utilisation éventuelle de méthodes semi-quantitatives ; nous y reviendrons ci-dessous.

La concentration analysée sur chaque prélèvement doit être représentative de la concentration en place. L'homogénéisation de sols hétérogènes reste un problème fondamental ; néanmoins, il est possible d'approcher la variabilité des concentrations qui en découle par la réalisation d'analyses multiples en quelques points. La variabilité des concentrations à petite distance pouvant être importante pour des sols hétérogènes, il est nécessaire de préciser le comportement de la variable à cette échelle en resserrant l'échantillonnage par une ou deux croix de sondages. On évitera si possible leur implantation dans des zones particulièrement polluées - par exemple si des traces d'une pollution élevée sont visibles.

La difficile combinaison de volumes échantillonnés différents doit inciter le praticien à choisir autant que possible un protocole - notamment un support - d'échantillonnage identique lors de toute la campagne. En outre, rappelons que le tamisage à 2 mm de l'échantillon avant analyse - norme AFNOR NF ISO 13877 - conduit à des volumes d'échantillon à analyser variables, et écarte d'emblée une proportion non négligeable de la pollution - de l'ordre de 25-30 % dans certains

cas, par exemple adsorbée sur des briques. Un échantillonnage s'arrêtant à une profondeur fixée a priori trop faible ou dès qu'une concentration négligeable est rencontrée est inadapté si un calcul de volume de sol pollué est envisagé.

L'échantillonnage doit permettre de répondre aux questions suivantes :
- Quels sont les produits présents ?
- Les concentrations fortes se répartissent-elles de façon homogène sur le site ? Le cas échéant, étudier séparément chaque population homogène peut s'avérer nécessaire.
- Existe-t-il des corrélations entre les différents produits ?
- Que valent les concentrations moyennes en polluants par rapport aux seuils d'intervention préconisés ?
- Les polluants présentent-ils une structure spatiale à l'échelle de reconnaissance, éventuellement par zone géographique homogène ?

Dans notre cas, les HAP présentent des concentrations statistiquement et spatialement très contrastées - typiquement une majorité de concentrations faibles, voire inférieures à un seuil de détection et quelques valeurs extrêmement élevées. Les corrélations entre les concentrations des différents HAP sont très bonnes, en particulier entre HAP de poids atomiques proches. Au vu de cela, une quantification des concentrations des sommes de HAP par nombre de cycles pourrait être suffisante à leur étude. Les HAP présentent des structures spatiales nettes dans certains cas, malgré une variabilité à petite distance importante et systématique. A titre d'exemple, dans le cas particulier du site Y, une maille à 20 m contenant la croix de sondages centrale, soit 28 à 30 points expérimentaux au lieu des 52 réalisés, aurait dans un premier temps suffi à détecter l'existence d'une structure spatiale pour les concentrations en HAP.

Echantillonnage complémentaire

Grâce aux réponses apportées par la première campagne, cette seconde phase permet d'optimiser les coûts alloués à l'investigation. En cas d'absence de structure spatiale, affiner l'échantillonnage n'est vraisemblablement pas utile. Au contraire, si une structure spatiale des concentrations est mise en évidence, un nouvel échantillonnage plus fin du site, par zones homogènes éventuellement, va améliorer la précision des estimations géostatistiques.

S'il existe une corrélation entre différents produits polluants, il est intéressant économiquement autant qu'en termes de connaissance de la pollution que l'échantillonnage complémentaire se focalise sur certains d'entre eux. En effet, si deux polluants sont corrélés, il est préférable d'en échantillonner un seul de façon beaucoup plus dense ; grâce à leur corrélation, l'estimation du second y gagnera en précision par rapport à l'analyse pour une même configuration d'échantillonnage des deux polluants : par exemple, à même coût d'échantillonnage, multiplier par 4 la densité d'échantillonnage du premier polluant peut être plus instructif que de multiplier par deux la densité d'échantillonnage des deux polluants simultanément. A coût d'analyse égal, on pourra privilégier l'échantillonnage du produit le plus toxique ; si les coûts d'analyse diffèrent, on affinera logiquement l'échantillonnage du polluant le moins coûteux.

Utilisation d'informations auxiliaires

L'échantillonnage complémentaire est également conditionné par l'existence de variables qualitatives et semi-quantitatives corrélées aux concentrations, dont la reconnaissance fine peut améliorer la précision des estimations. Ces informations sont de trois types : informations historiques, variables organoleptiques et qualitatives, mesures semi-quantitatives - analyses de gaz, kits chimiques. Le coût financier de l'investigation étant un facteur limitant, un lien entre ces informations peu coûteuses et les concentrations en polluants incitera à leur utilisation. Néanmoins, elles ne peuvent en aucun cas se substituer au premier échantillonnage systématique par analyses classiques :

- Les limites d'une utilisation a priori des informations historiques pour guider l'implantation des échantillons ont été illustrées sur les deux sites ; le manque d'informations relatives au fonctionnement de ces sites anciens et à leurs infrastructures entraîne en effet un risque important de non détection de certaines taches de pollution.
- Concernant les données qualitatives, il est primordial de s'assurer de l'effective corrélation avec les concentrations en HAP. La combinaison de différents indices qualitatifs doit également être envisagée.
- La mise en œuvre de méthodes semi-quantitatives ne peut se passer d'un recalage avec les concentrations en HAP dans les sols, l'estimation ultime et la sélection de zones à dépolluer concernant ces derniers, et non une analyse de gaz ou une réponse par kit chimique, à moins que la sélection ne porte sur la concentration totale des 16 HAP. Or, une concentration de 500 ppm en somme des 16 HAP pouvant correspondre à des risques de toxicité fort différents selon sa composition, s'intéresser uniquement à cette somme est dangereux d'un point de vue sanitaire.

L'analyse de l'historique - disponible avant l'investigation - et des autres variables auxiliaires - échantillonnées dès la première campagne - doit répondre aux questions suivantes :

- Les concentrations des produits polluants sont-elles cohérentes avec les informations historiques ? Si c'est le cas, ces informations historiques gagnent à être utilisées ; elles permettent de délimiter l'extension de certaines zones de forte concentration, et d'éviter des prélèvements complémentaires inutiles à proximité de points que l'on saurait être fortement pollués, ou au contraire sans risque de pollution.
- Existe-t-il une corrélation entre les analyses classiques des concentrations et les variables qualitatives ou semi-quantitatives ? Si c'est le cas, un échantillonnage dense de ces variables auxiliaires précédant l'échantillonnage complémentaire permettra d'orienter celui-ci et d'affiner à moindre coût la connaissance de l'état de pollution du site ; cela a été illustré sur le site Y, où la présence de goudron et une combinaison des différents indices se sont révélées être bien corrélées aux concentrations ; c'est également le cas des kits chimiques.

En fonction des réponses et du rapport entre coût d'échantillonnage et d'analyse et information apportée, l'échantillonnage complémentaire du sol en vue d'analyses classiques est réalisé.

Estimation des concentrations en place et sélectivité

En cas d'absence de structure spatiale même à petite distance, aucune sélectivité n'est possible. Cela signifie que la séparation entre teneurs fortes et faibles à petite distance est illusoire. La concentration moyenne constitue alors la meilleure estimation accessible de la concentration en place, et elle seule peut orienter la décision d'intervenir sur l'ensemble du site, en fonction des risques qui en découlent.

L'existence de structures spatiales permet l'estimation locale des concentrations en place utilisant les corrélations avec les variables auxiliaires. Celle-ci doit préciser la localisation des différents polluants par classes de teneurs, ce qui en fait un guide précieux pour l'étude de risque. Les méthodes d'estimation utilisées donnent une indication sur l'incertitude de l'estimation des concentrations en place.

Pour la délimitation des zones à traiter, les informations historiques sont utilisées si elles sont cohérentes avec les concentrations analysées. En cas de traitement in situ, l'estimation des concentrations en place fournit déjà une répartition de la pollution par classes de teneurs. Lorsqu'une sélection des terres est nécessaire, on a rappelé le risque d'un seuillage de la carte krigée et la nécessité d'une modélisation de la loi spatiale qui conduise au calcul de probabilités de dépassement du seuil d'intervention en chaque point du site ou sur des blocs ; la taille du support utilisé lors de la dépollution doit être prise en compte dans ce calcul.

Dans le cas d'une dépollution sélective, la structure spatiale permet de proposer une maille d'échantillonnage en dépollution. L'évaluation des probabilités de dépassement de seuil montre que toute sélection de zones à dépolluer - excepté en cas d'excavation de l'intégralité du site - engendre une probabilité non nulle de laisser en place des blocs de concentration supérieure au seuil choisi. L'estimation de cette probabilité constitue en ce sens un outil d'aide à la décision important.

Perspectives

S'il vise à fournir au praticien quelques recommandations et outils utiles pour aborder un site industriel pollué et enrichir sa compréhension de la pollution, ce travail n'épuise pas l'ensemble des questions liées à la caractérisation de pollutions industrielles de sols.

Mettre en œuvre et valider les conseils méthodologiques apportés sur des sites sujets au même type de pollution est nécessaire. Notamment, quelle économie permet en pratique l'échantillonnage dense d'informations qualitatives liées à la pollution ? Par ailleurs, la réflexion sur l'échantillonnage requis face à un site - densité, rapport entre effort d'échantillonnage et information apportée - doit être poursuivie. La question de l'homogénéisation, qui dépasse largement le cadre de ce travail, reste fondamentale. Comment l'améliorer, et à quel coût ? Se tourner vers des méthodes permettant une meilleure représentativité des concentrations en place est envisageable, mais le recalage avec les analyses chimiques classiques est inévitable. Dans quelle mesure finalement la méthodologie diffère-t-elle si l'on se tourne vers d'autres types de pollution de sols : solvants chlorés, pollutions métalliques ?

Bibliographie

Bibliographie

[1] Alfaro Sironvalle, M. (1979). *Etude de la robustesse des simulations de fonctions aléatoires*. Thèse de Doct.-Ing. en Sciences et Techniques Minières, Option Géostatistique, Ecole des Mines de Paris.

[2] BRGM (2000a). Gestion des sites (potentiellement) pollués. Evaluation simplifiée des risques - Version 2. BRGM Editions.

[3] BRGM (2000b). Gestion des sites pollués. Evaluation détaillée des risques - Version 0. BRGM Editions.

[4] Belkessam, L. (1999). Détermination des HAP contenus dans les sols. Etude de l'influence de la préparation des échantillons et de l'extraction. Rapport CNRSSP 1999/10, Douai.

[5] Bez, N. (1997). *Statistiques individuelles et géostatistique transitive en écologie halieutique*. Thèse de Docteur en Géostatistique, Ecole des Mines de Paris.

[6] Bouchez, M., D. Blanchet, F. Haeseler et J.-P. Vandecasteele (1996). Les hydrocarbures aromatiques polycycliques dans l'environnement : propriétés, origine, devenir. *Revue de l'IFP*, **51(3)**, pp. 407-419.

[7] Bourgine, B. et I. Niandou (1993). Conception d'une croix de sondages pour une identification optimale du variogramme. *Cahiers de Géostatistique*, Fasc. 3, Ecole des Mines de Paris, pp. 167-180.

[8] Cazier, F., M. Duval, H.C. Dubourguier and E. Degans (1997). Identification of pollutants in soils of former coal industries by GC/MS and HPLC/MS. In *Proceedings of the International Symposium on Environmental Biotechnology*, Anvers.

[9] Chautru, J.-M. (1989). *Modélisation probabiliste de gisements de nodules polymétalliques*. Thèse de Docteur en Géostatistique, Ecole des Mines de Paris.

[10] Chauvet, P. (1999). *Aide-mémoire de géostatistique linéaire*. Les Presses de l'Ecole des Mines.

[11] Chilès, J.P. (1995). Quelques méthodes de simulation de fonctions aléatoires intrinsèques. *Cahiers de Géostatistique*, Fasc. 5, Ecole des Mines de Paris, pp. 97-112.

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

[13] Cressie, N. (1991). *Statistics for spatial data*. Wiley, New-York. Reprinted (1993).

[14] Demange, C., C. Lajaunie, C. Lantuéjoul and J. Rivoirard (2000). Global recoverable reserves : testing various changes of support models on uranium data. In *Geostatistical case studies*, G. Matheron and M. Armstrong (eds.), Reidel, Dordrecht, pp. 135-147.

[15] Demougeot-Renard, H. and Ch. de Fouquet (2000). Managing heterogeneous sampling data for geostatistical estimations of benzo(a)pyrene concentrations in a former gas works. In

GeoENV 2000, P. Monestiez, D. Allard and R. Froidevaux (eds.), Kluwer, Dordrecht. Sous presse.

[16] Deverly, F. (1984). *Echantillonnage et géostatistique*. Thèse de Doct.-Ing. en Sciences et Techniques Minières, Option Géostatistique, Ecole des Mines de Paris.

[17] Dubourguier, H.C. (1999). Définition d'une stratégie d'investigation pour les sites industriels pollués - Rapport final. Pôle de Compétence Nord - Pas de Calais sur les Sites et Sols Pollués.

[18] Fetter, C.W. (1993). *Contaminant Hydrogeology*. Macmillan, New York.

[19] Fouquet, Ch. de and D. Mandallaz (1992). Using geostatistics for forest inventory with air cover : an example. In *Geostatistics Tróia '92*, A. Soares (ed.), Kluwer, Dordrecht, Vol. 2, pp. 875-886.

[20] Fouquet, Ch. de (1996). Influence de la méthode d'estimation et de la maille de reconnaissance sur la quantification des pollutions. In *Echantillonnage et environnement*, J. Nicolas (ed.), Cebedoc, Liège, pp. 39-63.

[21] Fouquet, Ch. de (1999). Représentativité spatiale des mesures de pollution dans les sols : des échantillons aux estimations. *Les techniques de l'industrie minérale*, **3**, pp. 25-34.

[22] Freulon, X. (1992). *Conditionnement du modèle gaussien par des inégalités ou des randomisées*. Thèse de Docteur en Géostatistique, Ecole des Mines de Paris.

[23] Gailland, D. et Ch. Gissler (1994). Réhabilitation des sols pollués ; la fragilité des évaluations. *Environnement & Technique / Info-Déchets*, Juin 1994, **137**, pp. 37-40.

[24] Gordon, A.D. (1981). *Classification*. Chapman and Hall, London.

[25] Greenacre, M. (1984). *Theory and applications of correspondence analysis*. Academic Press, London.

[26] Guiblin, Ph., J. Rivoirard et E.J. Simmonds (1995). Analyse structurale de données à distribution dissymétrique : exemple du hareng écossais. *Cahiers de Géostatistique*, Fasc. 5, Ecole des Mines de Paris, pp. 137-159.

[27] Guiblin, Ph. (1997). *Analyse géostatistique de campagnes (acoustique et chalutage) sur le hareng écossais*. Thèse de Docteur en Géostatistique, Ecole des Mines de Paris.

[28] Gy, P. (1988). *Hétérogénéité, échantillonnage, homogénéisation : ensemble cohérent de théories*. Masson, Paris (Mesures physiques).

[29] Halleux, I. et Ph. Nix (1991). L'optimisation des schémas d'échantillonnage : nouveaux concepts, nouvelles méthodes. In *Techniques et stratégies de l'échantillonnage de l'environnement*, Miremont.

[30] Hu, L.-Y. (1988). *Mise en œuvre du modèle gamma pour l'estimation de distributions spatiales*. Thèse de Docteur en Géostatistique, Ecole des Mines de Paris.

[31] Hugon, J.-P. et P. Lubek (2000). *Rapport d'expertise et de propositions sur le dispositif juridique et financier relatif aux sites et sols pollués*, Ministère de l'économie, des finances et de l'industrie.

[32] Jeannée, N. et Ch. de Fouquet (1999). Informations qualitatives et inférence de la structure spatiale. In *Comptes-rendus des journées de géostatistique*, D. Renard (coord.), rapport N-8/00/G, Ecole des Mines de Paris, pp. 91-108.

[33] Jeannée, N. and Ch. de Fouquet (2000a). Characterization of soil pollutions from former coal processing sites. In *Geostats 2000*, W. Kleingeld and D. Krige (eds.). Sous presse.

[34] Jeannée, N. and Ch. de Fouquet (2000b). Which experimental variogram for the structural inference. In *GeoENV 2000*, P. Monestiez, D. Allard and R. Froidevaux (eds.), Kluwer, Dordrecht. Sous presse.

[35] Journel, A.G. and Ch.J. Huijbregts (1978). *Mining Geostatistics*. Academic Press, London.

[36] Lajaunie, C. (1993). *L'estimation géostatistique non linéaire*. Cours C-152, Centre de Géostatistique, Ecole des Mines de Paris.

[37] Lantuéjoul, C. (1990). *Cours de sélectivité*. Cours C-140, Centre de Géostatistique, Ecole des Mines de Paris.

[38] Lantuéjoul, C. (1994). Non conditional simulation of stationary isotropic multigaussian random functions. In *Geostatistical Simulations*, M. Armstrong and P.A. Dowd (eds.), Kluwer, Dordrecht, pp. 147-177.

[39] Lebart, L., A. Morineau et J.-P. Fénelon (1979). *Traitement des données statistiques : méthodes et programmes*. Dunod, Paris.

[40] Lecomte, P., H.C. Dubourguier, F. Cazier, Ph. Wavrer et N. Jeannée (1998). Stratégie d'investigation pour les sites industriels pollués par les HAP. Communication lors du colloque Pollutec 1998, Lyon, 4 nov. 1998, non publié.

[41] Lecomte, P. (1999). *Les sites pollués, traitement des sols et des eaux souterraines*. Lavoisier, Paris (2^{nde} éd.).

[42] Liao, H.-T. (1990). *Estimation des réserves récupérables de gisements d'or, comparaison entre krigeage disjonctif et krigeage des indicatrices*, Document du BRGM No. 202 (1991), Bureau de Recherches Géologiques et Minières, Orléans.

[43] Mc Bratney, A.B. and R. Webster (1983). How many observations are needed for regional estimation of soil properties ? *Soil Science*, **135(3)**, pp. 177-183.

[44] Maréchal, A. (1982). Local recovery estimation for co-products by disjunctive kriging. In*Proceedings of the* 17^{th} *APCOM international symposium*, T.B. Johnson and R.J. Barnes (eds.), Society of Mining Engineers of the AIME, New-York, pp. 562-571.

[45] Matheron, G. (1965). *Les variables régionalisées et leur estimation : une application de la théorie des fonctions aléatoires aux sciences de la nature*. Masson, Paris.

[46] Matheron, G. (1970). *La théorie des variables régionalisées et ses applications*, Cahiers du Centre de Morphologie Mathématique, Fasc. 5, Ecole des Mines de Paris.

[47] Matheron, G. (1974). Effet proportionnel et lognormalité ou : le retour du serpent de mer. Rapport N-374, Centre de Géostatistique, Ecole des Mines de Paris.

[48] Matheron, G. (1976). A simple substitute for conditional expectation : the disjunctive kriging. In *Advanced geostatistics in the mining industry*, M. Guarascio, M. David and C. Huijbregts (eds.), Reidel, Dordrecht, pp. 221-236.

[49] Matheron, G. (1978a). *Estimer et choisir*. Cahiers du Centre de Morphologie Mathématique, Fasc. 7, Ecole des Mines de Paris.

[50] Matheron, G. (1978b). *Le krigeage disjonctif et le paramétrage local des réserves*. Cours C-76, Centre de Géostatistique, Ecole des Mines de Paris.

[51] Matheron, G. (1982). La destructuration des hautes teneurs et le krigeage des indicatrices. Rapport N-761, Centre de Géostatistique, Ecole des Mines de Paris.

[52] Matheron, G. (1984). The selectivity of the distributions and the "second principle of geostatistics". In *2nd NATO ASI : "Geostatistics for natural resources characterization"*, G. Verly, M. David, A.G. Journel and A. Maréchal (eds.), Reidel, Dordrecht, pp. 421-433.

[53] Matheron, G. (1988). The internal consistency of models in geostatistics. In *Geostatistics*, M. Armstrong (ed.), Vol. 1, Kluwer, Dordrecht, pp. 21-38.

[54] Montgomery, J.H. (1996). *Groundwater Chemicals*. Desk Reference, Lewis Publishers (2nd ed.).

[55] Neilson, A.H. (1998). *The handbook of environmental chemistry*. Vol. 3 part I : PAH and related compounds - chemistry. Springer, Berlin.

[56] 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.

[57] Pellet, M. et L. Laville-Timsit (1993). Echantillonnage des sols pour la caractérisation d'une pollution : guide méthodologique. Rapport BRGM 37865.

[58] Pellet, M. et P. Lecomte (1993). Diagnostic et évaluation de sites contaminés. Comparaison de techniques d'échantillonnage - Expérimentations sur site et interprétation. Rapport ANTEA A00921.

[59] Petitgas, P. and J. Rivoirard (1991). Global estimation : σ^2/n and the geostatistical estimation variance. In *Workshop on the applicability of spatial statistical methods to acoustic survey data*, ICES/ CIEM, Reykjavik, 5-9 sept. 1991.

[60] Pardo-Igúzquiza, E. and M. Chica-Olmo (1993). The Fourier Integral Method : an efficient spectral method for simulation of random fields. *Mathematical Geology*, **25(2)**, pp. 177-217.

[61] Piel, A., P. Petit et Th. Hosay (1997). Cokeries et centrales : valorisation du charbon et impact environnemental. Document ISSEP.

[62] Pitout, C. (2000). *Conception et utilisation d'un système d'information géographique pour l'étude et le suivi de sites industriels pollués - Analyse spatiale 2D-3D, analyse multiparamètre*. Thèse de Docteur de l'Université des Sciences et Technologies de Lille.

[63] Ricour, J. (1993). Evaluation et diagnostic des sites contaminés. Communication au Colloque Pollutec "Maladie en sous-sol", non publié.

[64] Rivoirard, J. (1985). Convergence des développements en polynômes d'Hermite. *Sciences de la Terre*, Série Informatique Géologique, **24**, pp. 129-159.

[65] Rivoirard, J. (1991). *Introduction au krigeage disjonctif et à la géostatistique non linéaire*. Cours C-139, Centre de Géostatistique, Ecole des Mines de Paris.

[66] Rivoirard, J. (1995). *Concepts et méthodes de la géostatistique*. Cours C-158, Centre de Géostatistique, Ecole des Mines de Paris.

[67] Rivoirard, J. and Ph. Guiblin (1996). Global estimation variance in presence of conditioning parameters. In *Geostatistics Wollongong '96*, E.Y. Baafi and N.A. Schofield (eds.), Kluwer, Dordrecht, Vol. 1, pp. 246-257.

[68] Rivoirard, J. and N. Bez (1997). A 2D geostatistical analysis of northern blue whiting acoustic data west of the british isles. Conseil International pour l'Exploitation de la Mer, CM 1997/Y :13.

[69] Rivoirard, J. (1998). Les variogrammes pondérés. Rapport N-46/98/G, Centre de Géostatistique, Ecole des Mines de Paris.

[70] Rivoirard, J. (2000). Weighted Variograms. In *Geostats 2000*, W. Kleingeld and D. Krige (eds.). Sous presse.

[71] Saporta, G. (1990). *Probabilités, analyse des données et statistique*. Technip, Paris.

[72] Silverman, B.W. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall, London.

[73] Srivastava, R. and H. Parker (1989). Robust measures of spatial continuity. In *Geostatistics*, M. Armstrong (ed.), Vol. 1, Kluwer, Dordrecht, pp. 295-308.

[74] Steyer, E. (2000a). Ancienne cokerie du site X. Compte-rendu de la campagne d'échantillonnage mai, juin et juillet 2000. Rapport CNRSSP 2000/09, Douai.

[75] Steyer, E. (2000b). Caractérisation *in situ* de sites issus d'activité liées au charbon. Amélioration et applicabilité en vraie grandeur de tests rapides de terrain permettant la détection des HAP et des BTEX. $2^{\text{ème}}$ rapport d'avancement de thèse. Rapport CNRSSP 2000/21, Douai.

[76] Wackernagel, H. (1995). *Multivariate geostatistics - an introduction with applications*. Springer, Berlin.

[77] Wavrer, Ph. (1996). *Apport à la théorie de l'échantillonnage des solides hétérogènes. Application à des grandeurs mesurées sur matières premières, secondaires ou ultimes*. Documents du BRGM No. 265, Bureau de Recherches Géologiques et Minières, Orléans.

[78] Wavrer, Ph. (1997a). Ancienne cokerie du site de X. Compte-rendu de la campagne d'échantillonnage. Rapport CNRSSP 1997/08, Douai.

[79] Wavrer, Ph. (1997b). Méthodologies et stratégies d'échantillonnage de sols. Etude bibliographique. Rapport CNRSSP 1997/11, Douai.

[80] Wavrer, Ph. et N. Jeannée (1998). Ancienne cokerie du site de Y. Compte-rendu de la campagne d'échantillonnage. Rapport CNRSSP 1998/03, Douai.

[81] Wavrer, Ph. (1998). Echantillonnage de sites contaminés par des polluants organiques. Comparaison de différentes stratégies et outils. Rapport CNRSSP 1998/27, Douai.

Sixième partie

Annexes

Annexe A

Audit de sites

Sommaire

Cette annexe détaille le principe du diagnostic de sites pollués, ainsi que la méthodologie généralement mise en œuvre : enquête documentaire suivie éventuellement d'une investigation de terrain. L'importance de la représentativité d'un échantillonnage et les différentes sources d'hétérogénéité auxquelles celui-ci est exposé sont discutées.

Pour appréhender les difficultés liées aux spécificités des pollutions de sols, la France préconise une approche par *audit de site*, dont le principe a été présenté au paragraphe 1.1.

Bien que plus adaptée à la nature du problème qu'une approche normative, cette analyse des risques est par construction lourde et souvent longue. Pour un site donné, selon les points de vue et intérêts des uns et des autres, l'appréciation des mesures de dépollution nécessaires et des coûts associés peut faire l'objet d'une fourchette très large. Nous allons à présent rappeler les deux phases essentielles d'un audit de site : l'enquête documentaire, et l'investigation de terrain.

A.1 Enquête documentaire

Nous avons mentionné comme point de départ d'une étude de risque le "rassemblement des informations disponibles sur le site". Cela implique d'une part l'analyse de l'ensemble des activités passées et présentes sur le site - analyse historique -, et la connaissance générale du site et de son milieu naturel d'autre part. Cette phase documentaire est essentielle et doit exploiter au maximum les multiples sources d'information disponibles.

La phase historique doit récapituler les différentes utilisations du site au cours du temps, les produits potentiellement polluants qui y ont transité ainsi que la localisation des différentes installations. Il est dans certains cas peu aisé de retrouver de telles informations, notamment lorsqu'il s'agit d'industries anciennes telles que les activités charbonnières qui concernent plus particulièrement

ce travail. La connaissance des caractéristiques géologiques et hydrologiques du site permet d'estimer sa vulnérabilité par rapport aux risques de contamination provenant des activités qui s'y sont exercées.

La visite du site est finalement une étape indispensable, permettant de confronter à la réalité les informations recueillies lors de l'enquête historique, de se rendre compte des conditions particulières éventuellement insoupçonnées et de préparer si nécessaire une phase d'investigation par le repérage des zones accessibles, des zones de risque potentiel et des normes de sécurité à respecter [Lecomte (1999)].

A.2 Campagnes d'investigation

Lors de l'étude d'un site, une voire plusieurs campagnes d'investigation sont mises en place en fonction des informations collectées. Il existe pour cela un grand nombre de stratégies d'échantillonnage et de techniques de prélèvement.

Ces campagnes visent notamment à confirmer la présence de pollutions suspectées suite à l'enquête historique. Il serait néanmoins dangereux de s'en tenir là, l'étude historique ne permettant fréquemment pas, dans le cas de sites industriels anciens, une connaissance exhaustive des taches de pollution potentielles. Dans le cas d'une évaluation détaillée des risques, il sera intéressant pour l'analyse des risques et le choix d'une méthode de dépollution de pouvoir estimer les concentrations en place, voire les volumes de terrain contaminés.

On distingue deux groupes de méthodes d'investigation [Lecomte (1999)]. Les premières consistent à réaliser certaines mesures directement sur le site : géophysique, hydrogéologie, paramètres physico-chimiques, etc. Par ailleurs, il existe de nombreuses méthodes d'investigation indirectes et hors site : géotechnique, chimie, etc. Ces méthodes associent échantillonnages sur le site et mesures en laboratoire spécialisé. La combinaison de méthodes d'investigation différentes est fréquente, une campagne d'échantillonnage devant équilibrer de façon pertinente information nécessaire et moyens à mettre en place pour l'obtenir. Cet objectif difficile à atteindre dépend notamment de la spécificité du site rencontré, et seules quelques règles générales peuvent être appliquées.

La phase d'interprétation qui suit est essentielle mais cependant très délicate. Elle doit à la fois donner un sens aux données et établir le type, le niveau et la répartition de la pollution suspectée, afin de permettre une évaluation des dangers et de pouvoir juger de la nécessité d'un traitement.

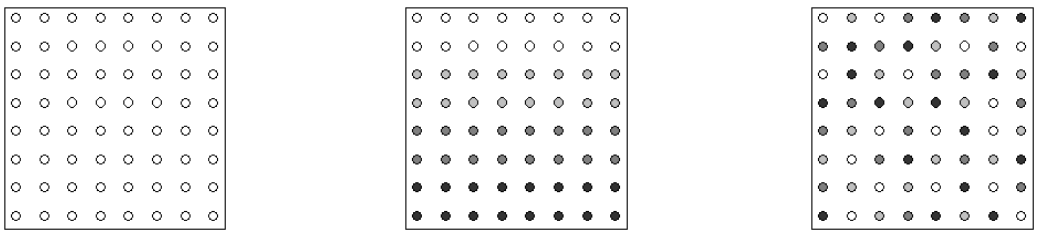
A.2.1 Hétérogénéités et représentativité d'un échantillonnage

Avant de présenter les principaux types de plans d'échantillonnage et techniques de prélèvement d'échantillons, il est nécessaire d'aborder la question de la *représentativité d'un échantillonnage*. Nous entendons donc ici par *échantillonnage* l'opération élémentaire qui consiste à prélever une certaine fraction du lot de matière à étudier [Wavrer (1996)]. Une campagne d'investigation vise à prélever des échantillons de taille souvent réduite qui seront analysés afin d'être interprétés et de servir de support pour toute prise de décision. Faute d'une investigation mal adaptée, mal conduite ou trop limitée, il arrive encore que certains sites présentant un degré faible de pollution

soient déclarés fortement contaminés ; plus grave, nombre de sites sont à l'inverse considérés comme "propres" alors qu'ils présentent en réalité un niveau de pollution important. Il découle de cela une des questions fondamentales lors de toute étude de site : "Que vaut un résultat ?" [Halleux & Nix (1991)].

Pour être valable, un échantillonnage doit présenter certaines propriétés et qualités [Gy (1988)] :
- Il doit tout d'abord permettre d'obtenir des échantillons *représentatifs* du sol étudié au regard de la grandeur suivie. Cela signifie d'une part l'absence d'un biais trop important - justesse - et d'autre part la reproductibilité - ou fidélité.
- Par ailleurs, un échantillonnage doit être réalisé de façon *correcte*, c'est-à-dire que tous les éléments constitutifs du sol doivent avoir la même probabilité d'être retenus dans l'échantillon. Un échantillonnage non correct ne sera jamais représentatif, et c'est donc dès sa conception et surtout sa mise en œuvre qu'il est nécessaire d'assurer cette correction.

Il est important de prendre conscience que malgré tout le soin apporté à la préparation d'une campagne d'échantillonnage et à sa réalisation, des erreurs d'échantillonnage sont irrémédiablement commises. Celles-ci, liées à l'hétérogénéité de la matière échantillonnée, peuvent être subdivisées en plusieurs composantes [voir par exemple Wavrer (1996)] :
- **Erreur fondamentale d'échantillonnage**
  Elle découle de l'*hétérogénéité de constitution* de la matière, qui résulte des particularités physico-chimiques des particules individuelles de la matière, dans l'état où celle-ci se trouve. L'hétérogénéité de constitution est indépendante de la distribution spatiale des particules. Le mélange et l'homogénéisation n'ont aucune influence sur elle, de sorte que l'erreur associée ne peut être annulée ni même réduite sans modification de l'état physique de la matière ; elle correspond donc à un minimum incompressible au-delà duquel il est illusoire de vouloir énoncer une quelconque précision sur la mesure effectuée. Afin d'évaluer les performances limites d'un échantillonnage donné, il est possible de calculer cette erreur.
- **Erreur de ségrégation**
  Liée généralement à l'*hétérogénéité de distribution* spatiale des différents constituants de la matière étudiée, cette erreur peut théoriquement être réduite voire annulée par homogénéisation du lot à analyser. Cette homogénéisation est néanmoins très difficile à assurer dans le cas de sols hétérogènes tels que ceux rencontrés sur d'anciennes friches industrielles[1]. Les deux types d'hétérogénéité de la matière sont illustrés à la figure A.1.



Homogénéité de constitution
Homogénéité de distribution

Hétérogénéité de constitution
Hétérogénéité de distribution

Hétérogénéité de constitution
Homogénéité de distribution

FIG. A.1 – Hétérogénéités de constitution et de distribution [d'après Wavrer (1996)].

[1] Il est possible d'améliorer cette homogénéisation par des techniques telles que le cryo-broyage ; les coûts générés par ces techniques sont néanmoins souvent prohibitifs.

– **Erreurs de préparation et de prélèvement**
Elles sont dues aux techniques et/ou procédures de prélèvement utilisées, ainsi qu'aux différentes manipulations - conditionnement, transport, analyse - pouvant affecter l'intégrité de l'échantillon. Par un soin particulier apporté lors de la constitution de l'échantillon et de sa manipulation, il est possible de les diminuer sensiblement voire de les éliminer.

A.2.2 Plans d'échantillonnage

La conception d'un plan d'échantillonnage doit permettre une reconnaissance efficace de l'ensemble des zones potentiellement intéressantes. Un certain réalisme est ici essentiel, la reconnaissance exhaustive d'un site relevant de l'utopie. Il s'agit de trouver le meilleur compromis entre le coût de la campagne envisagée, les objectifs recherchés, la taille du site investigué, le type de pollution et le degré de précision attendu.

Le choix d'un schéma d'échantillonnage adapté aux objectifs fixés est prépondérant. On en distingue essentiellement trois types :

– **Echantillonnage aléatoire**
Une telle approche est fréquemment mise en œuvre lorsque les informations sur l'état du site, en termes de pollution, sont inexistantes ou insuffisantes pour une utilisation rationnelle. Les prélèvements sont répartis aléatoirement sur la zone à investiguer. Plus coûteux qu'une autre méthode, en ce sens qu'un plus grand nombre d'échantillons est souvent nécessaire pour obtenir la même précision d'information que par une autre méthode, l'échantillonnage aléatoire est néanmoins adapté à la caractérisation d'une zone homogène par rapport au paramètre mesuré [Patata & Mastrolilli de Angelis (1997), cité dans Wavrer (1997b)]. Ces conditions étant extrêmement rares dans le domaine des sols pollués, ce type d'échantillonnage n'est pas à conseiller.
L'échantillonnage aléatoire stratifié en est une variante ; il consiste à découper la zone à investiguer en entités plus homogènes que la zone initiale puis à procéder à un échantillonnage aléatoire au sein de chacune des strates géographiques ainsi constituées.
– **Echantillonnage systématique**
Cet échantillonnage consiste à effectuer les prélèvements aux nœuds d'une grille régulière. La maille de cette grille est le plus souvent carrée[2] et son origine est déterminée de façon à optimiser la couverture de la zone investiguée. Souvent utilisée pour déterminer l'extension d'une pollution ou la distribution d'un polluant donné, cette approche fournit le meilleur rapport entre le nombre d'échantillons à prélever - et donc le coût - et la précision attendue lors de l'interprétation des résultats.
– **Echantillonnage ciblé**
La détermination de la position des prélèvements est ici faite de façon tout à fait subjective, sur la base de l'étude historique et des visites préliminaires. Les échantillons peuvent par exemple être groupés à proximité d'anciennes mares à goudrons, ou en considérant des critères d'odeur, de texture, etc. Cet échantillonnage, s'il permet de déterminer les produits présents, qu'il sera intéressant de suivre dans une éventuelle campagne extérieure, ne prémunit

[2]Une maille triangulaire voire hexagonale est également envisageable. En théorie meilleures que la maille carrée, en ce sens que pour un même nombre d'échantillons elles permettent une estimation plus précise, le repérage plus délicat des points explique leur utilisation plus rare.

absolument pas contre le risque de rater les éventuelles taches de pollution que l'enquête historique n'a pas mis en évidence.

La détermination du nombre d'échantillons à prélever est un problème difficile. Celui-ci dépend fortement de la complexité du milieu, de la précision attendue de la campagne, de l'importance des connaissances déjà acquises. Certains calculs statistiques sont possibles afin de déterminer le nombre d'échantillons minimum pour atteindre un niveau de précision donné, mais nécessitent des hypothèses peu réalistes, par exemple que les valeurs échantillonnées suivent une loi normale [Cline (1944), cité dans Mc Bratney & Webster (1983)]. Notons que la géostatistique peut également apporter une réponse à ce problème. Le nombre d'échantillons à prélever découle souvent directement du financement alloué à la campagne.

A.2.3 Techniques de prélèvement

Avant d'effectuer les prélèvements, il est évidemment nécessaire de sélectionner la technique ou le matériel susceptible de convenir le mieux au contexte de l'étude - type de sol, profondeur à atteindre, etc. Le coût étant variable selon la technique, ce choix a également une incidence sur le nombre d'échantillons que l'on peut financièrement prélever. On distingue cinq grands types de techniques généralement utilisées pour le prélèvement d'échantillons de sol [Wavrer (1997b)] :

- **Pelles, truelles, cuillères**
  Simples d'utilisation, peu coûteux, ils sont adaptés aux sols granulaires et pour de faibles profondeurs - jusqu'à 1 m, voire 1.50 m -, mais à déconseiller pour le prélèvement de sols contaminés par des polluants volatils.
- **Tarières légères**
  A mains ou à moteur, à gouge ou à spirale (ou vis), ces tarières sont utilisables jusqu'à des profondeurs de 2 m en moyenne - 6 à 8 m pour les tarières à moteur. D'utilisation aisée et rapide, elles sont à déconseiller si l'on désire une description des changements de faciès lithologiques ou des profondeurs, à cause du remaniement important.
- **Tubes ou sondes de prélèvement**
  L'intérêt est ici d'éviter le remaniement de l'échantillon grâce au tube ou à la sonde. Ces outils sont donc bien adaptés pour obtenir des informations géologiques. Leur utilisation peut être délicate dans le cas de sols meubles, à granulométrie fine ou encore en présence de remblais pouvant contenir des blocs de grosse taille. Sensiblement identiques, les forages à l'aide de foreuses à tarière, à foret rotatif ou à percussion (carottier) permettant d'atteindre des profondeurs plus importantes (voir photos à la figure A.2).
- **Creusement de fosses ou de tranchées à l'aide d'un engin d'excavation**
  Cette technique (voir photos à la figure A.3) permet l'observation d'une coupe de sol en place avant d'effectuer le prélèvement. Outre l'utilisation d'un matériel plus lourd - souvent une pelle mécanique -, cette méthode présente certains risques - éboulement des parois - lorsque les fosses excavées sont profondes. Par ailleurs, le risque de dégagements toxiques en fond de fosse, associé à un certain confinement, oblige à prendre certaines mesures de sécurité. Ces dégagements permettent néanmoins l'analyse des gaz intersticiels (voir figure A.4).

(a) (b)

FIG. A.2 – Carottier à percussion (a) fonctionnant sans adjonction d'eau ni d'air et utilisant une sonde à lumières (b).

(a) (b)

FIG. A.3 – Pelle mécanique (a) permettant le creusement des fosses (b).

FIG. A.4 – Capot en tôle muni d'un tuyau en caoutchouc, posé sur les fosses immédiatement après leur creusement et relié à un détecteur à photo-ionisation de modèle MultiwarmPID qui évalue la teneur en polluants organiques volatils totaux.

Annexe B

Compléments sur les cokeries et l'analyse des HAP

Sommaire

Cette annexe détaille le fonctionnement d'une cokerie avant de présenter le protocole de préparation et d'analyse des échantillons prélevés lors des campagnes d'investigation.

B.1 Cokeries

B.1.1 Cokéfaction

La figure B.1 décrit les flux de matières au sein d'une cokerie. Les charbons utilisés lors d'une cokéfaction se présentent en grains de taille inférieure au centimètre et ont une humidité comprise entre 8 et 10 %. Certaines caractéristiques sont requises pour l'obtention d'un coke de bonne qualité, et il est rare qu'un seul type de charbon les réunisse toutes. On recourt donc, pour la cokéfaction, au mélange de 2 ou 3 types de charbons, appelé *pâte à coke* [Piel et al. (1997)].

La cokéfaction a lieu dans des fours à coke, constitués de fours accolés, encore appelés batteries ou groupes de cellules. Chaque groupe réunit 25 à 80 cellules de carbonisation - Longueur : 13 à 16 m, largeur : 35 à 50 mm, hauteur : 5 à 7 m -, séparées par un piédroit, paroi creuse où circulent les gaz dont la combustion assure le chauffage de la batterie. Le chargement du four se fait par pilonnage - latéralement - ou par gravité - par le haut du four à travers des bouches d'enfournement. La pâte à coke est enfournée et chauffée par les piédroits, à l'abri de l'air, jusqu'à 1000°C environ, pendant 14 à 18 heures. Le chauffage provoque d'abord la vaporisation de l'eau contenue dans la pâte à coke. Ensuite, le charbon devient plastique entre 350 et 420°C, ce qui provoque un dégagement de produits goudronneux qui se propage depuis les parois jusqu'au milieu de la charge. Une resolidification intervient vers 450°C, se traduisant par la formation de semi-

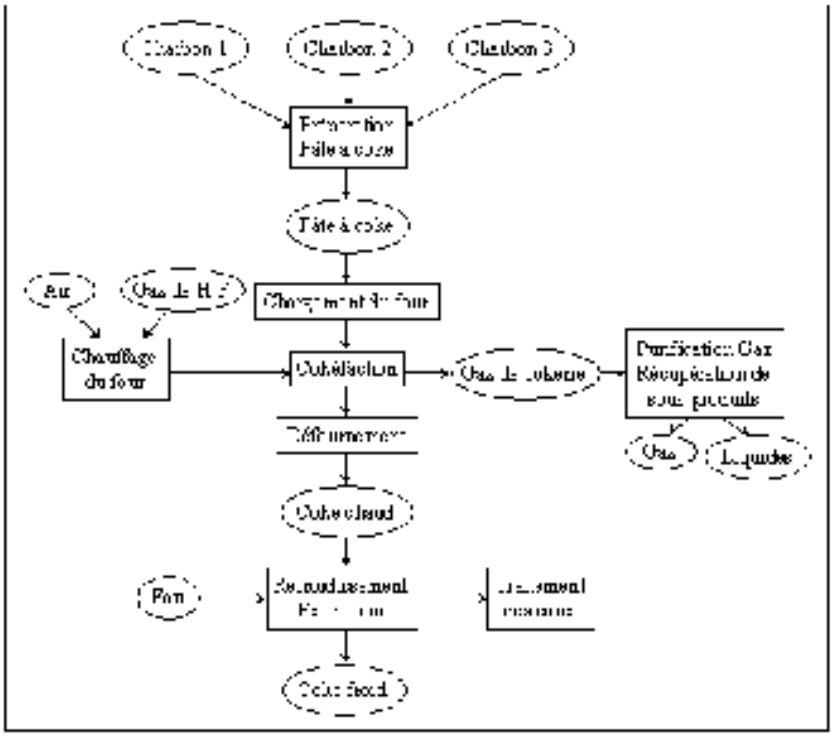


FIG. B.1 – Flux de matières au sein d'une cokerie [Piel et al. (1997)].

coke, lequel produira finalement lors de la poursuite du chauffage le coke de haute température. Le défournement se fait par une défourneuse, système qui pousse le coke à travers les portes latérales des fours, vers le wagon de refroidissement qui amène le coke chaud à la tour de refroidissement.

B.1.2 Impact environnemental

Un four à coke ne constitue pas un réacteur continu et étanche. Il y a par conséquent émission d'hydrocarbures :

- par fuites (portes des fours, tampons d'enfournement, ...),
- à la suite des opérations discontinues réclamées par le procédé : chargement et défournement des fours, extinction humide du coke, maintenance, etc. Ces émissions de gaz et de poussières sont essentiellement diffuses ; nature physico-chimique et température de ces émissions varient.

On peut décrire, étape par étape, les pollutions engendrées par une cokerie [Piel et al. (1997)] :

- manutention et stockage du charbon : dégagements de poussière, dépendant de la granulométrie et du taux d'humidité du charbon.
- chargement des fours : émission irrégulière de gaz par les différentes ouvertures libres, principalement de l'azote, mêlé à du CO_2 et du CO, ainsi qu'à de faibles quantités d'oxygène . Ces gaz peuvent entraîner de la poussière de charbon. Lors du chargement, le charbon commence sa distillation, ce qui engendre un dégagement de vapeur d'eau, de méthane, de CO ainsi que d'hydrocarbures non saturés et produits benzéniques.

- cokéfaction : émissions diffuses de gaz au niveau des portes, du toit des fours, du système de récupération des gaz de distillation et du système de traitement de ceux-ci. La composition de ces gaz varie d'une installation à l'autre, mais les produits majeurs restent identiques.
- défournement : il peut engendrer des émissions importantes lorsque la cokéfaction n'a pas été poussée assez loin. Si elle est complète, des émissions de particules peuvent être générées par le flux ascendant d'air chaud au-dessus du coke.
- extinction : le transport du four vers la tour d'extinction entraîne l'émission de vapeurs d'eau et de poussières.
- manutention : elle entraîne exclusivement l'émission de poussières.

B.2 Préparation et analyse des HAP

L'efficacité de la caractérisation d'une pollution dépend de la qualité de la campagne d'investigation et de l'analyse chimique des polluants qui lui succède. Nous avons vu au chapitre A les différentes stratégies d'échantillonnage possibles ainsi que les techniques de prélèvement. Une fois l'échantillon prélevé, il est important que la chaîne de préparation et d'analyse de cet échantillon soit la plus précise possible, pour un coût acceptable, et ne soit pas source de variabilités importantes pour les concentrations. En outre, il est recommandé que le délai de conservation soit court, afin d'éviter une évolution de la pollution au sein du prélèvement[1]. On distingue trois étapes :

- **Préparation de l'échantillon**
  Les échantillons prélevés sont tout d'abord homogénéisés. Cette phase, essentielle si l'on ne veut perdre une partie de la représentativité de l'échantillon, reste cependant souvent délicate à réaliser dans le cas de matériaux hétérogènes [Belkessam (1999)]. Ensuite, conformément à la norme AFNOR NF ISO 13877, les échantillons sont séchés à l'air libre puis broyés au mortier et finalement tamisés. Seule la fraction granulométrique inférieure à 2 mm est analysée. Cela peut prêter à caution ; par exemple, Belkessam (1999) montre que dans le cas d'un sol limoneux faiblement pollué le tamisage à 2 mm entraîne une sous estimation d'environ 50 % des concentrations par rapport au traitement de la totalité de l'échantillon.
- **Extraction**
  Cette phase consiste à extraire de la matrice solide les substances désirées en utilisant des solvants adaptés. Les autres composés sont éliminés pour éviter toute risque d'interférence. Pour les HAP analysés sur les deux sites discutés dans ce travail, on a eu recours à l'extraction accélérée par solvant (ASE), qui est une extraction à chaud et sous pression. Un mélange de dichlorométhane / acétone est utilisé comme solvant.
- **Dosage**
  Le dosage est effectué dans notre cas par chromatographie liquide à haute performance (HPLC), et la détection à la fin de la colonne se fait par U.V. à barette de diodes. Cette méthode donne de meilleurs résultats pour les HAP lourds que la chromatographie en phase gazeuse (GC) [Cazier et al. (1997)], mais des résultats plus variables pour les HAP légers. Par ailleurs, elle présente des risques de confusion entre les couples Phe-Ant, Baa-Cry, Bbf-Bkf, Dba-coronène [Neilson (1998)].

[1] Dans notre cas, ce délai a été de plusieurs mois, ce qui est à proscrire lors d'un audit de site. Néanmoins, l'objectif se situant au niveau de la recherche, cela ne s'est pas avéré gênant, la procédure ayant été appliquée de façon similaire à tous les prélèvements.

Annexe C

Rappels de géostatistique

Sommaire

Destinée au lecteur non familier à la géostatistique, cette annexe en présente les notions et concepts nécessaires à la compréhension du mémoire de thèse. Il n'a aucune prétention d'exhaustivité et est volontairement synthétique et intuitif. L'essentiel de ce qui est présenté provient de Matheron (1965, 1970), Chilès & Delfiner (1999), Rivoirard (1995) et Chauvet (1999).

C.1 Notions statistiques

Ces rappels sont pour l'essentiel issus de Rivoirard (1995). On s'intéresse à la concentration en un polluant donné dans un sol. Pour cela, on procède à l'échantillonnage du sol en dix points alignés. L'analyse de la concentration de chaque échantillon nous donne les valeurs 1.7, 1.2, 1.5, 3., 2.5, 1.8, 3.5, 7., 10. et 5.9 ppm, que nous noterons $x_1, \ldots, x_{10}$. En classant ces valeurs par ordre croissant et en les regroupant par classes[1], on obtient l'histogramme des concentrations, qui illustre graphiquement les fréquences d'observation des différentes concentrations (voir figure C.1).

La *moyenne* m des concentrations vaut

$$m = \frac{1}{10} \sum_{i=1}^{10} x_i = 3.81$$

Cependant, la concentration des 10 échantillons varie autour de cette concentration moyenne de

[1] La détermination du nombre de classes de l'histogramme est une étape délicate. Un nombre trop faible de classes conduit à une perte d'information, tandis que des classes trop nombreuses donnent des graphiques souvent incohérents : classes vides, peu représentées [Saporta (1990)]. Aucune règle absolue n'existant dans ce domaine, on procède empiriquement en testant différents nombre de classes et en retenant le choix le plus probant.

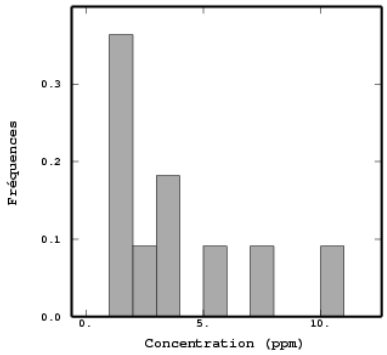


FIG. C.1 – Histogramme des concentrations en polluant.

3.81. Cette variabilité se mesure par la *variance* des concentrations

$$\sigma^2 = \frac{1}{10}\sum_{i=1}^{10}[x_i - m]^2 = 7.60$$

Définie comme une moyenne de carrés, cette variance est toujours positive. Par ailleurs, faisant intervenir des carrés de concentration, c'est une grandeur qui s'exprime en ppm^2. Par commodité, on utilise souvent l'*écart-type* σ, racine carrée de la variance, afin d'avoir une information sur la variabilité de la concentration qui soit homogène à celle-ci ; ici, l'écart-type est de 2.75. Dans le cas de variables dissymétriques, on a fréquemment recours au *coefficient de variation* $\frac{\sigma}{m}$; celui-ci sera d'autant plus élevé que l'histogramme de la variable est dispersé, i.e. la variable fort contrastée. En outre, il a l'avantage d'être sans unité et relatif à la moyenne.

Une version probabiliste de cela s'obtient en considérant la concentration comme une *variable aléatoire* X, dont nous observons 10 réalisations, les 10 valeurs analysées. La moyenne, ou espérance, valeur probable de X, est notée $\mathrm{E}[X]$. Cette espérance est estimée par la moyenne des valeurs prises par nos observations, c'est-à-dire $m = 3.81$. On peut également estimer l'espérance de la variable aléatoire X^2, $\mathrm{E}[X^2] = 22.11$. D'autre part, la variance σ_X^2 de X, notée $\mathrm{Var}[X]$, vaut $\mathrm{E}[X - \mathrm{E}[X]]^2 = E[X^2] - E^2[X]$ et est estimée par la variance des données 7.60. En ajoutant à X une constante a, on obtient $\mathrm{E}[a + X] = a + \mathrm{E}[X]$, mais cela ne modifie pas la variance.

Soient deux variables aléatoires X et Y. L'espérance de leur somme est égale à la somme de leurs espérances (additivité)

$$\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$$

On aura fréquemment recours à la covariance

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]$$

qui vaut également $\mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]$. En valeur absolue, la covariance est toujours inférieure au produit des écart-types $\sigma_X\sigma_Y$. Le degré de corrélation entre deux variables est souvent mesuré à l'aide du *coefficient de corrélation*

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_X\sigma_Y}$$

sans dimension et toujours compris entre -1 et 1. Plus ce coefficient est élevé et plus les variables sont dites corrélées ; inversément, lorsqu'il est négatif, on parle d'anti-corrélation.

Considérons deux constantes a et b, et la combinaison linéaire $aX+bY$ de X et Y. Des formules précédentes il découle simplement

$$\begin{aligned} \mathrm{E}[aX+bY] &= a\mathrm{E}[X]+b\mathrm{E}[Y] \\ \mathrm{E}[(aX+bY)^2] &= a^2\mathrm{E}[X^2]+b^2\mathrm{E}[Y^2]+2ab\mathrm{E}[XY] \\ \mathrm{Var}[aX+bY] &= a^2\mathrm{Var}[X]+b^2\mathrm{Var}[Y]+2ab\mathrm{Cov}(X,Y) \end{aligned}$$

Dans le cas de deux variables aléatoires X et Y d'espérances nulles, nous avons en particulier

$$\begin{aligned} \mathrm{Var}[X] &= E[X^2] \\ \mathrm{Cov}(X,Y) &= \mathrm{E}[XY] \end{aligned}$$

Ces "moments d'ordre 1" $\mathrm{E}[X]$ et $\mathrm{E}[Y]$ et "moments d'ordre 2" $\mathrm{E}[X^2]$, $\mathrm{E}[Y^2]$, $\mathrm{E}[XY]$ de X et Y nous suffiront pour la mise en œuvre de la géostatistique dite linéaire, cette dernière ne faisant intervenir que des combinaisons linéaires de ces moments.

Par ailleurs, d'autres statistiques sont intéressantes lors de l'étude d'une variable régionalisée. C'est le cas des *quantiles*, obtenus de la façon suivante : après rangement par ordre croissant des valeurs, le quantile à p %, noté q_p, est la valeur de la donnée telle que p % des valeurs lui sont inférieures. Par exemple, le quantile à 0 % est la valeur minimale des données, celui à 100 % la valeur maximale. On distingue en particulier la *médiane*, qui est le quantile à 50 %. L'intérêt de cette statistique est sa faible sensibilité, sa "robustesse" par rapport aux variations des valeurs fortes. Dans le cas d'une distribution symétrique, médiane et moyenne sont égales ; si la distribution est dyssimétrique avec une queue de distribution vers les valeurs fortes - nous ne traiterons que ce cas de dissymétrie -, alors la moyenne, sensible aux valeurs fortes, devient supérieure à la médiane. Les quantiles à 25, 50 et 75 % sont également appelés premier, deuxième et troisième quartile. L'*intervalle interquartile* $|q_{0.75}-q_{0.25}|$ est un indicateur de la dispersion de la distribution [Saporta (1990)].

Pour l'instant, aucune considération de l'implantation spatiale des concentrations n'a été faite. La carte d'implantation de la figure C.2 reprend la localisation des prélèvements effectués, la taille des croix étant proportionnelle à la concentration analysée. Cette représentation permet la visualisation des zones de forte et/ou faible concentration. Avec les mêmes concentrations analysées, et donc avec une moyenne et une variance identiques, d'autres implantations des concentrations auraient pu survenir, comme celles illustrées à la figure C.3. Ces différentes implantations ne correspondent cependant pas à la même structure spatiale, et c'est justement dans l'analyse de ces structures spatiales et leur utilisation que réside l'intérêt premier de la géostatistique.

C.2 Variables régionalisées

De façon très synthétique, la géostatistique vise à décrire des phénomènes naturels corrélés dans l'espace et éventuellement le temps et à quantifier l'incertitude liée à leur estimation. Les méthodes

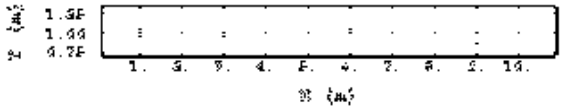


FIG. C.2 – Carte d'implantation des concentrations en polluant.

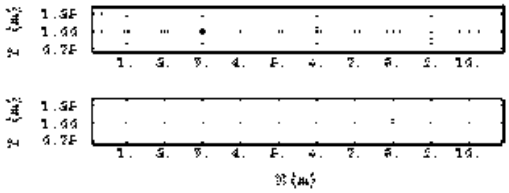


FIG. C.3 – Cartes d'implantations alternatives.

mises en œuvre ont commencé à apparaître dans les années 50, tout d'abord dans le domaine minier. Depuis, les domaines d'application se sont fortement diversifiés : pétrole, hydrogéologie, halieutique, topographie, météorologie, pédologie, finance ainsi que l'environnement, qui nous intéresse plus spécifiquement dans ce mémoire.

Nous nous intéressons à un phénomène dit *régionalisé* qui se déploie dans l'espace, en pratique sur un champ borné S. Ce phénomène régionalisé est par exemple la concentration d'un polluant sur une parcelle d'un site, ou bien encore la hauteur piézométrique d'une nappe, la topographie d'une région géographique, etc. Pour l'étudier, "nous supposons que ce phénomène se laisse décrire de manière satisfaisante par la donnée d'une (ou éventuellement plusieurs) fonction z définie sur S, que nous appellerons, d'un terme général, Variable Régionalisée" [Matheron (1978a)].

Généralement, cette variable n'est connue qu'en quelques points du champ auquel le praticien s'intéresse, par exemple en chaque point d'une grille plus ou moins fine recouvrant le champ. Tout le travail du géostatisticien va dès lors consister à prévoir, à partir de ces quelques données et de l'information qualitative dont il peut disposer sur la variable - par exemple le comportement assez lisse d'une variable telle que la piézométrie -, le comportement de la variable sur l'ensemble du champ.

Il faut garder à l'esprit que l'efficacité de la démarche est entièrement conditionnée par le fait que les données que nous possédons soient représentatives du phénomène étudié, ce qui peut ne pas être le cas si par exemple l'échantillonnage est mal adapté : comment par exemple détecter des anomalies décamétriques à partir de mesures espacées d'un kilomètre [Rivoirard (1995)] ?

C.3 Fonctions aléatoires

Considérons la concentration d'un polluant en un point du sol. Cette concentration peut être considérée comme une variable aléatoire. A présent nous nous intéressons non pas à la concentration en un point du sol, mais en N localisations notées $x_1, \ldots, x_N$. En notant ces concentrations $z(x_1)$, $z(x_2)$, ..., $z(x_N)$, nous allons considérer ces valeurs comme des réalisations d'un ensemble de variables aléatoires $Z(x_1)$, $Z(x_2)$, ..., $Z(x_N)$. De façon plus synthétique, on considérera l'ensemble des valeurs $z(x)$ comme une réalisation d'une variable aléatoire $Z(x)$ indexée par la position dans l'espace [Rivoirard (1995)]. C'est cette variable aléatoire indexée par la position que nous appellerons *fonction aléatoire* (FA). Nous réserverons l'usage de Z pour la fonction aléatoire et celui de z pour la variable régionalisée qui en est une réalisation.

Une fonction aléatoire se caractérise par sa *loi spatiale*, qui est la fonction

$$f(x_1, \dots, x_N) = \mathrm{P}[Z(x_1) \leq z_1, \dots, Z(x_N) \leq z_N]$$

connue quelque soit le nombre N de points et l'implantation de ces N points. Par souci de légèreté, la notation abrégée z_i sera souvent utilisée à la place de $z(x_i)$.

C.4 Propriétés fondamentales

C.4.1 Stationnarité

Une fonction aléatoire est dite *stationnaire* si sa loi spatiale est invariante par translation des $(x_i)_{i=1,\dots,N}$ d'un vecteur h, c'est-à-dire si

$$\mathrm{P}[Z(x_1) \leq z_1, \dots, Z(x_N) \leq z_N] = \mathrm{P}[Z(x_1 + h) \leq z_1, \dots, Z(x_N + h) \leq z_N]$$

Intuitivement, cela signifie que le phénomène tout entier est homogène dans l'espace. On se rend bien compte que la stationnarité stricte est une hypothèse extrêmement forte. Pour cela, on se limite fréquemment à une *stationnarité d'ordre deux*, c'est-à-dire l'invariance par translation des moments d'ordre 1 et 2

$$\begin{aligned} \mathrm{E}[Z(x)] &= m \\ \mathrm{Cov}(Z(x), Z(x+h)) &= \mathrm{E}[(Z(x) - m)(Z(x+h) - m)] = \mathrm{E}[Z(x)\, Z(x+h)] - m^2 = C(h) \end{aligned}$$

pour x et $x+h$ appartenant au champ d'étude. Une fonction aléatoire est donc stationnaire d'ordre deux (FASt2) si son espérance est constante et si la covariance pour deux points séparés d'une distance h est une fonction C dépendant uniquement de cette distance. En particulier, pour $h = 0$, nous avons que $\mathrm{Cov}(Z(x), Z(x)) = \mathrm{Var}[Z(x)] = C(0)$, et donc que la variance est constante sur le champ.

Par stationnarité nous sous-entendrons stationnarité d'ordre deux, sauf lorsque le contraire est précisé. Nous avons vu que la covariance est bornée supérieurement par le produit des écart-types. Cela entraîne que, pour tout h, $C(h) \leq \mathrm{Var}[Z(x)]$, qui est $\mathrm{Cov}(Z(x), Z(x)) = C(0)$. Donc, la covariance est toujours inférieure à sa valeur $C(0)$ à l'origine, qui est la variance.

Il est important de se rendre compte que la stationnarité est une propriété de la fonction aléatoire. Cependant, si nous nous plaçons au niveau de la variable régionalisée, observée en quelques points d'un champ, on perçoit bien que la stationnarité supposée du modèle associé va dépendre de la taille du champ, de la fenêtre d'observation que nous avons de ce phénomène.

Il est encore possible d'affaiblir les hypothèses, en supposant que ce n'est plus $Z(x)$ qui est stationnaire d'ordre deux, mais l'accroissement $Z(x + h) - Z(x)$ pour tout vecteur h. Ainsi, une fonction aléatoire $Z(x)$ est dite *intrinsèque* (FAI) si les accroissements $Z(x + h) - Z(x)$ vérifient pour tout h les propriétés suivantes :

- $\mathrm{E}[Z(x + h) - Z(x)] = 0$
  Cette propriété signifie qu'il n'y a pas de *dérive* du phénomène, autrement dit que l'espérance de la FA reste constante.

– $\mathrm{Var}[Z(x+h) - Z(x)] = 2\gamma(h)$
Cette condition signifie que la variabilité entre deux points x et $x+h$ est une fonction, le *variogramme*, qui ne dépend que de la distance h séparant les points, et non de la position de x.

Toute fonction aléatoire stationnaire est également intrinsèque, l'inverse n'étant cependant pas vrai.

En toute généralité, nous avons pour première propriété d'une FAI que $\mathrm{E}[Z(x+h) - Z(x)] = m(h)$, où $m(h)$ représente une dérive linéaire. Nous distinguons le cas d'une dérive non nulle en le qualifiant de fonction aléatoire avec *dérive*. En effets, certains phénomènes n'ont pas une moyenne constante sur l'ensemble de leur champ d'étude ; cela est par exemple fréquemment le cas de la piézométrie dans l'étude d'une nappe aquifère.

C.4.2 Ergodicité

Formellement, on dira qu'un processus stationnaire satisfait à l'*hypothèse ergodique* (en moyenne) si la moyenne spatiale de Z sur un domaine D converge vers son espérance mathématique $E[Z]$ lorsque D tend vers l'infini

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

où $|D|$ représente la surface de D. Cette hypothèse est essentielle, car de sa validité dépend le fait que nous puissions, à partir d'une unique réalisation d'une fonction aléatoire stationnaire, approcher l'espérance de cette dernière par la moyenne d'une réalisation. Le plus fréquemment, nous ne disposons que d'une seule réalisation. Il nous est dès lors impossible de vérifier la validité de cette hypothèse ergodique. C'est par conséquent par *choix* que nous décidons de modéliser Z comme une fonction aléatoire ergodique d'espérance égale à la limite de l'intégrale d'espace C.1.

Plus que cette ergodicité en moyenne, il est également important de savoir si nous pouvons, à partir d'une seule réalisation, déterminer la covariance d'une fonction aléatoire. On parle d'*ergodicité d'ordre deux*. On montre sous certaines conditions que la convergence de $C(h)$ vers 0 lorsque $h \to \infty$ entraîne l'ergodicité d'ordre deux. De la même façon que pour l'ergodicité en moyenne, nous choisissons de considérer comme covariance sous-jacente la limite des covariances obtenues expérimentalement. Il y a cependant plus. En effet, nous travaillons sur un champ de taille finie, qu'il n'est pas possible de faire tendre vers l'infini. Dès lors, si le domaine n'est pas assez grand pour que l'on puisse observer aux grandes distances une covariance qui tend vers 0, cela signifie que l'estimation de la moyenne est difficile ; dans ces conditions, mieux vaut l'éviter tout à fait et ne considérer que des incréments de Z. Cela justifie fréquemment l'utilisation du variogramme plutôt que de la covariance [Chilès & Delfiner (1999)].

C.5 Variogramme

C.5.1 Variogramme expérimental, régional, théorique

Soit une variable régionalisée $z(x)$ sur un champ D, interprétée comme une réalisation d'une fonction aléatoire $Z(x)$, pour l'instant supposée intrinsèque. En théorie, ce dont nous avons besoin pour mener à bien les différentes estimations est le variogramme $\gamma(h)$ de la fonction aléatoire $Z(x)$

$$\gamma(h) = \frac{1}{2}\mathrm{Var}[Z(x+h) - Z(x)] \tag{C.2}$$

Ce *variogramme théorique* illustre comment évolue la dissimilarité entre $Z(x)$ et $Z(x+h)$ lorsque h grandit. Si cette dissimilarité ne dépend que du module de h et non de sa direction, le variogramme est *isotrope*. Cela signifie intuitivement que le phénomène se développe de façon analogue dans chaque direction. Il est simple de vérifier que, pour tout h, $\gamma(-h) = \gamma(h)$, $\gamma(h) \geq 0$ et $\gamma(0) = 0$. Par ailleurs, il est nécessaire d'imposer certaines conditions sur ce variogramme, qui ne peut être une fonction quelconque.

Ce variogramme théorique n'a cependant aucune existence réelle [Matheron (1965)]. On s'intéresse en pratique au variogramme de la variable régionalisée, ou *variogramme régional*

$$\gamma_R(h) = \frac{1}{2|D \cap D_{-h}|} \int_{D \cap D_{-h}} [z(x+h) - z(x)]^2 \, dx \ , \tag{C.3}$$

où $D \cap D_{-h}$ représente l'ensemble des points tels que x et $x+h$ appartiennent à D. C'est ce variogramme régional que nous désirons modéliser pour obtenir le variogramme théorique. Par définition, l'espérance du variogramme régional est le variogramme théorique $\gamma(h)$. $\gamma_R(h)$ est calculable si nous connaissons $z(x)$ en tout point de D, ce qui est en pratique extrêmement rare. En général, nous connaissons $z(x)$ en N points $x(i)$, $i = 1, \ldots, N$, et le variogramme régional est approché par le *variogramme expérimental*

$$\gamma^\star(h) = \frac{1}{2N(h)} \sum_{x_j - x_i \sim h} [z(x_j) - z(x_i)]^2 \ , \tag{C.4}$$

$N(h)$ représentant le nombre de couples de points distants de h. Si le nombre N de points est suffisant et que ceux-ci sont correctement répartis sur D, $\gamma^\star(h)$ sera proche de $\gamma_R(h)$. Plus formellement, si les données appartiennent à une grille régulière, le variogramme expérimental $\gamma^\star(h)$, calculé pour des paires séparées par des distances rigoureusement égales à h, est un estimateur sans biais de $\gamma_R(h)$ et de $\gamma(h)$. Néanmoins, si les données ne sont pas régulières, leur regroupement par classes de distances lors du calcul du variogramme expérimental peut introduire un biais.

Par opposition à d'autres estimateurs du variogramme expérimental souvent utilisés pour leur plus grande robustesse[2] et discutés au chapitre 2, ce variogramme est souvent appelé *estimateur classique du variogramme* ou de façon plus synthétique *variogramme classique*.

[2] Plutôt que ce calcul expérimental suivi d'un ajustement, il est également possible de travailler directement à partir des données. C'est par exemple le cas de la méthode du maximum de vraisemblance [Kitanidis & Lane (1985) cité dans Chilès & Delfiner (1999)] ; cependant, cette méthode repose sur des hypothèses gaussiennes et conduit souvent à des biais. Nous lui préférons le calcul expérimental d'un variogramme si possible robuste et ajusté manuellement, cet ajustement pouvant être a posteriori amélioré par validation croisée (voir paragraphe C.5.6).

Deux problèmes distincts sont associés aux versions expérimentale, régionale et théorique du variogramme :

- D'une part l'*estimation* du variogramme régional à partir du variogramme expérimental, déduit d'un nombre fini d'échantillons. Plus l'échantillonnage est dense, et plus on peut espérer que le variogramme expérimental soit proche du variogramme régional. L'efficacité de cette estimation dépendra également de la distribution de la variable régionalisée.
- D'autre part la *fluctuation* du variogramme régional par rapport à son espérance dans le modèle théorique, qui est indépendante de tout échantillonnage. Supposons en effet que nous connaissions exhaustivement la variable régionalisée sur le champ d'étude D. Cela impliquerait la connaissance du variogramme régional $\gamma_R(h)$ pour tout h. Ce dernier présentera certainement de nombreuses variations de détail, et il sera nécessaire d'en déduire une version "simplifiée", exprimable sous une forme simple : le variogramme théorique. Il est tout à fait envisageable qu'une variable régionalisée proche de la première conduise après calcul de son variogramme régional et simplification de ce dernier au même variogramme théorique que la première. C'est précisément le sens donné à la notion de fonction aléatoire dont nous possédons une réalisation, la variable régionalisée. Ces réalisations conduiraient à des variogrammes régionaux probablement légèrement différents mais ayant tous la même espérance : le variogramme théorique. Ces fluctuations sont importantes lorsque l'on s'intéresse à des simulations de la fonction aléatoire, afin de contrôler ces dernières. Cette problématique ne nous concernera que très peu, et nous ne nous attarderons pas sur ce point.

C.5.2 Modèles de variogrammes

Il existe différents modèles de variogrammes reproduisant différentes caractéristique structurales. Parmi les plus courants, on notera les modèles sphérique, exponentiel et gaussien qui font intervenir un paramètre a lié à la distance à laquelle ils se stabilisent et un second paramètre C correspondant au niveau de variabilité lorsque le modèle est stabilisé. Ils sont repris à la figure C.4, tout comme le modèle linéaire, exemple de modèle non stationnaire.

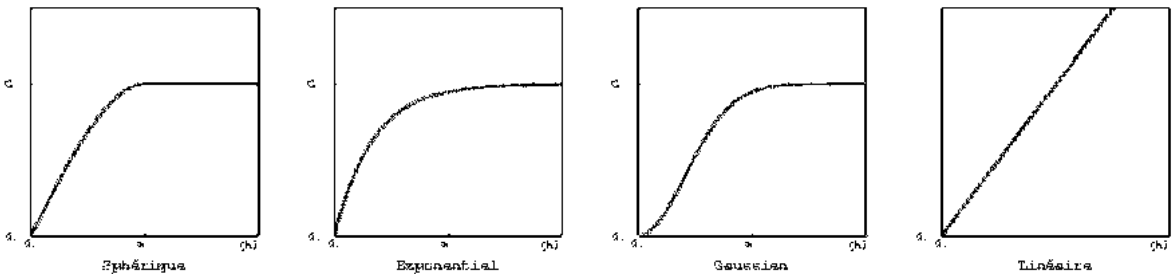


FIG. C.4 – Exemples de modèles de variogramme courants.

Un variogramme théorique peut combiner différentes structures à différentes échelles ; on parle de *structures emboîtées*.

C.5.3 Cas stationnaire

Plaçons-nous à présent dans le cas où $Z(x)$ est une FASt2. Comme cela a été vu, une FASt2 est caractérisée par sa moyenne m et sa covariance $C(h)$, où h est un vecteur. Tout comme pour le variogramme, cette covariance est dite isotrope si elle dépend uniquement du module de h et non de sa direction. Par ailleurs, $C(h) = C(-h)$ et $|C(h)| \leq C(0)$, qui est la variance de la FASt2. La covariance doit être une fonction définie positive, mais nous n'insisterons pas sur ce point ici.

Une fonction aléatoire stationnaire étant également intrinsèque, elle possède un variogramme. Ce dernier est lié à la covariance par la relation [Matheron (1965)]

$$\gamma(h) = C(0) - C(h) \tag{C.5}$$

où $C(0)$ est calculé une fois pour toutes, contrairement à $C(h)$. Dans ce contexte, la connaissance du variogramme est dès lors équivalente à celle de la covariance. Néanmoins, toute FASt2 étant une FAI, le variogramme est un outil plus général. Fréquemment, même dans le cas stationnaire, nous ne connaissons pas la moyenne et celle-ci est estimée à partir des données, ce qui peut introduire un biais si nous utilisons la covariance. Le variogramme, quant à lui, filtre cette moyenne et n'est donc pas affecté par ce biais. Ces raisons justifient l'emploi du variogramme comme outil structural même dans le cas d'une fonction aléatoire stationnaire d'ordre deux.

C.5.4 Pratique de l'analyse structurale

Supposons que l'on ait réalisé l'analyse exploratoire d'une variable régionalisée connue en quelques points. Considérons l'exemple d'un polluant analysé en dix points le long d'une ligne ; l'implantation des points et les concentrations analysées sont illustrées à la figure C.5.

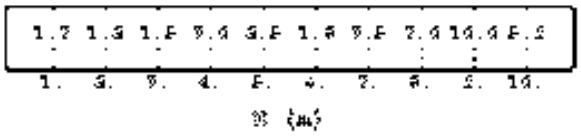
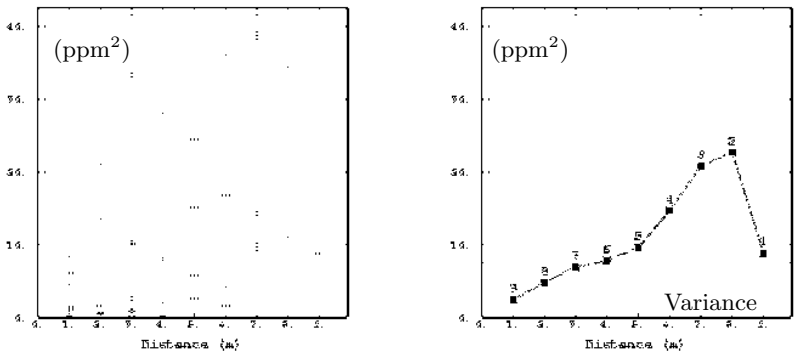


FIG. C.5 – Exemple de variable régionalisée à 1 dimension : concentration en polluant le long d'une ligne.

Nous intéressant à la variabilité entre les couples de points distants de h, il va tout d'abord être instructif de se faire une idée de cette variabilité en calculant, pour tous les couples de points, le carré de la différence de leur concentration. Autrement dit, cela revient à représenter la *nuée variographique* constituée des points $(|x_j - x_i|, \frac{1}{2}(z(x_j) - z(x_i))^2)$, pour tout $i, j = 1, \ldots, 10$ (voir figure C.6(a)). Le facteur $\frac{1}{2}$ est utilisé par référence à la définition du variogramme. La nuée variographique nous renseigne donc sur la dispersion des écarts entre les points en fonction de la distance qui les sépare. Elle permet aussi de détecter les couples apportant la contribution la plus importante au nuage. Ces couples font intervenir une valeur forte et une valeur faible. A plus d'une dimension, le calcul de cette nuée variographique dans différentes directions est une première manière de détecter des anisotropies, qui existent lorsque la dispersion du nuage diffère selon la direction.

Une fois la nuée variographique calculée, l'obtention du variogramme expérimental est immédiate. En effet, celui-ci n'est de par sa définition C.4 que la moyenne pour chaque classe de distance des $\frac{1}{2}(z(x_j) - z(x_i))^2$ de la nuée variographique. L'abscisse du variogramme expérimental est alors la moyenne des $|x_j - x_i|$ pour la classe de distance correspondante.



(a) Nuée variographique (b) Variogramme expérimental

FIG. C.6 – (a) Nuée variographique et (b) variogramme expérimental de la concentration, pas de calcul égal à 1 m.

Lorsque le pas d'échantillonnage est régulier, comme c'est le cas pour nos 10 points alignés, le choix du pas de calcul du variogramme est aisé ; on le prend égal au pas d'échantillonnage. Dans le cas contraire, il est nécessaire de choisir ce pas de façon judicieuse afin notamment que le nombre de points par classe de distances soit homogène. On calcule alors le variogramme expérimental pour chaque multiple du pas de calcul, la distance résultante étant assortie d'une tolérance, souvent égale à la moitié du pas.

C.5.5 Interprétation physique

Un variogramme présente en général les caractéristiques suivantes : il commence à l'origine car $\gamma(0) = 0$ et croît avec le module de h. Cette croissance peut se poursuivre, ou s'arrêter à un certain niveau de variabilité. Analysons cela plus en détail, sans oublier qu'en pratique l'analyse du variogramme se fait en relation avec la connaissance que l'on a du contexte de la variable - géologie par exemple.

Le comportement du variogramme à l'origine est lié au degré de continuité et de régularité de la variable. Bien que valant 0 à l'origine, il est possible que le variogramme tende vers une valeur non nulle C_0 lorsque h tend vers 0. Cette discontinuité porte le nom d'*effet de pépite* par référence au domaine minier, et peut avoir différentes origines :

– Existence d'une microstructure, c'est-à-dire d'une composante du phénomène d'échelle inférieure à la taille du support d'échantillonnage. Il en découle une très forte variabilité entre des échantillons même très proches. Le terme *effet de pépite* vient de là : en mine, dans l'étude de gisements d'or, on observe que l'or se présente généralement sous forme de pépites très pures et de taille réduite. Par conséquent, même à très petite distance, il existe une variabilité très

importante entre la teneur très élevée due à la pépite et les teneurs faibles environnantes.
- Structure existant à une échelle inférieure à la plus petite distance entre deux échantillons, et par conséquent dont l'échantillonnage ne peut rendre compte.
- Existence d'erreurs de mesure ou de positionnement des échantillons. Par exemple, il peut exister des erreurs lors de l'analyse de la concentration en un polluant.

Il arrive fréquemment que ces différentes origines coexistent, et seul un échantillonnage resserré au moins localement permet de distinguer l'importance de chacune.

Un variogramme ne présentant pas d'effet de pépite correspond à une variable spatialement continue. Son comportement à l'origine peut être parabolique, ce qui traduit une grande régularité de la variable régionalisée. En effet, intuitivement, cela signifie que près de l'origine la variabilité entre deux points ne grandit que très lentement lorsque la distance entre ces points augmente ; c'est le cas du modèle gaussien (voir figure C.4). Ce genre de comportement survient parfois dans l'étude de variables non stationnaires, comme typiquement la piézométrie d'une nappe. Il est néanmoins plus fréquent que l'on observe un comportement linéaire à l'origine, comme par exemple le modèle sphérique ou exponentiel.

Souvent, le variogramme se stabilise à un niveau de variabilité appelé *palier*, et cela pour une distance appelée *portée* du modèle. Cette portée correspond à la distance à partir de laquelle deux échantillons n'ont plus d'influence l'un sur l'autre.

C.5.6 Validation d'un modèle structural

Il est en pratique important de pouvoir évaluer la performance de l'ajustement d'un modèle de variogramme. Une méthode intéressante pour cela est la *validation croisée*. Pour un ensemble de données $(z(x_i))_{i=1,\dots,N}$, cette méthode consiste intuitivement à estimer la variable régionalisée successivement en chaque point x_i en retirant au préalable la donnée correspondante. On obtient une mesure de l'erreur d'estimation faite en ce point en considérant la différence entre l'estimation $z^\star(x_i)$ effectuée par krigeage des $N-1$ points restants et la valeur vraie $z(x_i)$. Les écart-types d'estimation σ_i sont également calculés. Il est alors possible de tester à l'aide de quelques graphiques si le modèle choisi est approprié :
- Le nuage des couples $(z_i, z_i^\star)_{i=1,\dots,N}$ entre les données vraies et estimées ; le modèle de variogramme sera d'autant meilleur que ce nuage est resserré autour de la première bissectrice.
- L'histogramme des erreurs standardisées $\frac{z_i - z_i^\star}{\sigma_i}$; cet histogramme sera idéalement symétrique et centré en 0.
- Le nuage entre valeurs estimées et erreurs standardisées.
- La carte d'implantation des données, avec indication des points pour lesquels l'erreur standardisée est forte.

Ces graphiques sont également un complément précieux à l'analyse exploratoire, permettant la détection de points mal expliqués par leur voisinage, d'anomalies.

Plusieurs statistiques sont calculables. Citons par exemple :
- L'erreur moyenne $\frac{1}{N}\sum_i (z_i - z_i^\star)$;
- L'erreur quadratique moyenne $\frac{1}{N}\sum_i (z_i - z_i^\star)^2$;

– L'erreur standardisée moyenne $\frac{1}{N}\sum_i \left(\frac{z_i - z_i^\star}{\sigma_i}\right)$;

– L'erreur quadratique standardisée moyenne $\frac{1}{N}\sum_i \left(\frac{z_i - z_i^\star}{\sigma_i}\right)^2$.

L'erreur moyenne et l'erreur standardisée moyenne permettent d'apprécier la qualité du modèle. La variance de l'erreur standardisée est également utile, car elle correspond au rapport entre les variances d'estimation expérimentale et théorique. On préconisera un modèle ayant une erreur faible et tel que la variance de l'erreur standardisée soit proche de 1. Dans le cas de variables fortement contrastées, ces statistiques devront néanmoins être utilisées prudemment à cause de leur manque de robustesse.

C.6 Estimation locale

Nous abordons ici un de nos objectifs essentiels : l'estimation locale de la variable, sa cartographie. L'estimateur géostatistique utilisé pour cela porte le nom de *krigeage*. Différentes expressions existent, selon les hypothèses faites sur la fonction aléatoire. Afin de ne pas alourdir cette partie, nous choisissons de présenter ce formalisme uniquement dans le cas d'une FASt2 de moyenne m constante mais inconnue. Nous présentons le cas d'une estimation ponctuelle, bien que le krigeage puisse être utilisé pour estimer la teneur sur des blocs v. La teneur moyenne estimée sur le bloc est en outre la moyenne des estimations des teneurs ponctuelles sur ce bloc :

$$Z^\star(v) = \frac{1}{v}\int_v Z^\star(x)\,dx$$

C.6.1 Etapes du krigeage

Supposons donc que nous cherchons à estimer Z en un point x. Compte-tenu de nos hypothèses, nous ne pouvons en géostatistique linéaire manipuler que des combinaisons linéaires de la fonction aléatoire étudiée. La première étape consiste donc à exprimer l'estimateur $Z^\star(x)$ qui nous intéresse comme combinaison linéaire des données disponibles (*contrainte de linéarité*)

$$Z^\star(x) = \sum_{i=1}^{N} \lambda_i Z(x_i)$$

L'étape suivante consiste à assurer que cet estimateur n'est pas biaisé. Cette *contrainte d'universalité* revient à exprimer que l'erreur d'estimation est d'espérance nulle

$$\mathrm{E}[Z^\star(x) - Z(x)] = 0$$

Elle conduit à la condition suivante sur les λ_i : $\sum_{i=1}^{N} \lambda_i = 1$

La dernière étape consiste à trouver, parmi toutes les combinaisons linéaires vérifiant la contrainte d'universalité, celle qui minimise la variance de l'erreur d'estimation.

$$\mathrm{Var}\left[\sum_{i=1}^{N} \lambda_i Z(x_i) - Z(x)\right] = C(0) - 2\sum_{i=1}^{N} \lambda_i C(x_i - x) + \sum_{i,j=1}^{N} \lambda_i \lambda_j C(x_i - x_j)$$

Cette variance est couramment appelée *variance d'estimation* ; elle ne dépend que du modèle structural choisi et de la géométrie de l'échantillonnage[3]. En introduisant le paramètre de Lagrange μ, on montre [voir par exemple Chauvet (1999)] que les poids λ_i cherchés sont solutions du *système de krigeage*

$$\begin{cases} \sum_{j=1}^{N} \lambda_j C(x_i - x_j) + \mu = C(x_i - x) & i = 1, \dots N \\ \sum_{i=1}^{N} \lambda_i = 1 \end{cases}$$

En conclusion, le krigeage permet de donner aux divers échantillons les poids les meilleurs en fonction de la structure de la variable et de leur configuration géométrique. En remplaçant dans l'expression de la variance d'estimation les poids par les solutions du système de krigeage, on obtient la *variance de krigeage*. Cela est essentiel, car la variance de krigeage nous renseigne sur l'incertitude liée à l'estimation.

En pratique, il peut être intéressant de ne pas considérer toutes les données pour l'estimation en un point x, mais uniquement celles situées dans un *voisinage glissant* centré en x. C'est notamment le cas lorsque l'on est confronté à un nombre de données élevé, le temps de calcul pouvant devenir prohibitif. Rien n'empêche que le krigeage conduise en certains points à une estimation négative, même lorsque la variable est positive comme c'est le cas pour une concentration. Ce point est à surveiller, un modèle conduisant à une proportion de valeurs estimées négatives trop importante n'étant pas approprié ; à cette fin, l'observation des variances de krigeage est instructive.

C.6.2 Quelques propriétés

Tout d'abord, le krigeage ponctuel est un *interpolateur exact*. Cela signifie que si l'on veut procéder à l'estimation en un point confondu avec un point de donnée, cette estimation par krigeage sera identique à la valeur de la donnée. La variance de krigeage est alors nulle en ce point.

Il est important de réaliser que "le krigeage lisse". Cela signifie que la variable régionalisée estimée $z^\star(x)$ présente moins de fluctuations et d'aspérités que la variable régionalisée vraie $z(x)$ [Chauvet (1999)]. Cela est voulu, car l'objectif d'une estimation n'est pas de reproduire la variabilité de la variable - les simulations existent pour cela - mais de donner en chaque point, compte tenu de la structure et de la configuration des échantillons, la meilleure estimation linéaire de cette variable, c'est-à-dire celle qui minimise la variance de l'erreur d'estimation.

C.7 Aspects multivariables

La géostatistique multivariable permet de tirer profit de corrélations entre variables régionalisées afin d'en améliorer l'estimation.

[3] La variance d'estimation, qui renseigne sur la précision de l'estimation d'une quantité par une autre, est à distinguer de la *variance de dispersion*, qui décrit la variabilité des valeurs à l'intérieur d'un domaine.

Soient z_1 et z_2 deux variables régionalisées. L'estimateur de Z_1 par *cokrigeage*, analogue multivariable du krigeage, prend la forme

$$Z_1(x)^{CK} = \sum_i \lambda_{1i} Z_1(x_i) + \sum_j \lambda_{2j} Z_2(x_j)$$

La connaissance des variogrammes de chacune des variables n'est plus ici suffisante, et il est nécessaire de modéliser la structure croisée entre les variables, par l'intermédiaire de leur *variogramme croisé*

$$\gamma_{12}(h) = \frac{1}{2}\mathrm{E}[(Z_1(x+h) - Z_1(x))(Z_2(x+h) - Z_2(x))]$$

Contrairement aux variogrammes simples, ce variogramme croisé peut prendre des valeurs négatives ; c'est le cas par exemple pour des variables anti-corrélées.

Le cokrigeage permet notamment d'améliorer l'estimation d'une variable par l'utilisation d'une autre qui serait mieux échantillonnée. Nous devons ici prendre garde à ce que les ajustements de variogrammes ou covariance croisés n'entraînent pas de variances négatives. Le recours à des *modèles de corégionalisation* permet de garantir cela. Le *modèle linéaire de corégionalisation*, qui est le plus simple, consiste à décomposer l'outil structural utilisé pour chacune des variables en composantes élémentaires. Les structures croisées entre deux variables s'expriment alors comme une combinaison linéaire des composantes élémentaires des deux variables, où cependant les composantes propres uniquement à l'une des deux variables n'apparaissent pas. Le principe est généralisable à plus de deux variables.

C.8 Méthodes non linéaires

La géostatistique linéaire permet, en manipulant exclusivement des combinaisons linéaires de la variable étudiée, d'estimer celle-ci en chaque point du champ d'étude à partir de données expérimentales, et de fournir une variance de l'erreur d'estimation associée. Cependant, elle devient généralement insuffisante lorsque l'on s'intéresse non plus à la variable $Z(x)$ échantillonnée mais à une fonction $f[Z(x)]$ de celle-ci. Considérons par exemple une variable $Z(x)$ pouvant prendre les valeurs 0, 1, 2 ou 3. Supposons qu'en un point x l'estimation $Z^\star(x)$ vale 1. Si la variance de l'erreur d'estimation est faible, il est fort probable que la valeur vraie $Z(x)$ ne dépasse pas 2. Mais si au contraire cette variance est élevée, la valeur vraie peut aussi bien valoir 1 que 0, 2 ou 3. Le fait que $Z^\star(x)$ soit inférieur à 2 ne signifie donc pas que la vraie valeur $Z(x)$ ne puisse être supérieure à 2, ce dernier évènement étant représenté par l'indicatrice $\mathbb{1}_{Z(x)\geq 2}$ [d'après Rivoirard (1991)].

Plus généralement, on s'intéressera au dépassement d'un seuil s fixé par la variable *réelle*, et non par son estimation ; ce dépassement est représenté par l'indicatrice

$$\mathbb{1}_{Z(x)\geq s} = \begin{cases} 1 & \text{si } Z(x) \geq s \\ 0 & \text{si } Z(x) < s \end{cases}$$

Afin d'alléger les notations, nous noterons cette indicatrice $\mathbb{1}_s(x)$ lorsque cela ne sera pas source d'ambiguïté. Nous nous fixons donc comme objectif, pour une fonction aléatoire **stationnaire** $Z(x)$, l'estimation du dépassement d'un seuil s à partir d'observations $(Z_i)_{i=1,\dots,N}$. Nous nous

intéressons d'abord à un dépassement *ponctuel*, avant d'envisager un *modèle de changement de support* nécessaire à un calcul par bloc.

Pour estimer cette indicatrice, une première solution consiste à procéder au krigeage de la FA indicatrice $\mathbb{1}_s(x)$. Malgré sa séduisante simplicité apparente, ce *krigeage d'indicatrice* possède certains inconvénients : outre la perte d'information que le passage en indicatrice génère à la base, il nécessite de réaliser la modélisation du variogramme de l'indicatrice chaque fois que le seuil est modifié et ne garantit pas la cohérence de l'estimation entre les différents seuils : pour un seuil $s' > s$, rien ne garantit que $\mathbb{1}^\star_{s'}(x) \leq \mathbb{1}^\star_s(x)$. Afin de remédier à cela, il est nécessaire de considérer simultanément les indicatrices à différents seuils ; néanmoins, ce *cokrigeage d'indicatrices* nécessite la modélisation des covariances simples et croisées correspondant aux indicatrices aux différents seuils, modélisation assez lourde et complexe qui est à l'origine de sa rare utilisation en pratique.

Or, la connaissance des covariances simples et croisées pour tous les seuils possibles est équivalente à la connaissance de la loi bivariable $\mathrm{P}[Z(x) < s \text{ et } Z(x+h) < s']$ du couple $(Z(x), Z(x+h))$. Donc, outre la modélisation du variogramme ou de la covariance, nous allons avoir ici besoin de préciser la loi spatiale bivariable, qui va en particulier décrire comment s'effectue la transition entre valeurs faibles et valeurs fortes. Il est primordial que la modélisation soit *adaptée* au phénomène décrit. De ce point de vue, un premier guide dans le choix d'un modèle est l'existence d'*effets de bord* (voir figure C.7), i.e. l'existence d'un passage progressif des valeurs faibles à fortes. Reprenons l'exemple de la fonction aléatoire de valeurs 0, 1, 2 et 3. En notant $\gamma_1(h)$ le variogramme de l'indicatrice $\mathbb{1}_{Y(x)\geq 1}$ et $\gamma_{12}(h)$ le variogramme croisé des indicatrices $\mathbb{1}_{Y(x)\geq 1}$ et $\mathbb{1}_{Y(x)\geq 2}$, on peut montrer que pour $h \neq 0$

$$\mathrm{P}[Y(x) \geq 2 \,|\, Y(x) \geq 1, Y(x+h) < 1] = \frac{\gamma_{12}(h)}{\gamma_1(h)}$$

Cette grandeur est la probabilité, sachant que le couple $(x, x+h)$ recouvre la coupure 1, que x dépasse également la seconde coupure 2. S'il y a existence d'effets de bord, on attend de cette grandeur qu'elle augmente avec h tant que la portée n'est pas atteinte. Par conséquent, il est possible de tester grâce aux variogrammes simples et croisés d'indicatrices l'existence d'effets de bord [Rivoirard (1995)].

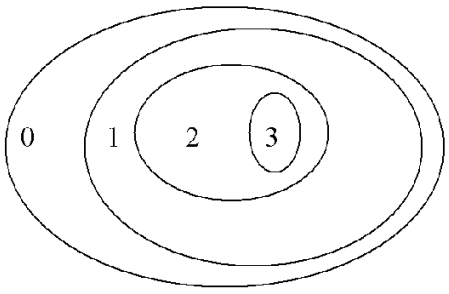


FIG. C.7 – Exemple de variable présentant des effets de bord.

Moyennant cette modélisation de la loi spatiale bivariable, nous obtiendrons une première méthode d'estimation de l'indicatrice : le *krigeage disjonctif*. Différents *modèles de diffusion* existent pour la loi spatiale en présence d'effets de bord, parmi lesquels le plus simple est le modèle gaussien. Il est également possible de n'avoir d'effets de bord que dans un sens, en allant vers les valeurs

fortes ou en les quittant - modèle à résidus d'indicatrices orthogonaux -, ou bien encore pas d'effets de bord du tout - modèle mosaïque - [Rivoirard (1991)]. Nous présentons ici le principe du krigeage disjonctif pour un modèle bigaussien. $Z(x)$ n'étant dans la grande majorité des cas pas de distribution gaussienne, il sera au préalable nécessaire de se ramener à ce cas favorable, au moyen d'une transformation, appelée *anamorphose gaussienne*.

C.8.1 Anamorphose

Intuitivement, l'anamorphose gaussienne est une transformation consistant à déformer l'histogramme de la variable étudiée $Z(x)$ pour se ramener à un histogramme gaussien réduit. Supposant que sa distribution n'est pas gaussienne, on considère la FA stationnaire $Z(x)$ comme une fonction de la gaussienne centrée réduite $Y(x)$:

$$Z(x) = \Phi[Y(x)]$$

où la fonction d'anamorphose Φ peut se déterminer par les coefficients ϕ_n de son développement en polynômes d'Hermite :

$$Z(x) = \sum_{n=0}^{+\infty} \phi_n \, H_n[Y(x)]$$

En considérant les d valeurs expérimentales $(z_i)_{1 \leq i \leq d}$ de $Z(x)$ rangées par ordre croissant, on peut associer à leurs fréquences cumulées des valeurs gaussiennes y_i de même fréquence cumulée

$$\mathrm{P}[Z(x) < z_i] = \sum_{j=1}^{i-1} \underbrace{\mathrm{P}[Z(x) = z_j]}_{=p_j} = G(y_i)$$

Il en découle que $Z(x)$ vaudra z_i lorsque $y_i < Y(x) < y_{i+1}$, ce qui permet également de déterminer les coefficients ϕ_n de l'anamorphose. Connaître les valeurs anamorphosées $y_i = \Phi^{-1}(z_i)$ associées aux teneurs est parfois indispensable, notamment pour réaliser un krigeage de $Y(x)$. La connaissance des y_i permet également de tester la validité de l'hypothèse bigaussienne sous-jacente au modèle gaussien anamorphosé présenté ci-dessous. La valeur choisie pour $y_i = \Phi^{-1}(z_i)$ est

$$y_i = \mathrm{E}[Y(x) \mid \underbrace{u_{i-1}}_{=G^{-1}\left(\sum_{j=1}^{i-1} p_j\right)} \leq Y(x) < u_i] \qquad (1 \leq i \leq n)$$

soit la valeur moyenne de G sur $]u_{i-1}\,,\, u_i[$. En particulier,

$$y_1 = \mathrm{E}[Y(x) \mid Y(x) < G^{-1}(p_1)]$$

Ce choix permet de conserver la moyenne nulle de g. Cependant, les variances résultantes peuvent être légèrement inférieures à 1, en particulier en présence d'atome. Lorsque la variable d'intérêt présente une proportion p non négligeable de valeurs z_1 nulles - effet zéro - ou égales à un seuil de détection, il convient de modéliser cela de façon appropriée. Un moyen de le faire est de considérer une fonction d'anamorphose Ψ du type

$$Z(x) = \Psi[Y(x)] = \mathbb{1}_{Y(x) > y_1} \, . \, \Phi[Y(x)]$$

où $y_1 = G^{-1}(p)$ est la valeur du seuil de détection sur la gaussienne et Φ est une fonction d'anamorphose sur $\{y \text{ t.q. } y > y_1\}$ telle que $\Phi(y_1) = z_1$ [Freulon (1992)].

C.8.2 Modèle gaussien anamorphosé

Si la variable gaussienne obtenue par anamorphose obéit à une hypothèse bigaussienne, *i.e.* si les lois bivariables $(Y(x), Y(x+h))$ sont bigaussiennes, la covariance $C(h)$ de Z se déduit de celle $\rho(h)$ de Y par la relation

$$C(h) = \sum_{n\geq 1} \phi_n^2 \, [\rho(h)]^n \tag{C.6}$$

car $\mathrm{Cov}[H_n(x), H_n(x+h)] = [\rho(h)]^n$. Les $[\rho(h)]^n$ se destructurant rapidement lorsque n augmente, seul un petit nombre (au plus quelques dizaines) de polynômes est nécessaire en pratique.

Une fois les coefficients d'anamorphose déterminés, la connaissance de $C(h)$ devient équivalente à celle de $\rho(h)$; plutôt que d'ajuster directement $C(h)$ à partir des structures expérimentales, il est préférable pour assurer la cohérence du modèle de spécifier la structure gaussienne en s'assurant que la structure $C(h)$ correspondante ajuste bien les structures expérimentales [Liao (1990)]. Ainsi, si nous ne nous intéressons qu'à la structure spatiale de $Y(x) = \Phi^{-1}[Z(x)]$, on peut éviter le calcul des valeurs anamorphosées.

C.8.3 Krigeage disjonctif

Dans le modèle gaussien anamorphosé, si nous considérons un seuil z_c sur $Z(x) = \Phi[Y(x)]$, alors pour le seuil correspondant sur la transformée gaussienne $y_c = \Phi^{-1}(z_c)$ nous avons que

$$Z(x) \geq z_c \Leftrightarrow Y(x) \geq y_c$$

Autrement dit, $\mathbb{1}_{Z(x)\geq z_c} = \mathbb{1}_{Y(x)\geq y_c}$ et on montre que le krigeage disjonctif (KD) recherché s'écrit

$$\left[\, \mathbb{1}_{Y(x)\geq y_c} \right]^{KD} = 1 - G(y_c) - \sum_{n\geq 1} \frac{1}{\sqrt{n}} \, H_{n-1}(y_c)\, g(y_c) \; [H_n[Y(x)]]^K$$

Seul le krigeage des polynômes d'Hermite est donc nécessaire à l'obtention de l'estimateur par KD. Si la stationnarité semble trop exigeante, il est possible d'introduire une condition de non-biais. Néanmoins, l'optimalité de l'estimateur n'est alors plus garantie [Rivoirard (1991)].

C.8.4 Espérance conditionnelle

Le meilleur estimateur de l'indicatrice $\mathbb{1}_s(x)$ est l'espérance conditionnelle

$$\mathrm{E}[\, \mathbb{1}_{Z(x)\geq s} \,|\, Z(x_i) = z_i, i = 1, \ldots, N] = \mathrm{P}[Z(x) \geq s \,|\, Z(x_i) = z_i, i = 1, \ldots, N]$$

Cette espérance conditionnelle n'est en pratique calculable que dans le cas d'une fonction aléatoire *multigaussienne*, i.e. pour laquelle toute combinaison linéaire des $Z(x)$, $Z(x_1)$, ..., $Z(x_N)$ est encore gaussienne. En considérant une gaussienne réduite U, on peut alors écrire

$$\begin{aligned} [\, \mathbb{1}_{Y(x)\geq y_c}]^{EC} &= \mathrm{E}[\, \mathbb{1}_{Y(x)\geq y_c} \,|\, Y(x_i) = z_i, i = 1, \ldots, n] \\ &= 1 - G\left(\frac{y_c - Y^K(x)}{\sigma^K(x)} \right) \end{aligned} \tag{C.7}$$

L'avantage de ce calcul est son extrême rapidité, ne nécessitant que le krigeage de la transformée gaussienne ; le prix à payer est cependant une hypothèse nettement plus forte sur la loi spatiale que pour le krigeage disjonctif. La validation pratique de cette hypothèse est délicate, et il est fréquent que l'on s'en tienne seulement à vérifier que les lois bivariables sont bigaussiennes. Une hypothèse de stationnarité est par ailleurs nécessaire, excepté dans le cas lognormal pour lequel il est possible d'introduire une condition de non-biais.

C.8.5 Modèle gaussien discret

Si, plutôt que d'estimer une probabilité de dépassement ponctuel d'un seuil nous nous intéressons à cette probabilité pour un support de taille supérieure - par exemple cohérent avec une méthode de dépollution à mettre en œuvre -, il est nécessaire de prendre en compte l'effet de support vu précédemment en envisageant un modèle de changement de support. La présentation ci-dessous d'un tel modèle, le *modèle gaussien discret*, est pour l'essentiel issue de Lantuéjoul (1990).

Nous avons vu précédemment que $Z(x)$ peut s'exprimer comme une fonction $\Phi[Y(x)]$ de la gaussienne réduite $Y(x)$, où la fonction d'anamorphose Φ peut se déterminer par les coefficients ϕ_n de son développement en polynômes d'Hermite

$$Z(x) = \sum_{n=0}^{+\infty} \phi_n \, H_n[Y(x)]$$

De la même façon, si nous considérons à présent x uniforme dans un bloc v, il existe une fonction d'anamorphose de bloc Φ_v telle que $Z(v) = \Phi_v(Y_v)$, où Y_v est une variable gaussienne centrée réduite.

Le modèle gaussien discret repose sur l'hypothèse que le couple (Y, Y_v) suit une loi bigaussienne. La détermination de ce modèle nécessite l'évaluation de la fonction d'anamorphose de bloc Φ_v de $Z(v)$ ainsi que du coefficient de corrélation r entre Y et Y_v. On montre que Φ_v se décompose également en polynômes d'Hermite :

$$\Phi_v(y) = \sum_{n=0}^{+\infty} \phi_n \, r^n \, H_n(y)$$

D'autre part, sachant que la variance de $Z(v)$ est connue :

$$\mathrm{Var}[Z(v)] = \mathrm{Var}\left[\frac{1}{|v|}\int_v Z(x)\,dx\right] = \frac{1}{|v|^2}\int_v\int_v C(x-y)\,dx\,dy$$

il est possible de déterminer le coefficient r, qui doit être positif, par la formule

$$\mathrm{Var}[Z(v)] = \mathrm{Var}[\Phi_v(Y_v)] = \sum_{n=1}^{+\infty} \phi_n^2 r^{2n}$$

Nous pouvons alors procéder à l'estimation de l'indicatrice $\mathbb{1}_{Z(v)>z_s}$ du dépassement d'un seuil z_s sur un bloc, moyennant la spécification des covariances entre point et bloc, nécessaires pour tenir

compte de la différence de support. Ces covariances se déduisent des covariances ponctuelles, et par suite il est également possible d'obtenir celles des gaussiennes associées, qui permettent l'estimation par krigeage disjonctif.

Par ailleurs, l'espérance conditionnelle s'obtient de la même façon que dans l'équation C.8 correspondant au cas ponctuel, en y remplaçant le krigeage simple de la gaussienne par celui de la gaussienne des blocs $Y^K(v)$:

$$[\mathbb{1}_{Y(v)\geq y_{s_v}}]^{EC} = 1 - G\left(\frac{y_{s_v} - Y^K(v)}{\sigma^K(v)}\right)$$

Krigeage disjonctif et espérance conditionnelle nécessitent une hypothèse de stationnarité dès que l'on est en présence d'un changement de support. Le modèle gaussien discret n'est par ailleurs plus valide dans le cas d'un effet zéro important, qui doit donc être traité au préalable.

Annexe D

Données du Pôle de Compétences

Sommaire

Outre celle du CNRSSP, différentes campagnes d'échantillonnage ont été réalisées dans le cadre du Pôle de Compétence Nord-Pas-de-Calais sur les Sites et Sols Pollués : analyses de gaz, campagne géophysique, reconnaissance par sondages profonds le long de profils. L'objectif de cette annexe est de les présenter et de les comparer aux analyses "classiques".

Plusieurs campagnes d'échantillonnage ont été réalisées sur l'ancienne cokerie du site X dans le cadre du Pôle de Compétence Nord-Pas-de-Calais sur les Sites et Sols Pollués [Dubourguier (1999)] : reconnaissance par analyse de gaz passive et active, campagne géophysique dense et réalisation de 44 sondages. Nous nous attachons ici à leur comparaison avec les données régulières du CNRSSP. En effet, bien que ces campagnes n'aient pas été réalisées au même moment, il est instructif de comparer leurs résultats dans la mesure où nous nous intéressons essentiellement aux HAP, polluants qui - exception faite des plus légers - ont tendance à rester en place.

La figure D.1 illustre l'implantation des campagnes d'échantillonnage du Pôle et du CNRSSP. Celles-ci diffèrent sensiblement, et rares sont les échantillons proches entre le CNRSSP et les données du Pôle ; connaissant la variabilité des concentrations en HAP, même à petite échelle, nous pouvons d'emblée noter qu'il sera délicat d'établir des comparaisons fines. En effet, une migration[1] des données du CNRSSP vers la grille des données du Pôle est nécessaire. Une telle migration à 5 m fournit uniquement 16 points d'échantillonnage communs, tandis qu'une migration à 10 m conduit à 48 points. Nous envisagerons les deux cas, malgré les réserves que nous avons émises. Vu le nombre de points de comparaison et les conditions de comparaison de ces données, nous nous limiterons aux résultats qui semblent significatifs. Rappelons que sur les données CNRSSP les zones les plus polluées se situaient aux alentours des points 55, 37 et ses voisins vers le Sud-Est ainsi qu'au Sud-Ouest.

[1] Par *migration*, nous entendons ici le déplacement des échantillons d'une des deux grilles d'échantillonnage sur l'autre d'une distance inférieure à un seuil fixé, afin d'autoriser le calcul de corrélations entre les deux grilles pour ces points.

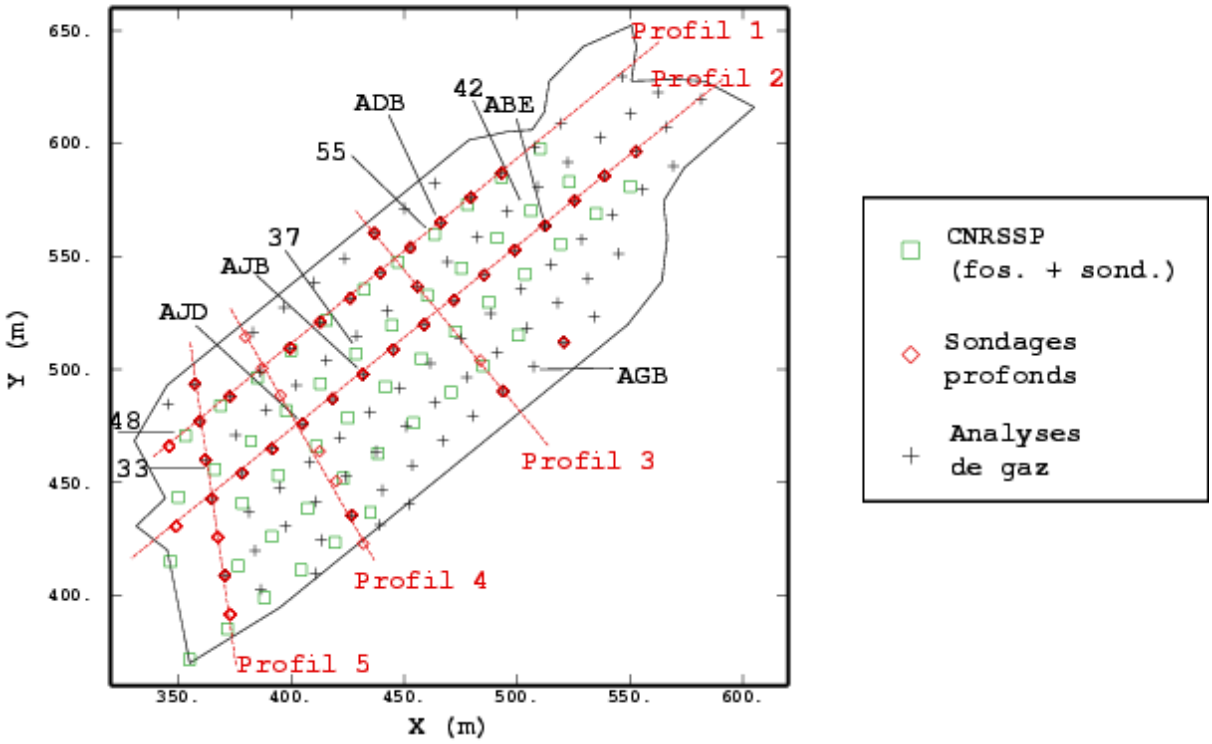


FIG. D.1 – Implantation comparée des différentes campagnes d'échantillonnage, avec indication de certains points.

D.1 Analyses de gaz

Les deux campagnes d'analyse de gaz ont été réalisées selon un même maillage triangulaire de 15 m (hauteur de maille) recouvrant le site. La première technique, passive (GORE), consiste à poser des cartouches dans le sol et à laisser les gaz s'adsorber pendant une période allant de quelques jours à quelques semaines. Pour la méthode active (CREID), les gaz sont pompés pendant quelques minutes. Les concentrations en HAP et BTEX[2] ont été analysées dans les deux cas. Les points ADB et AJD ressortent particulièrement pour les deux méthodes.

La comparaison des analyses de gaz CREID et GORE avec les analyses de gaz prélevées sur les mini-fosses lors de la campagne du CNRSSP n'est pas pertinente, ces dernières s'étant avérées fortement dépendantes des conditions climatiques (voir paragraphe 7.2.3). Nous comparons les mesures d'analyse de gaz aux accumulations entre 0 et 1.5 m des échantillons de sol du CNRSSP ; en effet, la profondeur d'investigation des mesures de gaz étant d'environ 1.50 m, nous ne pouvons les comparer aux fosses (0 - 0.80 m) ou aux sondages (0.50 - 1.50 m) CNRSSP seuls. Nous considérons donc pour le CNRSSP les variables

$$\mathrm{HAP}_{\mathrm{accumulé}} = [\mathrm{HAP}]_{\mathrm{fosse}} \times \mathrm{Profondeur}_{\mathrm{fosse}} + [\mathrm{HAP}]_{\mathrm{sondage}} \times \mathrm{Hauteur}_{\mathrm{sondage}}$$

[2]Les BTEX sont des hydrocarbures aromatiques composés d'un seul cycle de benzène ; ils sont plus légers que les HAP, plus volatils et nettement plus solubles [Fetter (1993)].

représentatives de la profondeur 0 - 1.50 m. Pour l'implantation des points chauds, nous observons une bonne concordance entre les concentrations des points 55 (CNRSSP) et ADB (CREID-GORE) (voir figure D.1). Les résultats sont moins probants pour les autres zones polluées. Les meilleures corrélations, à la fois pour une migration à 5 m et à 10 m, sont systématiquement dûes à un seul couple de points : ADB-55 ou AJB-37, ce qui n'est pas étonnant vu l'hétérogénéité des zones de concentrations fortes.

Concernant les structures spatiales, seule la méthode active CREID présente, pour les HAP Fle et Phe (3 cycles), une structure nette en logarithme translaté. Les accumulations de ces variables présentent pour les données CNRSSP une structure pépitique.

D.2 Données géophysiques

La géophysique commence à être fréquemment utilisée lors d'études de pollution. Tandis que celle-ci a fait ses preuves pour la détection d'objets ou d'anomalies - fûts, fissures dans des réseaux, etc -, il n'existe dans la littérature que peu de validations de son efficacité pour la quantification d'une pollution diffuse sur un site.

La conductivité apparente du terrain a été mesurée en 936 points sur une maille triangulaire de hauteur égale à 5 m recouvrant l'ensemble du site, soit un échantillonnage très dense (voir figure D.2). La mesure a été effectuée par un appareil EM31, selon deux profondeurs d'exploration : 3 m et 6 m, correspondant aux lettres H et V, et selon deux orientations afin de tenir compte d'éventuelles anisotropies, notées 1 et 2.

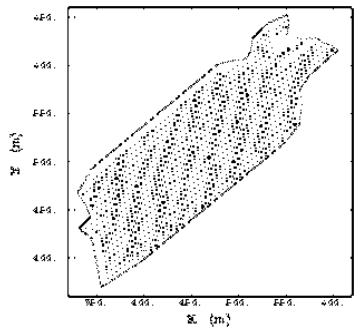


FIG. D.2 – Implantation des données de conductivité.

Comme l'illustrent le tableau D.1 et la figure D.3, il n'y a aucune corrélation entre les données du CNRSSP et les données de conductivité 0-3 m - horizontales H1 et H2, moyenne Hm, l'anisotropie Hani se mesurant comme la différence en valeur absolue entre H1 et H2 divisée par la plus grande des deux valeurs. Les concentrations élevées sont sans anomalie de conductivité associée, excepté pour quelques points présentant une anomalie de conductivité dûe à la proximité d'infrastructures liées à la pollution.

HAP	H1	H2	Hani	Hm
Nap	0.16	0.04	-0.08	0.12
Ant	0.11	0.13	-0.02	0.14
Phe	0.07	0.00	-0.08	0.04
Bap	0.19	0.21	-0.01	0.24
Inp	0.23	0.23	-0.04	0.27
Phl	0.39	0.14	0.01	0.31

TAB. D.1 – Corrélations entre données de conductivité 0-3 m et données CNRSSP.

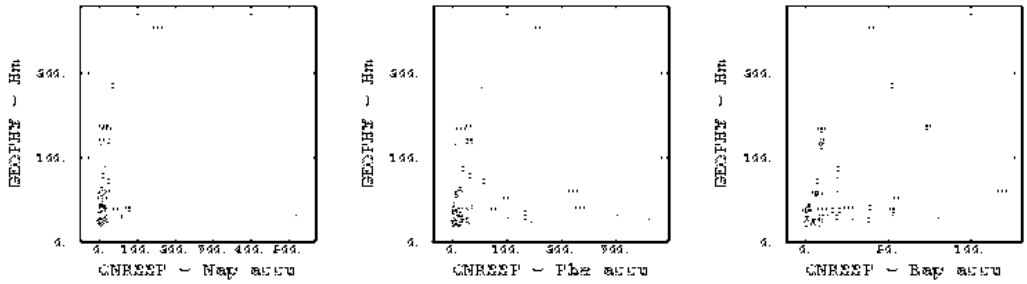


FIG. D.3 – Nuages de corrélation entre accumulation de trois HAP (CNRSSP) et conductivité moyenne 0-3 m.

D.3 Sondages ISA

Les cinq axes de prélèvement, définissant 44 points de sondages (voir figure D.1), ont été choisis après la phase d'exploration rapide que constituent l'acquisition des données de gaz et de conductivité. Les sondages ISA présentent des traces importantes de pollution jusqu'à 5 m. Cela montre qu'il est illusoire d'espérer obtenir un tonnage des zones contaminées en HAP sur le site à partir des données du CNRSSP, qui s'arrêtent à 1.5 m. Qualitativement, quelques similitudes entre les zones de fortes concentrations des données CNRSSP et des sondages ISA sont reprises au tableau D.2.

CNRSSP	Sondages ISA
Gazomètre le plus au Nord (fosses 33, 48)	Extrémités Sud-Ouest des profils 1 et 5 (éch. pollués en surface)
42 pollué (sondages) pour HAP légers	ABE (Profil 2) pollué en surface
Point 37 pollué	AJB (Profil 2) pollué en surface pour Bap

TAB. D.2 – Similitude de zones polluées entre données CNRSSP et sondages ISA.

La différence entre les grilles d'échantillonnage ne permet pas de comparer quantitativement les deux types de mesures : la migration à 5 m ne laisse que 6 données communes, celle à 10 m 17 données. Aucune structure spatiale n'est mise en évidence sur l'accumulation entre 0 et 1.5 m des données ISA.

D.4 Synthèse

Tout d'abord, il est important de noter que le choix de ce site pour un travail de comparaison entre diverses méthodes d'échantillonnage n'était peut-être pas le plus favorable, de par l'importance des remaniements successifs qui y ont eu lieu. Ensuite, il est dommage que les différentes campagnes n'aient pas été réalisées sur des grilles d'échantillonnage compatibles. En effet, nous pouvons avoir pour objectif de comparer différentes méthodes sur une même grille, ou inversement vouloir comparer pour une même méthode différentes grilles d'échantillonnage ; **il n'est cependant pas étonnant que la comparaison de différentes méthodes réalisées sur des grilles d'échantillonnage différentes ne permette que des considérations qualitatives, vu l'hétérogénéité des concentrations.**

Les méthodes semi-quantitatives d'analyse de gaz retrouvent bien certains des pics de pollution les plus élevés, mais il n'est pas raisonnable d'en attendre plus. Par ailleurs, la mise en œuvre de la géophysique apporte peu d'information pour le suivi de polluants organiques sur ce site, ne détectant que les infrastructures ; la connaissance de ces dernières présente néanmoins l'intérêt de compléter une information historique souvent assez pauvre sur ces sites ayant un long passé industriel. Il existe finalement une cohérence qualitative entre les points chauds détectés par les sondages ISA et les données CNRSSP.

Caractérisation géostatistique de pollutions industrielles de sols
Cas des hydrocarbures aromatiques polycycliques
sur d'anciens sites de cokeries

L'estimation des concentrations en hydrocarbures aromatiques polycycliques dans les sols de friches industrielles présente de nombreuses difficultés pratiques, liées aux propriétés des polluants et à la formation de ces sites :

- prélèvement et préparation des échantillons dans des sols fortement hétérogènes,
- forte variabilité à petite distance, notamment pour des terrains remaniés,
- fort contraste des teneurs, compliquant l'inférence du variogramme.

Actuellement, la reconnaissance est souvent guidée par l'historique du site, et la recommandation d'une reconnaissance systématique préconisée pour les estimations géostatistiques apparaît souvent excessive.

En s'appuyant sur l'étude détaillée de deux anciens sites de cokeries, différentes méthodes d'estimation géostatistiques sont comparées pour l'estimation (i) des concentrations en place et (ii) de la probabilité de dépassement d'une valeur guide. De nombreuses questions pratiques ou méthodologiques sont examinées :

- propriétés de différents modes de calcul du variogramme expérimental et validité des résultats ;
- utilisation d'informations auxiliaires telles que l'historique de site, des relevés organoleptiques, des mesures semi-quantitatives, en vue d'améliorer la précision des estimations ;
- discussion des plans d'échantillonnage usuels, au vu de la répartition verticale des teneurs ou de l'historique du site.

La mise à disposition de mesures multiples à partir d'un même prélèvement permet d'approcher l'ordre de grandeur de l'erreur d'échantillonnage - au sens large. Des reconnaissances à petite distance montrent les difficultés d'un tri sélectif des terres en l'absence de structure spatiale. Plusieurs études de sensibilité sont menées en vue de quantifier l'apport d'une information auxiliaire dense pour l'estimation des teneurs.

En se basant principalement sur des modélisations existantes, ce travail vise à fournir au praticien des recommandations pratiques pour la caractérisation de pollutions de sols.

Geostatistical characterization of soil pollution at industrial sites
Case of polycyclic aromatic hydrocarbons at former coking plants

Estimating polycyclic aromatic hydrocarbons concentrations in soil at former industrial sites poses several practical problems on account of the properties of the contaminants and the history of site :

- collection and preparation of samples from highly heterogeneous material,
- high short scale variability, particularly in presence of backfill,
- highly contrasted grades making the variogram inference complicated.

The sampling strategy generally adopted for contamined sites is based on the historical information. Systematic sampling recommended for geostatistical estimation is often considered to be excessive and unnecessary.

Two former coking plants are used as test cases for comparing several geostatistical methods for estimating (i) in situ concentrations and (ii) the probability that they are above a pollution threshold. Several practical and methodological questions are considered :

- the properties of various estimators of the experimental variogram and the validity of the results ;
- the use of soft data, such as historical information, organoleptical observations and semi-quantitative methods, with a view to improve the precision of the estimates ;
- the comparison of standard sampling strategies, taking into account vertical repartition of grades and the history of the site.

Multiple analyses of the same sample give an approximation of the sampling error. Short scale sampling shows the difficulty of selecting soils in the absence of a spatial structure. Sensitivity studies are carried out to assess how densely sampled soft data can improve estimates.

By using mainly existing models, this work aims at giving practical recommendations for the characterization of soil pollution.